Pedestrian Detection in Surveillance Video Based on Time Series Model

Hui Liu[1(✉)] and Liyi Xie[2]

[1] College of Information Engineering, Fuyang Normal University, Fuyang 236041, China
liuhongmei318@126.com
[2] Shandong Sport University, Jinan 250014, China

Abstract. To solve the problem of low detection accuracy caused by the occlusion of downlink people in complex scenes, a pedestrian detection method based on time series model in surveillance video is proposed. The gray values of pixels at the same location are regarded as a time series, and the mixed Gaussian model is used to recognize the pedestrian foreground. The threshold segmentation method is used to segment the image. The threshold segmented image is projected vertically to the X axis, and the overlapped pedestrian trough is used as the segmentation point. The bimodal feature window is projected vertically to segment the region of interest. Mark the feature points of the real image dataset, build an enhanced feature point detection network model, and obtain the descriptor detection results. The time-domain and frequency-domain information is represented as a symbol sequence, and the target is clustered using the equal length overlapping time window segmentation method, so that the location center of gravity does not change and the specified convergence degree is achieved. Balance the data fusion features to determine the pedestrian detection results in the surveillance video. The experimental results show that this method can detect all pedestrians, the maximum accumulated error of target recognition is 21%, and the maximum average accuracy of target matching is 90%, which proves that the detection effect is good.

Keywords: Time Series Model · Monitoring Video · Pedestrian Detection · Region of Interest

1 Introduction

Pedestrian detection is the first step in a large number of applications, such as intelligent video surveillance, auxiliary driving system, human-computer interaction, military applications and intelligent digital management. Due to the differences in light, color, scale, posture and dress, pedestrian recognition is a challenging problem. Pedestrian detection in images has a long history. in the past decade, people have great interest in pedestrian detection. Now we have entered a video surveillance society, and video surveillance phenomena can be seen everywhere in life. People are increasingly interested in pedestrian detection algorithms in video surveillance. The Internet of Things

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 533, pp. 104–118, 2024.
https://doi.org/10.1007/978-3-031-50574-4_8

technology is introduced into surveillance, and an intelligent monitoring system integrating identification, positioning, tracking, monitoring, early warning and management is established. In practical application, the system can accurately and comprehensively identify the changes of people, objects and environment within the monitoring range, and conduct intelligent analysis to achieve the identification of intrusion behavior. At present, most of the public security video monitoring systems and special video monitoring systems are used in daily operations and public security business, which is difficult to adapt to the new police needs [1]. In addition, the narrow scope of monitoring, low degree of intelligence, and incomplete auxiliary forensics system are also the main problems that lead to the monitoring system being unable to meet the new demand for security. The video surveillance system uses the method of computer automatic detection, tracking and recognition to obtain useful information from a large number of surveillance videos, and to understand and analyze it without relying on people as much as possible. With the development of video surveillance technology, the problem of single camera should be solved first, that is, multiple cameras should be used to replace a single camera, so as to solve the problem of large-scale surveillance. However, in the video surveillance system, how to judge the consistency of moving objects is a new difficulty. Among the current research methods, reference [2] proposes a multi-target tracking algorithm combining target detection and feature matching. Determine whether there is an obstacle according to the feature difference between the target itself and the current frame, and then track and detect it according to the remaining feature information after occlusion; Reference [3] proposes a real-time tracking system for moving objects based on binocular vision. The binocular stereo matching method is used to determine whether there is occlusion, then determine whether there is occlusion through matching error, and finally use gray correlation matching for follow-up tracking. Reference [4] proposes an end-to-end anomaly behavior detection network, which takes video packets as input and outputs anomaly scores. After the spatiotemporal encoder extracts the spatiotemporal features of video packets, it uses the attention mechanism based on hidden vectors to weight the packet-level features, and finally uses the packet-level pool to map the video packet scores to achieve behavior tracking. However, these three methods are vulnerable to the impact of dynamic targets, resulting in poor detection results. Therefore, a pedestrian detection method based on time series model in surveillance video is proposed. In the first section of this paper, based on the time series, the mixed Gaussian model is used to recognize the pedestrian foreground. The second section of the article marks the feature points of the real image data set, establishes an enhanced feature point detection network model, and realizes the region of interest division based on Gaussian background modeling. In the third section of the article, the time domain and frequency domain information are expressed as a time series, and the target is clustered using the equal-length overlapping time window segmentation method, so that the positioning center of gravity remains unchanged and reaches the specified convergence degree. The fourth section of the article balances the data fusion features and realizes the image feature balance processing based on data fusion. The last section obtains the pedestrian detection results in the surveillance video to achieve pedestrian detection in the surveillance video.

2 Pedestrian Foreground Recognition Based on Time Series Model

A pedestrian foreground recognition method based on time series model is proposed to solve the problem of pedestrian background brightness change and repetitive motion in video. For a video image, the gray value of a pixel at the same position is regarded as a time series, and the probability observation value of the pixel at time t can be expressed as:

$$G(I_t) = \sum_{i=1}^{n} \omega_{i,t} \times \rho\left(t_i, s_{i,t}, \sigma_{i,t}^2\right) \tag{1}$$

In formula (1), $\omega_{i,t}$ is the weight of the i Gaussian distribution at time t, and ρ is the Gaussian probability density function; t_i is a time series; $s_{i,t}$ is the expected value of the i-th Gaussian distribution at time t; $\sigma_{i,t}^2$ is the standard deviation of the i-th Gaussian distribution at time t.

Using formula (2) to mine surveillance video pedestrian information, which can be described as:

$$x_n = \beta_0 + \sum_{i=1}^{n} v_i t_i + \sum_{j=0}^{\mathrm{n}} v_j t_j \tag{2}$$

In formula (2), β_0 represents the dimension level information; v_i represents the information mining speed; v_j means the monitoring operation and maintenance management speed; t_i represents the information mining scalar time series; t_j represents the operation and maintenance management scalar time series; and n represents the mining times [5].

The association rule mining algorithm is used to feature mine the surveillance video pedestrian information and analyze the abnormal data mined by the surveillance video pedestrian information. Mining abnormal data using association rules to locate the surveillance video pedestrians. The time series model of the surveillance video pedestrian information collection constructed based on this is:

$$s(t) = \sum_{n=1}^{t} h_{x_n} g_n(t) \tag{3}$$

In formula (3), a_{mn} represents the potentially useful information amplitude; g_{mn} represents the multi-layer conjugate authentication coefficient.

According to the above mixed Gaussian model and time series model, whether a pixel is the background can be judged. The process is: first initialize the Gaussian function, and then analyze a new pixel. If the pixel observation value is within 2.5 standard deviations of a Gaussian function in the mixed Gaussian model, the pixel matches the corresponding Gaussian function; If there is no match, replace the Gaussian function with the lowest probability with a new one. If it matches, update the weight of all Gaussian functions [6]. Then, conduct normalization processing, normalize the newly generated ownership value, and update its matched Gaussian function parameters to:

$$s_{i,t} = (1 - \alpha) s_{t-1} + \alpha(t_i) \tag{4}$$

$$\sigma_{i,t}^2 = (1-\alpha)\sigma_{t-1}^2 + \alpha(t_i - s_t)^t(t_i - s_t) \tag{5}$$

In the above formula, α represents the learning parameters. Set the background threshold f_0 and take the first j Gaussian functions with the sum of weights greater than f_0 as the background model:

$$B = \arg\min_j \left(\sum_{i=1}^{j} f_i > f_0 \right) \tag{6}$$

In formula (6), f_i represents the number [7] of Gaussian functions included in a mixed Gaussian model. After determining the background model, the pixels in the image can be classified. If the observed pixel matches one of the first j Gaussian functions, the pixel is considered as the background, otherwise it is the foreground.

3 Monitoring Video Pedestrian Detection

3.1 Region of Interest Division Based on Gaussian Background Modeling

The luminance vertical projection curve of two independent pedestrians is also independent of each other; The vertical projection curve of brightness of pedestrians with occlusion also has adhesion [8]. The head of the adherent pedestrian is located at the peak of the two brightness projection curves respectively, and the adherent part is located at the valley bottom between the two peaks in the projection curve. Therefore, the specific steps of the direction projection method based on Gaussian background modeling to segment the region of interest are as follows, and the process is as follows:

Step 1: divide the video sequence image into foreground.

Step 2: threshold segmentation is performed on the foreground target. At this time, the foreground contains multiple pedestrians occluding each other. The threshold should be set in an adaptive way, and the corresponding segmentation threshold is:

$$u = \lambda \times m_{\max} + (1-\lambda) \cdot \overline{m} \tag{7}$$

In formula (7), λ represents the weighting coefficient; $m_{\max}$ is the image gray maximum; $\overline{m}$ represents the image gray mean.

Step 3: vertically project the thresholded image to the X axis, and vertically project the windows with bimodal features by taking the wave valley of the adherent pedestrian as the segmentation point to obtain the gray level vertical projection curves of the three images. The first is that the gray-scale area of the human body shows a convex peak without adhesion [9]; The second is to determine the midpoints of the ascending and descending curves on both sides of the convex peaks in the projection curve for quantitative calculation, and take these two midpoints as the starting and ending points of a brightness band respectively, so as to obtain a series of brightness bands perpendicular to the X axis, while the possible areas of the human body are included in the brightness band; The third is the adhesion of vertical projection curves, which needs to be handled separately.

Step 4: horizontally project the brightness band obtained from vertical projection to the Y axis. The starting and ending points of the brightness band are selected in the same way as for vertical projection.

Step 5: Put the brightness band obtained from vertical projection and horizontal projection into the corresponding position in the original image at the same time. At this time, the original image can be divided into many high brightness rectangular areas. The above method is suitable for independent and unobstructed object interest determination. The pedestrian can be detected by inputting the interest directly into the trained convolutional neural network. However, this method cannot accurately deal with the situation of pedestrians blocking each other, and the detection rate of the system is low [10–12]. Therefore, the directional projection method based on Gaussian background modeling is used to determine the interest of occluded pedestrians. The region of interest is determined. The two people who occlude and adhere to each other are respectively included in two different rectangular brightness boxes, that is, they are divided into two regions of interest.

3.2 Construction of Enhanced Feature Point Detection Network Model

Segmenting occluded and non occluded regions of interest, outputting multi view data fusion features, and balancing pedestrian detection network can achieve pedestrian detection in video. Firstly, the input multi view images are matched to form a complete image, and then the target detection network is used to train the fused image to improve the accuracy of occlusion and long-distance small pedestrian detection.

3.2.1 Real Image Dataset Feature Point Annotation

The workflow of the multi view data fusion model of self supervised learning is as follows: image acquisition, self supervised feature point and descriptor extraction, feature matching, and finally multi view image fusion. In the process of multi view data fusion, it is difficult to use manual annotation to extract feature points from data sets. For the annotation of traditional detection and segmentation tasks, given an image, the semantic truth value can be determined by marking the rectangular box or the outline of the object. However, for the feature point detection task, it is difficult to determine which pixel can be used as the feature point manually. Therefore, the basic dataset containing only simple geometric shapes and the self collected dataset are used to automatically label the dataset. The specific process is as follows:

Model pre training using simple geometric shape data set simple geometric shape data set is composed of some images whose feature points are easy to determine, such as line segments, polygons, cubes, etc. The true values of data sets and feature points can be obtained by using scale invariant feature transformation to extract feature points from basic data sets. Because the feature points of the basic geometric shape images such as line segments and triangles are subsets of the real image feature points, a primary feature point detection network is obtained by training the feature point detection network using the labeled simple geometric shape data set. Compared with traditional algorithms such as scale invariant feature transformation, the primary feature point detection network trained in simple geometric shape data sets has certain advantages in accuracy, but

when extracting feature points from real image data sets, there will be some missing feature points, and the detection accuracy is low. Therefore, a new model is obtained by using homography adaptive transformation and primary feature point detection network training to improve the accuracy of feature point extraction of real images.

The input image is processed by multiple composite geometric transformations, and the super parameter is set to 80 frames, that is, the super parameter is the original image without composite geometric transformation, and the remaining 79 frames are the images formed by the original image through randomly generated composite simple geometric transformation. The generated primary feature point detection network extracts the pseudo feature points of the real image data set, maps the 79 frames of images corresponding to the source image back to the feature points of the original image, and accumulates them to form a new source image feature point, thus completing the feature point annotation of the real image data set.

3.2.2 Construction of Enhanced Feature Point Detection Network Model

In the compound simple geometric transformation, 79 frames of the source image transformation image formed by the known transformation matrix are obtained, so 79 sets of image pairs of known pose transformation of the source image and its corresponding 79 frame images are obtained. In this way, the true value of the mapping relationship between the original image and the transformed image is obtained. The final self collected data set contains feature points and feature point descriptor truth values, which are used for joint training of feature point detection and descriptor detection network branches in the feature point detection network. In order to realize the joint training of feature point detection sub network and description sub network in the primary feature point detection network, the loss function values of the two detection sub networks are weighted and added to obtain a unified loss function. In order to fuse information from different perspectives, it is necessary to find the corresponding relationship between different perspectives. The adaptive homography transformation is used to solve the corresponding relation matrix of different perspectives. The composite simple geometric transformation matrix learned through self-monitoring is not all useful and needs to be selected. In order to select a composite simple geometric transformation matrix with good performance, truncated normal distribution is used to sample translation, scaling, in-plane rotation and symmetric perspective transformation within a predetermined range. Based on this, the constructed enhanced feature point detection network model is shown in Fig. 1.

After obtaining the true value of the mapping relationship between the original image and the real image of the data set according to Fig. 1, the automatic annotation of the real data set is completed, which realizes the automatic annotation of the real image data set that is difficult to label manually. The enhanced feature point detection network is used to train the previously acquired automatically labeled image dataset to improve the accuracy of feature point extraction.

Multilevel encoder: in order to give consideration to real-time and accuracy, the enhanced feature point detection network is designed into two branches to handle different tasks. The upper branch extracts the deep feature points of the original image through an asymmetric encoding and decoding network. The feature descriptor of the original

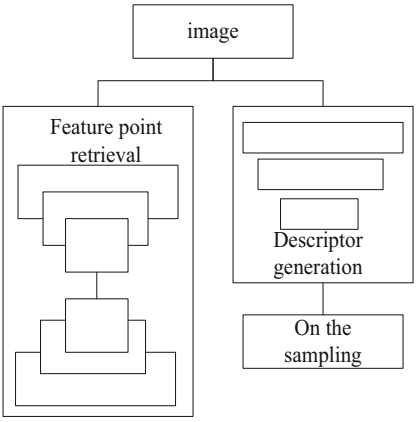

Fig. 1. Enhanced feature point detection network model

single view image is generated, and the surface feature description of the original image is extracted through a multi-channel, low-level encoder network.

Feature point detection: when detecting the network part of the feature point, the feature point of the image is obtained through a deep, few channel, asymmetric encoding and decoding network.

Fusion network: because the feature map of the network does not have the same channel and size, the features extracted by the descriptor generation network are shallow and contain a lot of location information, while the feature point detection network obtains deep feature points after multi-layer encoder, including information such as arm and face. In order to fuse features at different levels, the fusion network first realizes simple fusion of feature maps at different levels through Concatmate operation. In order to balance features of different sizes, the BatchNorm operation is used after Concatmate. The connected features are pooled globally $\times$ 1 Convolution to get a new weight. The purpose of this is to make a new feature selection and combination for the connected features. So far, the descriptor detection results are obtained.

3.3 Target Clustering Based on Time Series Segmentation

The method of piecewise symbolic linear representation is to segment the original time series and express the time domain and frequency domain information as symbol sequences. This method can not only ensure that the data pattern with long duration can be completely separated, but also maintain the dependence of the original time series data on time sequence. The specific implementation of the equal length overlapping time window segmentation method is as follows:

Let the sample set containing N multidimensional time series be marked as $T = \left\{T^1, T^2, \cdots, T^n, \cdots, T^N\right\}$, in which a single sample of length T can be represented as $T^n = \left(t_0^n, t_1^n, \cdots, t_1^n, \cdots, t_T^n\right)$. Each sample T^n in the sample set applies a sliding window with window size d and step length l. The division of the sample is represented as follows:

$$T^{ln} = f\left(T^{n}\right) = \begin{bmatrix} T_1^n \\ T_2^n \\ \vdots \\ T_m^n \end{bmatrix} = \begin{bmatrix} \left[T_1^n : T_{1+m}^n\right] \\ \left[T_{1+l}^n : T_{1+l+m}^n\right] \\ \vdots \\ \left[\mathbf{x}_{1+(i-1)l}^n : \mathbf{x}_{1+(i-1)l+m}^n\right] \\ \vdots \\ \left[\mathbf{x}_{1+(m-1)l}^n : \mathbf{X}_T^n\right] \end{bmatrix} \tag{8}$$

In formula (8), m represents the total number of time series after a slice.

The autoencoder can retain important features in the data while squeezing high-dimensional data. It uses multiple self-compilers to extract features in different time series fragments. Figure 2 is a schematic diagram of the process of the autoencoder training process.

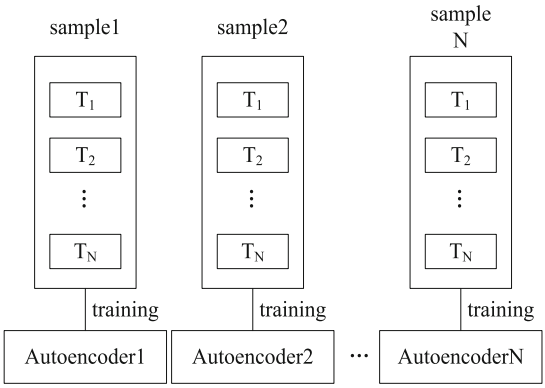

Fig. 2. Schematic diagram of the autoencoder training process

For normal time series sample segmentation processing, and all the samples in the same location of the data fragments, through the extraction in the nonlinear features, based on the principle, monitoring video pedestrian target clustering steps are as follows: first of the pre-processing of each sample, and then smooth, strengthen, and then using the super segmentation method, complete the monitoring video pedestrian feature extraction. For image segmentation processing, the interval between the target point and the hyperplane is calculated with the formula:

$$\gamma = \frac{1}{\|\overline{a}\|}\left|\overline{a}^{\Delta t} x_n + b\right| \tag{9}$$

In formula (9), $\overline{a}$ represents the normal vector; b the determination threshold; Δt the split time. According to this formula, the maximum conversion value of the calculation interval is $\frac{1}{\|\overline{a}\|}$, and judging from the maximum value of the norm equivalent, it can

continue to be converted to the minimum value of the normal vector to ensure that the classification spacing reaches the maximum.

In the extracted visual images, similar methods are used to integrate images with similar pixels to form a visual image library. The data of each cluster subset is averaged using the standard function of the sum of squares of errors, and then classified and optimized using iterative methods. The square error function is as follows:

$$e = \sum_{c=1}^{k} |N - O_c|^2 \tag{10}$$

In formula (10), N represents sample data; O_c represents the data cluster center; k represents the square error and the number of partitions; c represents the number of clusters. In a given sample, N initial centers are randomly and uniformly generated. The distance between each data sample point and K center centers is calculated respectively, and they are designated as the closest center, which is temporarily classified into one category. The mean value of each type of samples is calculated, and the center of gravity is repositioned. The clustering can be completed when the center of gravity does not change or the specified convergence degree is reached.

3.4 Characteristic Balancing Processing Based on Data Fusion

The enhanced feature point detection network is a single-stage target detection method. Different from the target detection framework of the RCNN series, the enhanced feature point detection network does not generate candidate boxes, but directly returns the location and category of the boundary box at the output layer. The enhanced feature point detection network draws on the ideas of residual network and feature pyramid network, adds cross layer jumping connection, and integrates the features of coarse and fine granularity, which can better realize the detection task. Multi scale prediction is added, that is, prediction is made on three characteristic layers of different sizes, and three anchor frames are predicted for each scale. The anchor frame is designed by clustering, and nine clustering centers are obtained, which are divided into three feature layers according to the size, and three features of different sizes are fused.

The feature extraction network of the enhanced feature point detection network is a balanced network, and its network structure is shown in Fig. 3.

It can be seen from Fig. 3 that the Convo logical in the feature balance network represents an activation function, and its operation process includes convolution layer, batch normalization layer and activation function. For the feature balanced network, the inseparable parts of the convolution layer and the hidden layer together constitute the minimum component, which uses deep features and shallow features for small-scale pedestrian detection.

After the difference between background and foreground is obtained, the object and background are completely segmented by binarization. The binary processing can eliminate the background drag retained by the differential processing, only retain the dynamic range of the background, and mainly focus on moving targets. There are still some less important interferences in the binary processing, which can be filtered in the

	Type	Number of convolution kernels	Convolution kernel size	Step length
1	convolution	32	1×1	1
	convolution	64	3×3	2
2	convolution	64	1×1	1
	convolution	128	3×3	2
3	convolution	128	1×1	1
	convolution	256	3×3	2
4	convolution	256	1×1	1
	convolution	512	3×3	2
5	convolution	512	1×1	1
	convolution	1024	3×3	2

Fig. 3. Feature-balanced network structure

subsequent process. After the binary processing, the thresholding method is used to determine whether the pixel is located on the detection target. Moving objects are prone to smear, and the median filter can effectively preserve the edge characteristics of the image without paying attention to details. Median filtering is a nonlinear window filtering algorithm, which can retain image details while removing noise. The basic principle of median filtering is: divide any pixel according to the pixels of adjacent areas, and replace the central pixel value with the median value. The median filtering has a good effect on noise processing. The interference caused by large discrete points between adjacent regions can be effectively eliminated by replacing pixel values with intermediate points. Compared with mean filtering, this method has better filtering effect and can retain image details better.

3.5 Surveillance Video Pedestrian Detection

Before detecting a target surveillance video pedestrian, the surveillance operator must divide the area of interest within the surveillance scene, which is usually a pre-set rectangular area. Let the known delimited region of interest be rectangular W_e, and the coordinates in the upper left corner of W_e be (x_1, y_1), and those in the lower right corner of W_e be (x_2, y_2). If the monitored pedestrian mass center of the video is a rectangle, and the external rectangle area of the video pedestrian exceeds the set threshold, the pedestrian is considered passing by and giving an alarm, otherwise the alarm will not be issued. The judgment formula is:

$$t = \begin{cases} true, if \quad Area_{E_e} \\ false, \quad else \end{cases} \tag{11}$$

In formula (11), *true* means video surveillance captured pedestrians; *false* indicates video surveillance missed pedestrians; and $Area_{R_e}$ indicates a rectangular area. In the

monitoring system, an alarm is raised when a pedestrian enters a predetermined area of interest, and not when the intended range is not reached. An alarm can be given during the entire detection process when entering the predefined area of interest rectangle through monitoring the pedestrian's center of mass, otherwise it will not be given.

4 Experiment

4.1 Experimental Environment

The experimental platform is ECS, the operating system is Ubuntu 16.04, the graphics card model is GeForce GTX 2080Ti, the video memory is 11 GB, the memory is 16 GB, the Cuda version is 10.0.130, and the OpenCV version is 3.2.00.

4.2 Experimental Data Set

The training and testing data sets used in this experiment are all from PASCAL VOC data sets. VOC2007train, valid and VOC2012 train, valid data sets are used for training. In order to verify the effectiveness of the method, the VOC2007 test data set is used for verification. The total training data is 22136 pictures, including 6496 pictures of pedestrians, and the total validation data is 4952 pictures, including 2097 pictures of pedestrians.

4.3 Experimental Parameter Setting

Only the pedestrian category is trained. The default size of the input image is 416 $\times$ 416, the number of input channels is 3, the set number of iterations is 50200, the batchsize is 64, and the learning rate is 0.001. When the number of iterations reaches 40000, the learning rate is updated to 0.01, and the processed dataset is trained in the same performance server. Under the same experimental environment and experimental parameters, the network is trained.

4.4 Experimental Evaluation Index

In order to evaluate the detection effect numerically, F_A and F_B are two indicators are used, in which F_A indicates the accumulation degree of target identification error, and F_B indicates the measurement of the average accuracy of target matching. The calculation formula is as follows:

$$F_A = 1 - \frac{P_1 + P_2}{\sum_t B_t} \tag{12}$$

$$F_B = \frac{\sum_{t_i i} e_{t,i}}{\sum_i \mathrm{D}_t} \tag{13}$$

In formula (12), P_1 represents the omission rate; P_2 represents the false alarm rate; and B_t represents the actual number of targets at time t. In formula (13), $e_{t,i}$ represents the i th target matching error at time t; D_t indicates the number of successfully matched targets. The smaller the results, the better the detection effect. The larger the calculation results are F_A, the better the detection results are F_B.

4.5 Subject Determination

Three different infrared image test sets were used for infrared human body test experiments. Test set 1 comes from the video monitoring results of traffic pedestrians. It is a multi-human test set, which includes pedestrians' adhesion (occlusion). Test set 2 comes from video monitoring results of traffic footpaths, a 2-person test set. Test set 3, derived from the video monitoring results of traffic footpaths, is a 1-person test set, and the subjects are shown in Fig. 4

(a) test set 1

(b) test set 2

(c) test set 3

Fig. 4. Experimental subjects

The test set shown in Fig. 4 is the standard test result to analyze the rationality of the surveillance video pedestrian detection method based on the time series model.

4.6 Experimental Results and Analysis

The design method, reference [2] method, reference [3] method and reference [4] method were used for comparative test. The test results are shown in Fig. 5.

(a) The reference [2] method (b) The reference [3] method

(c) The design method (d) The reference [4] method

Fig. 5. Comparative analysis of the detection results of the three methods

As can be seen from FIG. 5, the reference [2] method are used for feature matching, but lack of noise processing, and multi-target is taken as a background image occurs. Neither The the reference [3] nor reference [4] method identified all of the occluded targets. But use design method can identify all pedestrians through multi-target background recognition and noise processing.

To further verify the reliability of the surveillance video pedestrian detection method based on the time-series model, the calculation results of the three methods F_A and F_B were compared, as shown in Fig. 6.

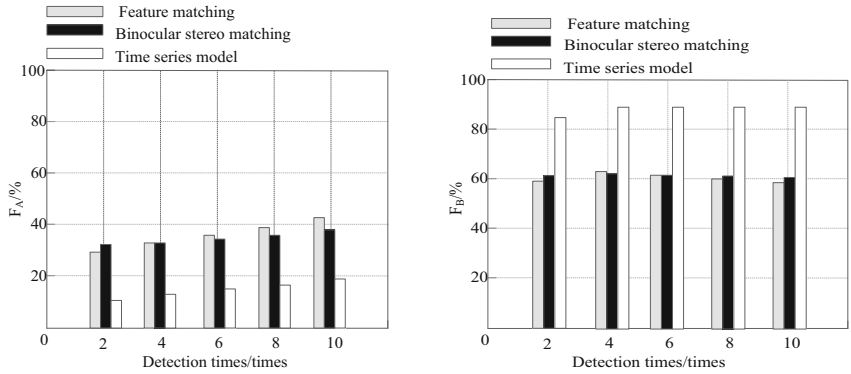

Fig. 6. Comparative analysis of the calculation results of the three methods F_A and F_B

According to Fig. 6, the maximum F_A is 42% based on feature matching, F_B and 62%, the detection method F_A 38% and F_B 61%, and the time series model based detection method F_A 19% and F_B 90%. According to the above analysis results, using the time-series-based model detection method is more effective.

5 Conclusion

The pedestrian detection method in surveillance video based on time series model is studied. First, foreground segmentation is performed to identify pedestrian foreground. Then the region of interest is segmented and the relevant features are extracted. Finally, an enhanced feature point detection network model is constructed, and pedestrian detection is performed by using time-domain and frequency-domain information as symbol sequences. The comparison experiment verifies the judgment effect of the method and increases the authenticity of the method research. Although the target detection has been greatly improved, there are still some defects. In the next step, we will focus on the following contents: Although the coverage of the image can be reduced by using segmentation methods, due to the existence of non relative moving objects, corresponding recognition algorithms must be used for segmentation and correction, which not only wastes a lot of time, but also reduces the segmentation accuracy. In the future research, we will focus on how to reduce the error and further improve the segmentation accuracy.

Aknowledgement. 1. 2018 Anhui Provincial University Natural Science Research Key Project: Research and Application of Intelligent Pedestrian Detection and Tracking Method Based on Digital Video (KJ2018A0669)

2. 2017 Anhui Provincial University Natural Science Research Key Project: Research on Key Technologies of High-Performance Computer Fault-tolerant Systems in Heterogeneous Environments (KJ2017A837)

References

1. Shen, X.: live in peace in white. The study for scene recognition of surveillance video based on semi-supervised feature fusion. Comput. Simulat. **38**(1), 394–399 (2021)
2. Ye, L., Li, W., Zheng, L., et al.: Multiple object tracking algorithm based on detection and feature matching. J. Huaqiao Univ. Nat. Sci. **42**(5), 661–669 (2021)
3. Zhang, J., Ji, F.: Moving target real-time tracking system based on binocular vision. China Comput. Commun. **34**(5), 122–124 (2022)
4. Xiao, J., Shen, M., Jiang, M., et al.: Abnormal Behavior Detection Algorithm With Video-bag Attention Mechanism in Surveillance Video. Acta Automatica Sinica **48**(12), 2951–2959 (2022)
5. Jian, Y., Ji, J.: Research on pedestrian identification model in optical sensor monitoring system Laser J. **41**(03), 82–85 (2020)
6. Zhang, B., Zhao, W., Duan, P., et al.: Surveillance Video Re-Identification with Robustness to Occlusion. J. Signal Process. Signal Process. **38**(06), 1202–1212 (2022)
7. You, F., Liang, J., Cao, S., et al.: Dense pedestrian crowd trajectory extraction and motion semantic information perception based on multi-object tracking. J. Trans. Syst. Eng. Inform. Technol. **21**(06), 42–54+95 (2021)
8. Liu, J., Li, X., Ye, L., et al.: Pedestrian detection algorithm based on improved RetinaNet. Sci. Technol. Eng. **22**(10), 4019–4025 (2022)
9. Qi, P., Wang, H., Zhang, J., et al.: Crowded pedestrian detection algorithm based on improved FCOS. CAAI Trans. Intell. Syst. **16**(04), 811–818 (2021)
10. Liu, S., He, T., Dai. J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mobile Netw. Appli. (2021)
11. Liu, S., He, T., Dai, J., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)
12. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-Learning and e-Education. Mobile Netw. Appli. (2021)