IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 5, October, 2022, pp. 1815-1821
Journal homepage: http://iieta.org/journals/ts

Pedestrian Identification Method Based on Multi-Scale Feature Learning in Surveillance Video Images

Aihua Li[1], Ling Pang[2*], Lei An[3], Zihui Che[1]

[1] College of Data Science and Software Engineering, Baoding University, Baoding 071000, China
[2] School of Computer and Information Engineering, Hebei Finance University, Baoding 071051, China
[3] College of Artificial Intelligence, Baoding University, Baoding 071000, China

Corresponding Author Email: pangling@hbfu.edu.cn

https://doi.org/10.18280/ts.390541

Received: 3 June 2022
Accepted: 29 August 2022

Keywords:
multi-scale feature learning, surveillance video, pedestrian identification

ABSTRACT

With the development of deep learning technology and people's demand for intelligent security, human-computer interaction, shopping guide and other technologies, computer vision technology for pedestrian identification shows great application value. In this paper, pedestrian identification method based on multi-scale feature learning in surveillance video images is studied. Firstly, the deep residual network ResNet and densely connected convolutional network DenseNet are introduced as baseline networks. A model is constructed based on hybrid hourglass network module, enhanced weighted feature pyramid fusion network module and post-processing module. The loss function is designed, which is unified with other traditional models, and the optimization objective of the loss function is respectively corresponding to three parts, namely, the prediction error of corresponding center point, the prediction error of offset and the prediction error of bounding box size. The experimental results verify the effectiveness of the proposed model.

1. INTRODUCTION

Better use of increasing audio, text, video and other information to assist human production and life has been a concern of scholars at home and abroad [1-7]. With the development of in-depth learning technology and people's demand for intelligent security, human-computer interaction, shopping mall guidance and other technologies, computer vision technology for pedestrian identity recognition shows great application value [8-13]. Influenced by various problems in the actual surveillance shooting scene, the appearance difference, obscured field of view, poor lighting conditions, unclear images or incomplete shooting and other problems limit the rapid application of pedestrian identification technology, which still requires in-depth research [14-21]. Innovating the traditional deep learning model, obtaining higher generalization ability and robustness, and achieving accurate extraction of pedestrian feature representation with high discrimination are the efforts of scholars to carry out relevant research.

Liu et al. [22] converted the cross-model human recognition task from visible infrared image to gray-scale infrared image, which is called minimum modal difference. A pyramid feature integration network (PFINet) is proposed. The network mines the distinctive detailed features of pedestrian images and combines high-level and semantic features to build a robust pedestrian representation. It is a challenging process to recognize others in multispectral images without spatial alignment. The main contribution of Brehar et al. [23] is the matcher based on key points, which uses a solution based on depth learning to find relevant key points of human posture, and uses local neighbor search to extract the best candidate object for identity recognition. Zhang et al. [24] studied a video-based re-ID model with non-local attention modules. Based on the residual module embedded in the three-dimensional convolutional neural network, a non-local attention module is added to enrich the pedestrian feature representation from both local and global aspects. Due to the differences in camera pixels, posture, lighting, occlusion and intra class changes between different cameras, the task of pedestrian identification is still challenging for computer vision scientists. Cao et al. [25] proposed a multi-task network based on even partition of network, which can simultaneously calculate the recognition loss and verification loss of two input images. A pair of images is given as input. The system predicts the identities of the two input images, and simultaneously outputs similarity scores to indicate whether they belong to the same identity. Li et al. [26] proposed a multi-path self-adaptive pedestrian alignment network (MAPAN) to learn distinguishing feature representation. The multi-path network directly learns features from data in an end-to-end manner, and self-adaptively aligns pedestrians without any additional manual annotation. In order to alleviate the cross modal difference between the visible domain and the infrared domain, the distinguishing features of the two modes are mapped to the same feature embedded space, and the identity loss and triplet state loss are combined into the total loss. A large number of experiments show that this method has superior performance compared with the existing technology.

From the perspective of research status, the current pedestrian identification methods in surveillance video are difficult to solve such problems as scale imbalance and imbalance of positive and negative samples. Therefore, this paper conducts research on pedestrian identification methods in surveillance video images based on multi-scale feature learning. In the second chapter, the paper first introduces the

deep residual network ResNet and the densely connected convolutional network DenseNet as the baseline network. Then in the third chapter, the model is built based on the hybrid hourglass network module, the pyramid fusion network module with enhanced weighted features and the post-processing module. The fourth chapter designs a loss function that is unified with other traditional models. The set optimization objective of the loss function is respectively corresponding to three parts: the prediction error of the center point, the prediction error of the offset and the prediction error of the boundary box size. The experimental results verify the effectiveness of the model built by our institute.

2. INTRODUCTION TO THE BASELINE NETWORK MODEL

The traditional method of pedestrian identification in surveillance video images uses ResNet, a deep residual network with shortcut connection, as the baseline network, to extract high-level features of surveillance video images while retaining some valuable low-level information. In order to obtain ideal experimental results, shortcut connections in residual blocks usually span two or more layers. Assuming that the calculation result of removing the shortcut connection from the residual module is represented by F (a), the calculation result after introducing the shortcut connection can be obtained by the following formula:

$$G(a) := F(a) - a \tag{1}$$

Assuming that the ownership value of the residual module is expressed by $\{q_i\}$, the following formula gives the expression of the calculation result of the residual module:

$$b = G\left(a, \{q_i\}\right) + a \tag{2}$$

Assuming that the residual mapping function is represented by $G(a, \{q_i\})$, if only two weight layers are considered, then:

$$G\left(a, \{q_i\}\right) = q_2 \phi\left(q_1 a\right) = q_2 \operatorname{Re} LU\left(q_1 a\right) \tag{3}$$

The calculation of the residual module requires that the dimensions of $G(a, \{q_i\})$ and a be consistent. If they cannot be consistent, then a linear transformation of xa is required. Generally, it is realized by introducing the weight matrix q_r, then:

$$b = G\left(a, \{q_i\}\right) + q_r a \tag{4}$$

In other studies, *DenseNet*, a densely connected convolutional network, is used as the baseline network for pedestrian identification of surveillance video images. The connection between any layers in *DenseNet* densely connected blocks can be a spanning way. Figure 1 shows the structure of the close connection module. Assuming that the monitoring video image feature map of the front k-1 layer is represented by $a_0, a_1, \ldots, a_{k-1}$, then the feature map of the k_{th} layer of the closely connected blocks can be spliced based on $a_0, a_1, \ldots, a_{k-1}$. Assuming that the feature connection corresponding to the feature map generated from layer 0 to layer k-1 is represented by $[a_0, a_1, \ldots, a_{k-1}]$, the block normalization module, modified linear module and 3×3 convolution module are processed equivalently to a composite function $F_k(.)$, then the relevant formula is:

$$a_k = F_k\left(\left[a_0, a_1, \ldots, a_{k-1}\right]\right) \tag{5}$$

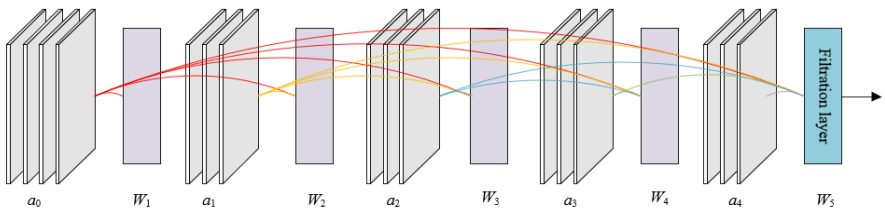

Figure 1. Structure of closely connected module

3. CONSTRUCTION OF PEDESTRIAN IDENTIFICATION MODEL BASED ON MULTI-SCALE FEATURE LEARNING

The breakthrough of deep learning has brought many advanced methods for the realization of pedestrian identification in surveillance video images. In order to maintain the recognition accuracy and solve the problems of unbalanced positive and negative samples and excessive super parameters, this paper constructs a pedestrian area detection framework for surveillance video images based on the idea of arx hor-free algorithm. Due to the imbalance of scales and the complexity of feature fusion module, the pedestrian identification effect directly using this algorithm will not be particularly ideal. Therefore, this paper is committed to building a lightweight feature extraction and fusion network model that can adapt to pedestrian areas in surveillance video images of different sizes and shapes, so as to effectively use multi-scale information.

Referring to the mainstream structure of the existing identity recognition framework, the model is built based on the hybrid hourglass network module, the pyramid fusion network module with enhanced weighted features and the post-processing module. Figure 2 shows the overall architecture of the model.

The hybrid hourglass network module is based on each of the internal hourglass modules to extract the whole body features of pedestrians in different scale surveillance video images, breaking through the limitation of single scale feature extraction in traditional models. Following the principle of feature extraction from coarse to fine, the order of each hourglass module of the hybrid hourglass network module is different, so the feature map information output by this module is also different. Such hierarchical design also realizes the full

use of top and bottom features from top to bottom with the help of the pyramid fusion network module with enhanced weighted features, which makes the semantic information and spatial information of the pedestrian area in the monitored video image more fully integrated, and the information contained in the output results more valuable. The output of the post-processing module contains the horizontal and vertical offsets corresponding to the center point position of the pedestrian area in the monitored video image. Here, it is necessary to reasonably set the size of the bounding box to obtain a more stable prediction of the size of the bounding box.

The key problems of pedestrian identification in surveillance video images based on depth learning are the imbalance of positive and negative samples and the imbalance of scales. Due to the difference in the size and shape of pedestrian areas in different surveillance video images, it is difficult to achieve ideal feature extraction results using only a single scale model. In order to make the model more robust for pedestrians with different body shape, action and attitude characteristics, it is necessary to adapt multi-scale detection and then summarize the recognition results of pedestrian areas in surveillance video images of different sizes.

In order to extract effective multi-scale features of pedestrian's shape, action and posture and to be compatible with multi-scale feature fusion modules, it is necessary to set different sizes for each hourglass block of the hybrid hourglass network module. Figure 3 shows the basic structure of the hourglass network. The input size of hourglass block decreases with the increase of network level. To simplify the calculation process of the model, only one residual module is set for each hop connection.

Before entering the hybrid hourglass network module, the original surveillance video image will be preprocessed with a convolution layer and a residual block. The volume layer size is 7×7, the step size is 2, and the number of channels is 128. The residual block down sampling multiple is 2, and the number of output channels is 256.

Then the preprocessed feature map is copied and the feature fusion module and the hybrid hourglass network module are input respectively. In order to retain the spatial information of pedestrian area in the surveillance video image, the first hourglass block will obtain the feature map with the most complete spatial information. If the most complete feature map of spatial information is input to the hourglass block with larger order, the probability of losing the pedestrian area spatial information in the surveillance video image will increase, especially the pedestrian area with smaller size in the surveillance video image.

After the feature extraction of the first hourglass block is completed, its output feature map will be further down sampled based on the maximum pooling operation. Then the feature map is copied and the feature fusion module and the next hourglass block are input respectively. Similarly, the previous operation is repeated for the second hourglass block. And so on, until all the output feature maps are obtained and the feature fusion module is input.

Although the features extracted by the hybrid hourglass network come from different layers, they are relatively independent and cannot share semantic and spatial information. In order to ensure the fusion effect of different scale surveillance video feature maps, this paper constructs an enhanced weighted pyramid fusion network module to fuse the features of all feature maps output by the mixed hourglass network.

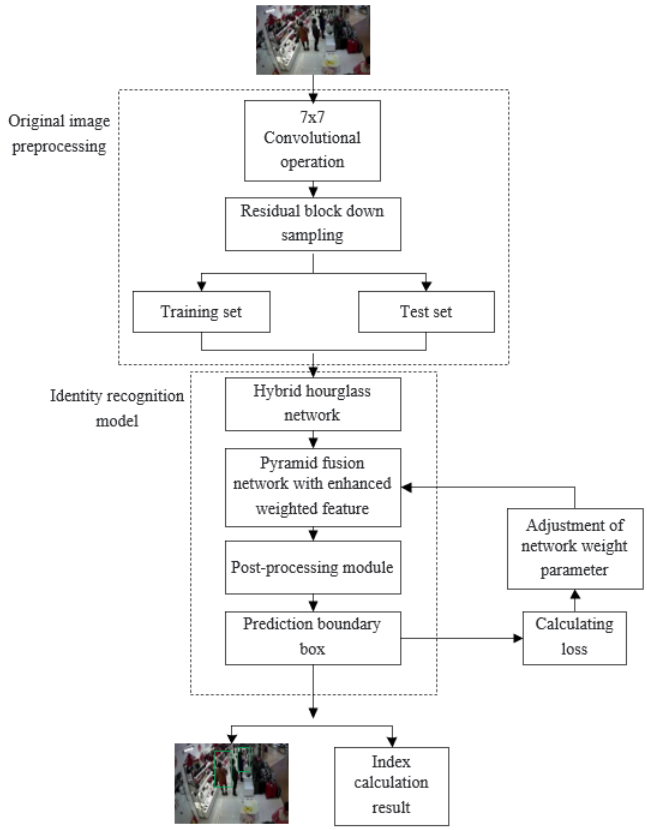

Figure 2. Overall architecture of the model

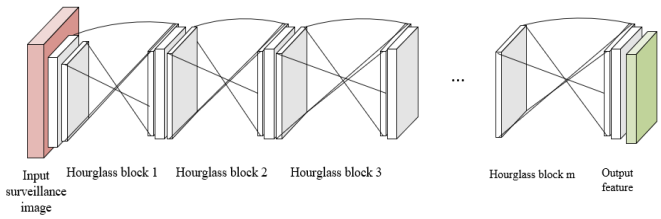

Figure 3. Basic structure of hourglass network

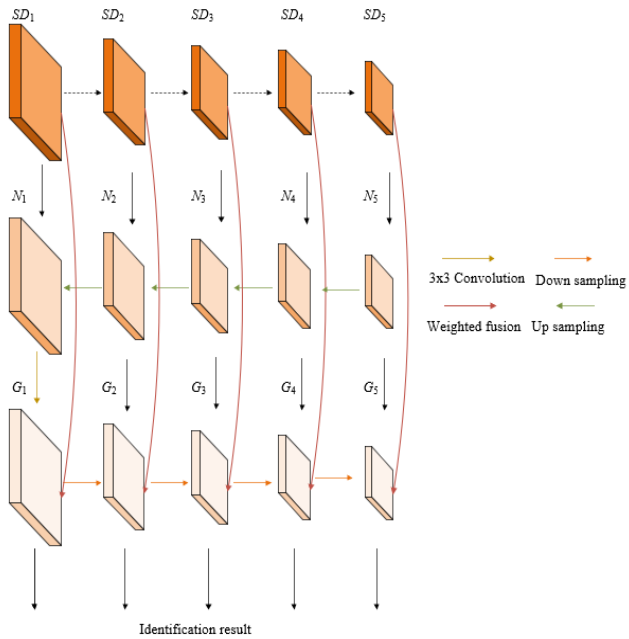

Figure 4. Enhanced weighted pyramid fusion network structure

The pyramid fusion network module with enhanced weighted features adopts the bottom-up feature fusion mode. Figure 4 shows the pyramid fusion network structure with enhanced weighted features. Specifically, the final output result corresponding to the layer 1 features output by the hybrid hourglass network consists of the output features of the input layer and the output features of the middle layer. The output features of the input layer need to be weighted to some extent. Assuming that the input characteristics of Layer i are represented by SD_i, the characteristics of Layer I of the middle top-down path are represented by N_i, and the weight coefficient of SD_i is represented by q_i, meeting the requirement of $q_i \geq 0$. q_i is set based on the contribution of each layer to SD_i, and the parameters to ensure network stability are represented by λ. The traditional convolution operation is represented by UJ, the depth separable convolution operation is represented by $FLCO$, and the down sampling operation is represented by DS, then

$$G_i = \begin{cases} FLCO\left(\dfrac{q_1 \cdot SD_1 + N_1 + UJ(N_1)}{q_1 + \lambda}\right), i = 1 \\ FLCO\left(\dfrac{q_i \cdot SD_i + N_i + DS(G_{i-1})}{q_i + \lambda}\right), i \geq 2 \end{cases} \quad (6)$$

Boundary box regression is another important task of pedestrian identification in surveillance video images, that is, pedestrian area size prediction in surveillance video images. In this paper, the prediction boundary box size is obtained based on the center point direct prediction method. In order to make the prediction results more accurate and stable, and adapt to different pedestrian targets, this paper introduces the linear transformation from center point to corner point as shown in the following formula. It is supposed that the coordinates of the predicted center point are represented by (a,b), any corner point of the truth value box is represented by (a_1,b_1), the number of layers of the feature pyramid is represented by k, and the normalization factor is represented by c, then there are:

$$e_q = log\left|\frac{2^k(a+0.5)-a_1}{c}\right|, e_f = log\left|\frac{2^l(b+0.5)-b_1}{c}\right| \quad (7)$$

To control the size of pedestrian target, the loss function shown in Equation 8 is set in the model built in this paper. If the area of the truth value box is represented by R_l and the area of the predicted boundary box is represented by R_l^*, then there are:

$$FE_{SI} = \frac{1}{M}\sum_{l=1}^{M}\left|R_l^* - R_l\right| \quad (8)$$

The post-processing module mainly processes the output characteristics of the hybrid hourglass network module. 3 different 1×1 convolution kernels are adopted to perform convolution operation on the output characteristics of each hourglass block, obtaining 1 thermodynamic diagram, M length-width vectors and M offsets, with the size of D×Q×W, 2D and 2D respectively. The number of pedestrian areas detected in the surveillance video image is represented by M, and D is the number of pedestrian area classification categories in the surveillance video image, which is set as 1.

The thermal map is mainly used to predict the category and coordinates of the center point in the pedestrian area in the monitoring video image. In order to filter the repeated center point, this paper first maximizes the pool operation of the thermodynamic diagram. Because the receptive field of each pixel in the feature map in the surveillance video is 4×4, the convolution kernel size is set as 3×3, which means that the center point in each 12×12 block will not be repeated in the input surveillance video image. Through the above operations, higher model execution efficiency can be obtained. The main reason is that the non-maximum suppression operation in traditional methods is abandoned.

The score of each pixel of the thermal map after the maximum pooling operation is compared with the neighboring 8 pixels. If it has the highest score, it is determined that it is the alternative center point. The top 100 centers with the highest score are obtained through this step.

The process of feature extraction for surveillance video images involves more down sampling and up sampling operations, which will make the center point position of pedestrian area in surveillance video images shift. This paper predicts the center point offset based on 100 center points obtained. The original feature map in the monitoring video goes through 1×1 convolution, and converts to the prediction diagram with a size of $2\times Q\times W$. It is a 2-channel prediction diagram, and the two channels respectively correspond to the offset of the x-axis and y-axis corresponding to the center point. Finally, in order to obtain a more stable and accurate prediction boundary box, the $C2C$ transformation operation is performed on the prediction size.

4. OPTIMIZATION OF LOSS FUNCTION FOR THE FACE RECOGNITION MODEL

The model constructed in this paper belongs to the pedestrian identification model based on key points. In order to compare the recognition performance with that of other existing models based on key points, this paper designs a loss function that is unified with other traditional models. The set optimization objective of loss function is divided into three parts: prediction error of corresponding center point, prediction error of offset and prediction error of boundary box size. Among them, the most critical part is the center point prediction error which determines whether the pedestrian area of the monitored video image can locate accurately. The pixel level logic regression loss function used is shown below. Assuming that the number of pedestrian areas in the original monitoring image is represented by M_{pe}, and the coordinates of the center point predicted in the thermal diagram are represented by (a, b). The super parameters used to control the weight of indistinguishable samples are represented by β and γ. The penalty around the truth value point is reduced by $(1-B_{ab})^{\gamma}$ and the Gaussian kernel is represented by B_{ab}, then there are:

$$DP_d = -\frac{1}{M_{pe}}\sum_{a=1}^{Q}\sum_{b=1}^{F}\begin{cases}\left(1-\hat{B}_{ab}\right)^{\beta}\log\left(\hat{B}_{ab}\right), B_{ab}=1 \\ \left(1-B_{ab}\right)^{\gamma}\left(\hat{B}_{ab}\right)^{\alpha} log\left(1-\hat{B}_{ab}\right), Otherwise\end{cases} \quad (9)$$

B_{ab} is dispersed on the thermal diagram with the true value point as the center. Assuming that the standard deviation adapting to the pedestrian target area size is represented by φ_o, the formula of B_{ab} distribution is provided in the following:

$$B_{ab} = exp\left(-\frac{(a-\delta\hat{\varphi})^2+(b-\delta\hat{\varphi})^2}{2\phi_o^2}\right) \quad (10)$$

The offset prediction error adopts L1 loss function, which is as follows:

$$DP_{OF} = \frac{1}{M}\sum_{l=1}^{M}\left|\hat{a}_l - \left(\frac{a_l}{S} - \left\lfloor\frac{a_l}{S}\right\rfloor\right)\right| + \left|\hat{b}_l - \left(\frac{b_l}{S} - \left\lfloor\frac{b_l}{S}\right\rfloor\right)\right| \quad (11)$$

It is assumed that S is related to the order of each hourglass block, that is, the number of down sampling operations in the process of feature extraction. Equation 9, Equation 10 and Equation 11 are put together, assuming that the weight coefficients are represented by μ1 and μ2 respectively, and the overall loss expression of the model is given by Equation 12:

$$FH = DP_c + \mu_1 DP_{OF} + \mu_2 DP_{SI} \quad (12)$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify that the constructed model achieves an ideal effect in multi-scale feature fusion of pedestrian shape, action and attitude and identity recognition, and improve the robustness of the model on different pedestrian identity recognition, this paper designs a relevant comparative experiment to analyze the performance of the constructed model on different surveillance video image data sets. In order to verify that the constructed model can solve the problem of imbalance between positive and negative samples, *ResNet* and *DenseNet* are selected as the baseline networks for some reference models. The identification performance evaluation indexes of the model include the average accuracy when IoU threshold is 0.6 and 0.8, the recognition recall rate, the recognition speed, and the sensitivity under different false positive numbers.

The comparison of test results on the surveillance video image data set is given in Table 1. It can be seen from the table that the identification performance of the model built in this paper exceeds that of other reference models. The main reason is that the model built in this paper is different from the traditional multi-scale feature fusion algorithm. It does not just predict a series of key points of pedestrian targets, but uses a network structure to extract multi-scale information to achieve the full use of multi-scale information of surveillance video images.

The comparison of test results on the surveillance video image data set is given in Table 1. It can be seen from the table that the identification performance of the model built in this paper exceeds that of other reference models. The main reason is that the model built in this paper is different from the traditional model which adopts multi-scale feature fusion algorithm. It does not just predict a series of key points of pedestrian targets, but uses a network structure to extract multi-scale information to achieve the full use of multi-scale information of surveillance video images.

This paper also compares different pedestrian area detection sensitivity of different sample images, and the comparison results are given in Table 2. It can be seen from the table that the identification sensitivity of almost all detectors on sample 3, sample 4, sample 6 and sample 9 is higher than that on the other four samples, and it is possibly because the difference between the monitoring background of these samples and the texture features of the pedestrian area is more prominent. In addition, compared with other reference models, the constructed model achieves the highest pedestrian identification sensitivity in almost all samples. This paper selects three most representative monitoring places, corresponding to the monitoring set at the entrance 1, 2 and 3 of a mall respectively, and the identification results of pedestrians passing by these places are analyzed in detail.

Through the analysis of pedestrian identification results in the above three most representative monitoring areas, it is found that the constructed model can effectively reduce the false positive generation and obtain accurate prediction of pedestrian target areas, which further verifies that the model constructed in this paper has certain performance advantages in the robustness of different pedestrian identification. In addition, four reference models and the proficiency rate-recall curve (P-R curve) of the model in this paper are drawn, as shown in Figure 5. It can be seen from the figure that the recognition effect of this model is much better than that of other models. When the recall rate is less than 0.5, the recognition accuracy of the model for pedestrian identity is more than 90%. However, the accuracy rate will slowly decline with the increase of recall rate, indicating that there are a few false positive identification results in the constructed model.

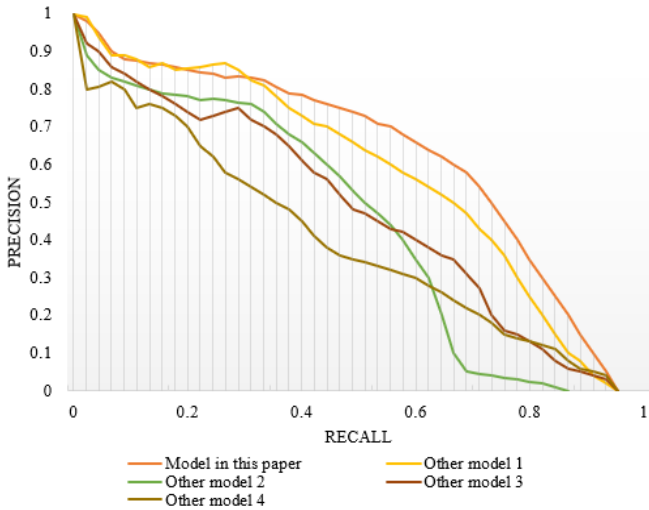

Figure 5. P-R curves for different models

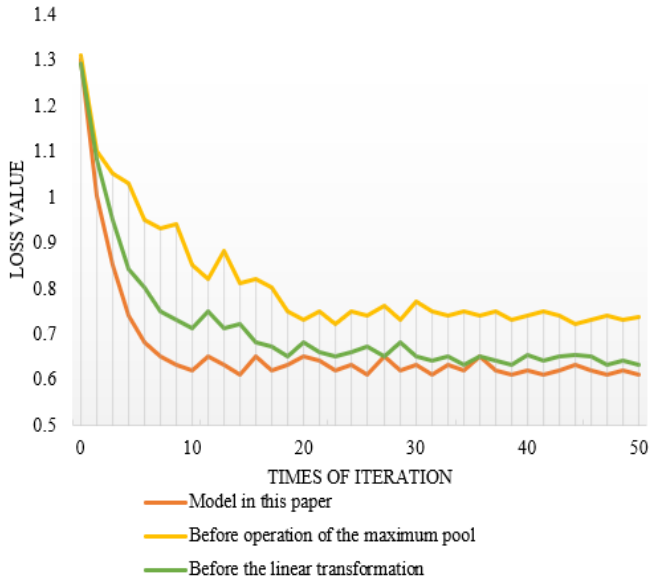

Figure 6. Training loss curves of different models

Table1. Comparison of test results of different models

Model	Reference model 1	Reference model 2	Reference model 3	Reference model 4	Model in this paper
Feature extraction network	*ResNet*-101-*FPN*	*DenseNet*-104 (52×2)	*DenseNet*-54 (18×3)	*DenseNet*-104 (52×2)	*DenseNet*-60 (16+20+24)
0.6 mean accuracy	51.6	59.1	55.8	62.4	66.9
0.8 mean accuracy	36.5	37.8	39.5	31.2	46.1
recall rate	52.7	58.6	59.1	53.7	69.5
speed	163	247	169	115	183
sensitivity1	58.2	67.4	69.5	74.8	79.1
sensitivity 2	69.5	61.8	74.1	85.6	81.9
sensitivity 4	64.7	68.2	76.8	82.4	84.7

Table 2. Comparison of recognition sensitivity of different models

Sample No.	Reference model 1	Reference model 2	Reference model 3	Reference model 4	Model in this paper
1	55.6	59.1	57.4	62.8	71.3
2	59.1	65.8	66.1	85.2	87.1
3	69.3	72.4	79.8	93.5	98.7
4	75.8	71.6	85.1	83.7	95.8
5	62.2	68.7	75.8	88.4	81.6
6	67.8	75.1	78.6	85.9	91.3
7	68.5	69.5	61.8	75.1	88.9
8	61.8	63.1	74.7	85.2	96.4

In order to measure the scientificity of the model design idea, this paper draws the loss curve of the model before introducing the maximum pool operation and linear transformation operation, so as to measure the training performance of different models. The training loss comparison results are shown in Figure 6. It can be seen from the figure that the training process of the model in this paper has gone through three stages of reduction, specifically including the first stage corresponding to the first 10 iterations with a fast decline speed, the second stage corresponding to the 10_{th}-25_{th} iterations with a decline in vibration, and the third stage with a stable number of iterations greater than 25. When the number of iterations of the model is greater than 25, the fluctuation of the curve tends to be stable, indicating that the training of the model in this paper is close to the optimal state. For the model before the introduction of maximum pooling operation and linear transformation operation, although it can also reach the stable state faster, the prediction error of center point, offset prediction error and boundary box size prediction error are significantly larger.

6. CONCLUSION

In this paper, a method of pedestrian identification in surveillance video images based on multi-scale feature learning is studied. First, the deep residual network ResNet and densely connected convolutional network DenseNet are introduced as the baseline network. The model is built based on the hybrid hourglass network module, the pyramid fusion network module with enhanced weighted features and the post-processing module. A loss function that is unified with other traditional models is designed. The optimization objective of the set loss function is divided into three parts: the prediction error of the corresponding center point, the prediction error of the offset and the prediction error of the boundary box size.

Relevant comparative experiments are designed to analyze the performance of the models built on different surveillance video image datasets, and the test results of different models are compared. It is verified that the models built achieve ideal multi-scale feature fusion of pedestrian shape, action, posture and identity recognition. The detection sensitivity of different pedestrian areas in different sample images is compared, and the comparison results are given. Through the analysis of pedestrian identification results in the above three most representative monitoring areas, it is found that the model can effectively reduce the generation of false positive and obtain accurate prediction of pedestrian target areas, which further verifies that the model constructed in this paper has certain performance advantages in the robustness of different pedestrian identification. The loss curve of the model before the introduction of maximum pool operation and linear transformation operation is drawn to measure the training performance of different models, and verify the scientificity of the model design idea.

ACKNOWLEDGMENT

This paper was supported by Hebei Province University Smart Finance Application Technology R&D Center (Grant No. IFDC2022020).

REFERENCES

[1] Li, Z., Jia, Z., Yang, J., Kasabov, N. (2020). Low illumination video image enhancement. IEEE Photonics Journal, 12(4): 1-13. https://doi.org/10.1109/JPHOT.2020.3010966

[2] Huang, Y.Y., Kuo, T.Y., Chen, H.H. (2020). Selecting representative thumbnail image and video clip from a video via bullet screen. In Companion Proceedings of the Web Conference 2020: 48-49. https://doi.org/10.1145/3366424.3382691

[3] Zou, W., Jin, Z. (2020). Algorithm for motion video based on basketball image. In The International Conference on Cyber Security Intelligence and Analytics, Springer, Cham, 792-799. https://doi.org/10.1007/978-3-030-43306-2_111

[4] Kaur, H., Jindal, N. (2020). Image and video forensics: A critical survey. Wireless Personal Communications, 112(2): 1281-1302. https://doi.org/10.1007/s11277-020-07102-x

[5] Jiang, J., Qin, C.Z., Yu, J., Cheng, C., Liu, J., Huang, J. (2020). Obtaining urban waterlogging depths from video images using synthetic image data. Remote Sensing, 12(6): 1014. https://doi.org/10.3390/rs12061014

[6] Droste, R., Jiao, J., Noble, J.A. (2020). Unified image and video saliency modeling. In European Conference on Computer Vision, Springer, Cham, 12350: 419-435. https://doi.org/10.1007/978-3-030-58558-7_25

[7] Tanzil, Y.T., Gamal, A. (2021). Elements identification for pedestrian comfort. In IOP Conference Series: Earth and Environmental Science, IOP Publishing, 673(1): 012026. https://doi.org/10.1088/1755-1315/673/1/012026

[8] Bala, A., Rani, A., Kumar, S. (2020). An illumination insensitive normalization approach to face recognition using locality sensitive discriminant analysis. Traitement du Signal, 37(3): 451-460. https://doi.org/10.18280/ts.370312

[9] Deng, X., Liao, K., Zheng, Y., Lin, G., Lei, H. (2021). A deep multi-feature distance metric learning method for pedestrian re-identification. Multimedia Tools and Applications, 80(15): 23113-23131. https://doi.org/10.1007/s11042-020-10458-8

[10] Wawage, P., Deshpande, Y. (2022). Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. Acadlore Transactions on AI and Machine Learning, 1(1): 22-29. https://doi.org/10.56578/ataiml010104.

[11] Ke, X., Lin, X., Qin, L. (2021). Lightweight convolutional neural network-based pedestrian detection and re-identification in multiple scenarios. Machine Vision and Applications, 32(2): 1-23. https://doi.org/10.1007/s00138-021-01169-7

[12] Zhang, X., Li, N., Zhang, R., Li, G. (2021). Pedestrian Re-identification method based on bilateral feature extraction network and re-ranking. In 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), pp. 191-197. https://doi.org/10.1109/CAIBDA53561.2021.00047

[13] Zhang, J., Cheng, L., Xin, Z., Chen, F., Wang, H. (2022). Video-based pedestrian re-identification with non-local attention module. In International Conference on Artificial Intelligence and Security, Springer, Cham, pp. 437-447. https://doi.org/10.1007/978-3-031-06767-9_36

[14] Yang, X., Wang, Q., Li, W., Zhou, Z., Li, H. (2022). Unsupervised domain adaptation pedestrian re-identification based on an improved dissimilarity space. Image and Vision Computing, 118: 104354. https://doi.org/10.1016/j.imavis.2021.104354

[15] Bai, X., Jiang, F., Zhao, Q. (2022). A Pedestrian re-identification algorithm based on 3d convolution and non_local block. In Proceedings of the 4th International Symposium on Signal Processing Systems, pp. 42-48. https://doi.org/10.1145/3532342.3532349

[16] Braidotti, L., Bertagna, S., Dodero, M., Piu, M., Marinò, A., Bucci, V. (2022). Identification of measures to contain the outbreaks on passenger ships using pedestrian simulations. Procedia Computer Science, 200: 1565-1574. https://doi.org/10.1016/j.procs.2022.01.357

[17] Moura, R.S., Sanches, S.R., Bugatti, P.H., Saito, P. (2022). Pedestrian traffic lights and crosswalk identification. Multimedia Tools and Applications, 81(12): 16497-16513. https://doi.org/10.1007/s11042-022-12222-6

[18] An, F.P., Liu, J.E. (2022). Pedestrian re-identification algorithm based on visual attention-positive sample generation network deep learning model. Information Fusion, 86: 136-145. https://doi.org/10.1016/j.inffus.2022.07.002

[19] Pan, Y., Liang, M., Liu, Z., Li, J., Zhu, J. (2022). Pedestrian re-identification based on multi-stream modal. In International Conference on Algorithms, Microchips and Network Applications, SPIE, 12176: 520-530. https://doi.org/10.1117/12.2636401

[20] Sun, Y.B., Zhang, W.J., Wang, R., Li, C., Zhang, Q. (2022). Pedestrian re-identification method based on channel attention mechanism. Beijing Hangkong Hangtian Daxue Xuebao/Journal of Beijing University of Aeronautics and Astronautics, 48(5): 881-889.

[21] Wong, P.K.Y., Luo, H., Wang, M., Cheng, J.C. (2022). Enriched and discriminative convolutional neural network features for pedestrian re-identification and trajectory modeling. Computer-Aided Civil and Infrastructure Engineering, 37(5): 573-592. https://doi.org/10.1111/mice.12750

[22] Liu, Y.J., Shao, W.B., Sun, X.R. (2022). Learn robust pedestrian representation within minimal modality discrepancy for visible-infrared person re-identification. Journal of Computer Science and Technology, 37(3): 641-651. https://doi.org/10.1007/s11390-022-2146-1

[23] Brehar, R., Marita, T., Negru, M., Nedevschi, S. (2021). Pedestrian identification in infrared and visible images based on pose keypoints matching. In Journal of Physics: Conference Series, IOP Publishing, 1780(1): 12033. 10.1088/1742-6596/1780/1/012033

[24] Zhang, J., Cheng, L., Xin, Z., Chen, F., Wang, H. (2022). Video-based pedestrian re-identification with non-local attention module. In International Conference on Artificial Intelligence and Security, Springer, Cham, pp. 437-447. https://doi.org/10.1007/978-3-031-06767-9_36

[25] Cao, Z., Lee, H.J. (2019). Multi-task network based pedestrian re-identification. In 2019 6th International Conference on Systems and Informatics (ICSAI), pp. 1324-1328. https://doi.org/10.1109/ICSAI48974.2019.9010442

[26] Li, B., Wu, X., Liu, Q., He, X., Yang, F. (2019). Visible infrared cross-modality person re-identification network based on adaptive pedestrian alignment. IEEE Access, 7: 171485-171494. https://doi.org/10.1109/ACCESS.2019.2955930