Hindawi
Scientific Programming
Volume 2021, Article ID 3279957, 10 pages
https://doi.org/10.1155/2021/3279957

Research Article

Pedestrian Behavior Recognition Based on Improved Dual-stream Network with Differential Feature in Surveillance Video

Yonghong Tan, Xuebin Zhou, Aiwu Chen, and Songqing Zhou

School of Intelligent Manufacturing, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China

Correspondence should be addressed to Yonghong Tan; tyh2977@huse.edu.cn

Received 16 July 2021; Accepted 11 August 2021; Published 25 August 2021

Academic Editor: Yi-Zhang Jiang

In order to improve the pedestrian behavior recognition accuracy of video sequences in complex background, an improved spatial-temporal two-stream network is proposed in this paper. Firstly, the deep differential network is used to replace the temporal-stream network so as to improve the representation ability and extraction efficiency of spatiotemporal features. Then, the improved Softmax loss function based on decision-making level feature fusion mechanism is used to train the model, which can retain the spatiotemporal characteristics of images between different network frames to a greater extent and reflect the action category of pedestrians more realistically. Simulation results show that the proposed improved network achieves 87% recognition accuracy on the self-built infrared dataset, and the computational efficiency is improved by 15.1%.

1. Introduction

Pedestrian action recognition is an important research direction in the field of computer vision, which has important research significance and application value in the military and civil fields such as video surveillance, intelligent transportation, motion analysis, navigation, and guidance [1–4]. Due to the poor image quality of low-cost cameras and the lack of stable and obvious feature information, the difficulty of pedestrian detection and behavior recognition is increased [5].

In order to improve the effect of pedestrian action recognition, scholars at home and abroad have also proposed many action recognition algorithms [6–12]. Li et al. proposed an action recognition method based on sparse coding and spatial pyramid feature extraction [6]. Fernando et al. proposed action recognition based on dual-tree complex wavelet transform, where support vector machine (SVM) was adopted to classify and recognize the samples' wavelet entropy [7]. In order to make full use of the complementary features in different modes, Varol et al. proposed an action recognition model based on multimodal feature fusion, which improved the recognition performance of low contrast object [8]. With the development of hardware technology in recent years, deep learning has been widely used in the field of image processing [9–12]. At present, pedestrian detection and action recognition algorithms are usually based on deep learning network, mainly using three-dimensional convolution network, long-term and short-term memory network (LSTM), and dual-flow network to learn high-dimensional spatiotemporal features and automatically classify and recognize [10].

Kuehne et al. [11] designed an action recognition algorithm based on convolutional neural network to meet the needs of assisted driving. Ioffe and Szegedy [12] proposed an action recognition network based on a multilevel segmentation model, which improved the detection accuracy of pedestrian actions in complex backgrounds by extracting deep features from suspected regions. Pedestrian action recognition based on deep networks mostly analyzes the detected pedestrians to realize the recognition of different simple actions, such as standing, walking, squatting, and running. However, the human body action is a sequence action, and only the introduction of temporal-domain features can help improve the accuracy of recognition. Wang et al. [13] extended the original two-dimensional convolution kernel to a three-dimensional convolution kernel and proposed an abnormal behavior model based on three-dimensional convolution, but such methods have complex parameter settings and a huge amount of parameters. LSTM uses convolutional networks to extract pedestrian features frame by frame and makes full use of the spatial-temporal characteristics of pedestrians to improve the characterization ability of behavioral actions. However, its multiscale and high-dimensional processing mode restricts the network operation speed.

As we all know, the visual cortex is mainly responsible for processing visual information in the cerebral cortex. It has two information output channels, dorsal stream and ventral stream, which correspond to spatial pathways and content pathways, respectively [14]. Inspired by this, Simonyan and Zisserman [15] creatively proposed action recognition based on dual-stream convolutional networks. The dual-stream convolutional neural network is a model that combines the spatial information network and the temporal information network. It uses the optical flow as the network input to compensate for the temporal-dimension information that the spatial network cannot capture and merges the results obtained from different models. The recognition accuracy of pedestrian behavior is improved, but the extraction process of optical flow takes a long time, which does not meet the real-time requirements of engineering development.

In order to improve the accuracy and efficiency of pedestrian detection and action recognition in video sequences under complex backgrounds, this paper proposes a fast and effective action recognition model on the basis of dual-stream convolutional networks. A comparative experiment was carried out to verify the practicability and effectiveness of the proposed algorithm. The contribution of this work can be summarized as follows:

(1) To improve the pedestrian behavior recognition accuracy of video sequences in complex background, an improved spatial-temporal two-stream network is proposed in this study

(2) The deep differential network is used to replace the temporal-stream network so as to improve the representation ability and extraction efficiency of spatiotemporal features

(3) The improved Softmax loss function based on decision-making level feature fusion mechanism is used to train the model, which can retain the spatiotemporal characteristics of images between different network frames to a greater extent and reflect the action category of pedestrians more realistically

2. Related Works

2.1. Dual-Stream Convolutional Networks. The dual-stream network structure is composed of two independent spatial stream networks and temporal-stream networks, which are used to learn the spatial position information between video frames and the temporal motion characteristics in optical-flow data, as shown in Figure 1. The two networks have the same structure; each structure is composed of 3 pooling layers and 3 convolutional layers, and a nonlinear layer is added after each convolutional layer. Although the two independent networks have the same structure, they play different roles in the dual-stream network. The input of the spatial stream network is the original image sequence, while the input of the temporal-stream network is the optical flow between adjacent data. In order to better characterize the temporal and spatial characteristics of video sequences, the dual-stream network structure designs two fusion layers, whose purpose is to fuse spatial and motion characteristics at spatial positions so that the channel responses at the same pixel positions are consistent.

The output result of the dual-stream network structure adopts cascade fusion. It is assumed that the output features of the two networks are represented as $x^A \in R^{H\times W\times D}$ and $x^B \in R^{H\times W\times D}$, respectively, where H, D, and W are the height, number of channels, and width of the feature map. The fusion operation stacks the two feature maps on the same spatial position (i, j) of the entire feature channel d, so $y_{i,j,d} = x^A_{i,j,d}$ and $y_{i,j,2\ d} = x^B_{i,j,d}$ can be obtained. Dual-stream network greatly improves the accuracy of behavior recognition, but it also has certain limitations. The temporal characteristic of the dual-stream network exists in the optical-flow map between adjacent frames, and the utilization information of the time dimension is limited, and the optical flow calculation complexity is relatively high. The dual-stream network cannot model the pixel-level relationship of spatiotemporal features.

2.2. Long Short-Term Memory (LSTM). LSTM network is a special recurrent neural network, which can solve the problem of long-term dependence and gradient vanishing problem of recurrent neural network [16]. The LSTM network adopts the "gate structure" to transfer the information processed at the current moment to the next moment, which can fully mine the effective information contained in the massive data. Its network structure is shown in Figure 2. The so-called gate structure refers to a sigmoid network layer (σ) and a bitwise multiplication operation.

As we all know, historical data that have occurred help to predict the probability of occurrence of the next moment. Most current recursive networks use the state of the last frame for feature representation, which obviously loses most of the dynamic information. Compared with extracting feature information from local frames, the overall features of the entire sequence can better present a global representation.

3. An Improved Dual-Stream Network for Behavior Recognition

3.1. Spatiotemporal Feature Propagation. As we all know, the dual-stream network structure is composed of two independent spatial stream networks and temporal-stream networks, respectively, inputting image sequences and optical-flow map. The calculation of optical flow is complicated and requires a lot of hardware resources, resulting in poor real-time engineering application. When the camera is fixed, the optical flow in the foreground is zero. In other words, the

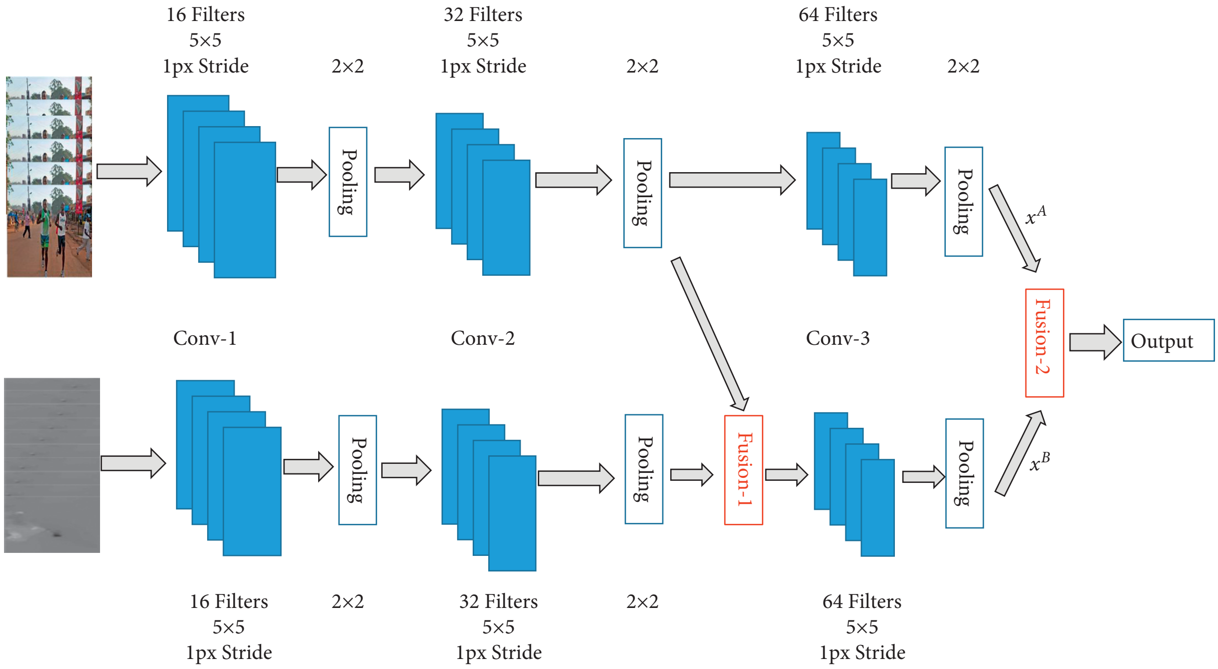

Figure 1: Two-stream network.

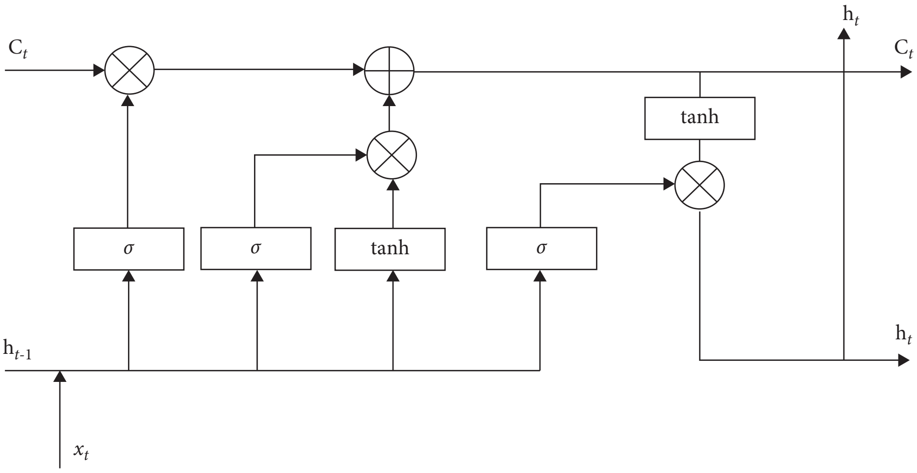

Figure 2: LSTM structure.

difference between images is similar to the result of optical flow. Therefore, this paper proposes a dual-stream network model based on deep differential, which uses deep differential network instead of temporal stream to obtain the interframe relationship and temporal relationship in the sequence. Deep differential is a network structure based on deep feature propagation [12]. The key frame-based feature propagation differential map can be used to replace the optical flow map as the input of the temporal-stream network, which can reduce the computational complexity and enhance the posture expression and category recognition capabilities of the feature propagation map for human actions. The improved dual-stream network structure is shown in Figure 3, where the size of the convolution kernel is 7×7 and 3×3, respectively.

The adjacent sequences' output by the camera has a high degree of similarity, and the optical flow characteristics obtained by it are very weak. In other words, the optical flow characteristics obtained by a large number of frame-by-frame optical flow calculations are not obvious. The differential key frame proposed in [17] can quickly obtain the difference between images and improve the performance of image compression. Differential key frames include the temporal relationship between adjacent frames in the video and have similar performance to optical flow map, but they have the advantages of fast generation speed and small computational operations. Due to the problems of large redundancy and high complexity between frames, our improved dual-stream network for behavior recognition first extracts frames

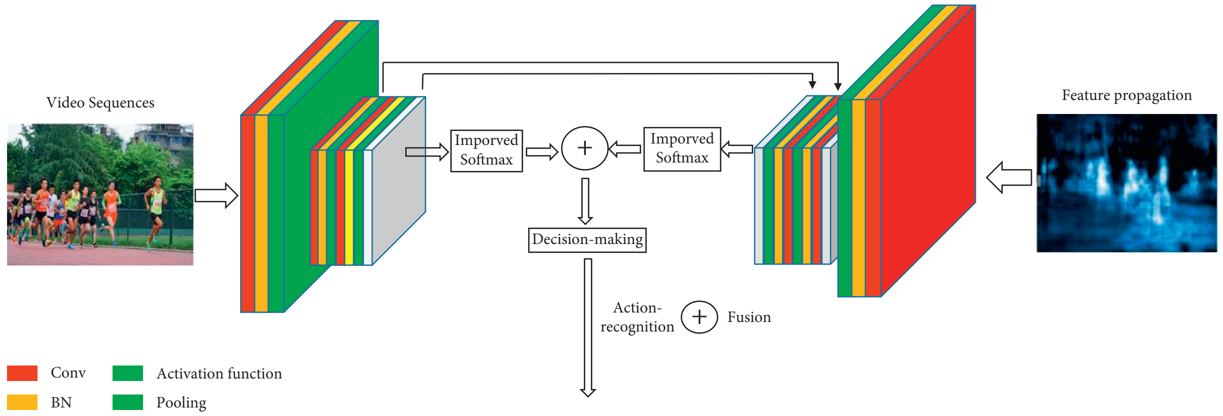

Figure 3: Improved two-stream network.

based on multiple time scales and uses differential feature propagation to obtain key frames of the sequence.

Assuming that an input video composed of a sequence of t frames is recorded as $\mathbf{X}$, each segment is first divided into $\mathbf{T}$ parts with the same time, and then, the key frame x_i is extracted from each part, so the entire video is recorded as $\mathbf{X} = \{x_1, x_2, \ldots, x_t\}$. These key frames can be converted into differential key frames by calculating the difference between adjacent frames and is denoted as $\mathbf{Y} = \{y_1, y_2, \ldots, y_t\}$; finally, the key frame and the differential key frame are input to the differential convolution network, respectively, to obtain the corresponding high-dimensional spatiotemporal feature vector $\{S_1, S_2, \ldots, S_t\}$, where $S_t \in R^d$, $i = 1, 2, \ldots, T$, and d is the feature dimension of the key frame.

The dual-stream network structure proposed in this paper can quickly extract the high-dimensional spatiotemporal features of the sequence and obtain the corresponding differential features based on the detected pedestrian area. Each convolution kernel is followed by a pooling operation. The pooling operation includes average pooling and maximum pooling. The adopted calculation formula is shown as

$$P_{i\longrightarrow j} = \frac{\left(S_1 \oplus S_{i+1} \oplus \cdots \oplus S_j\right)}{(j-i+1)}, \tag{1}$$

where $P_{i\longrightarrow j}$ represents the average pooling feature between key frames i to j. After the key frames are processed by convolution pooling and fully connected operation, the final output result of the deep difference network is a d-dimensional feature vector, and finally, the high-dimensional spatiotemporal information of the entire sequence is obtained. Each key frame is formed into a $1 \times 1 \times 1024$-dimensional vector after the global average pooling operation, and then, the final spatiotemporal feature is extracted through the last convolutional layer.

3.2. Decision-Level Fusion Mechanism with Improved Loss Function. The spatiotemporal dual-channel branch of the dual-stream network separately extracts features from the same sequence of different modal images to obtain spatial position information and temporal motion information. These two types of features enhance the characterization ability of pedestrian actions under the action of the fusion module, but the original dual-stream network only uses feature cascade for fusion. In addition, due to the complexity of the human action recognition problem in video sequential, its performance is often susceptible to interference from environmental noise and ultimately make wrong decisions and affect the output of the entire model. In order to improve the accuracy of the recognition model, this paper proposes a decision-level fusion mechanism. The fusion mechanism draws on the memory characteristics of the LSTM network, by modeling the previous output data and using a coupling mechanism to associate information in different dimensions. It has characteristic invariance in high-dimensional space.

A strong classifier based on improved Softmax logistic regression is designed in paper. The fused features are used to classify the pedestrian action and can get the highest classification probability, which can more effectively improve the accuracy of action recognition. In [15], assuming that the currently given sample sequences $x^{(i)}$ have k categories, the output is $y^{\{i\}} \in \{1, 2, \ldots, k\}$, where its training set is $\{(x^{(i)}, y^{(i)})\}, i \in \{1, 2, \ldots, k\}$. For a given sample feature x, the estimated conditional probability of the category j is $p(y = j|x)$, and the probability equation can be expressed as the following equation:

$$p\left(y^{(i)} = j \mid x^{(i)}; \theta\right) = \frac{e^{\theta_j^T x^{x(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{x(i)}}}. \tag{2}$$

Therefore, the classification probability of each category in Softmax logistic regression is written as follows:

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1 \mid x^{(i)}; \theta\right) \\ p\left(y^{(i)} = 2 \mid x^{(i)}; \theta\right) \\ \vdots \\ p\left(y^{(i)} = k \mid x^{(i)}; \theta\right) \end{bmatrix}. \tag{3}$$

Since the probability of each category satisfies the exponential family distribution [18], if the recognition probability $h_\theta(x^{(i)})$ obtained by equation (3) is expanded in series, we can obtain

$$h_{\theta}\left(x^{(i)}\right) = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(j)}}} \begin{bmatrix} e^{\theta_1^T x^{(j)}} \\ e^{\theta_2^T x^{(j)}} \\ \vdots \\ e^{\theta_k^T x^{(j)}} \end{bmatrix}. \tag{4}$$

The model parameter θ is a matrix with k rows, where each row represents the parameters of the corresponding category, so the model parameter matrix θ can be written as $\theta = [\theta_1^T, \theta_2^T, \ldots, \theta_k^T]$. $1/\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}$ in equation (4) is the normalization operation on the probability distribution, so as to quantify the output probability. By deriving the log-likelihood of the overall sample results, the loss function is rewritten as follows:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} I\left(y^{(i)} = j\right) \log \frac{e_j^T x^{(l)}}{\sum_{l=1}^{k} e_j^T x^{(l)}}, \tag{5}$$

where $I(y^{(i)} = j)$ is an indicative function. The value is 1 when there is a positive example, otherwise it is 0. In order to minimize the model parameter matrix θ, the obtained probability value by substituting equation (4) into equation (5) is expressed as

$$\begin{aligned} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} y^{(i)} \log\, h_{\theta}\left(x^{(i)}\right) \\ +\left(1 - y^{(i)}\right) \log\left(1 - \log\, h_{\theta}\left(x^{(i)}\right)\right). \end{aligned} \tag{6}$$

In order to minimize the loss function shown in equation (6), the gradient descent method is generally used for optimization, and the partial derivative of the loss function is calculated as follows:

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} I\left(y^{(i)} = j\right) - p\left(y^{(i)} = j \mid y^{(i)}; \theta\right)\right). \tag{7}$$

The first l partial derivative $\partial J(\theta)/\partial \theta_{jl}$ of the probability vector $\nabla_{\theta_j} J(\theta)$ indicates that the loss function takes the partial derivative of l parameters of category j. The gradient descent iteration of equation (8) is updated to determine the minimized loss function. The iterative operation includes the following equation:

$$\theta_j = \theta_j - \alpha \nabla_{\theta_j} J(\theta). \tag{8}$$

However, the update strategy of equation (8) used in Softmax logistic regression will affect the update effect of the parameters. Therefore, the multiobjective classification network, proposed in [16], is used to optimize the model, and the probability of equation (2) can be rewritten as $e^{(\theta_j - \varphi)^T x^{(i)}} / \sum_{j=1}^{k} e^{(\theta_j - \varphi)^T x^{(i)}}$, and the equation is expanded, and we get $p(y^{(i)} = j | x^{(i)}; \theta) = e^{\theta_j^T x^{(i)}} / \sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}$. That is to say, when all the hyperparameters θ are subtracted from $e^{\varphi^T x^{(i)}}$, the probability value of the loss function does not change, which shows that when Softmax classifies different samples, the result is not affected by the initial value, but this may cause the optimal solution to be nonunique. In order to solve this problem, this paper introduces a regular weight attenuation term λ in the loss function, constrains its optimal solution, and speeds up the convergence process. Therefore, the improved loss function in this paper can be rewritten as the following equation:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} I\left(y^{(i)} = j\right) \log \frac{e_j^T x^{(l)}}{\sum_{j=1}^{k} e_j^T x^{(l)}} + \frac{\lambda}{2} \sum_{j=1}^{k} \sum_{j=0}^{n} \theta_{ij}^2. \tag{9}$$

When λ is greater than 0, the partial derivative of equation (9) is denoted as follows:

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[x^{(i)} I\left(y^{(i)} = j\right) - p\left(y^{(i)} = j \mid y^{(i)}; \theta\right)\right] + \lambda \theta_j. \tag{10}$$

In order to solve the above improved Softmax logistic regression equation and obtain the classification probability w_i of pedestrian actions, this paper constructs a decision-level fusion mechanism, which makes decisions on category probabilities w_i^c and w_i^d under different samples. For different action categories, the recognition probability p_k of the input image can be obtained by using the principle of multiplication.

$$p_k = \frac{p_{k,w^d} \times p_{k,w^c}}{\sum_{k=1}^{k} p_{k,w^d} \times p_{k,w^c}}. \tag{11}$$

Obtain the recognition probability of multiple images through equation (11), and find the maximum value from it as the final recognition probability u of the current image sequence:

$$u = \arg\max_{k,i} p_{k,i} \quad k = 1, 2, \ldots, K, \tag{12}$$

where i is the number of videos included in each type of action and k is the total number of action types.

This paper proposes a cost function based on the decision-level feature fusion mechanism, which can retain the spatial and temporal information of images between different network on a larger scale, and adopts the principle of majority voting to increase the recognition probability of action categories under different key frame sequences, thereby improving the performance of human motion recognition.

4. Experimental Results and Analysis

4.1. Action Dataset. The scene image detected by the low-cost surveillance camera has no obvious texture details, and it is difficult to obtain fine behaviors such as playing ball and smoking through the image. For pedestrian detection and motion recognition tasks, most of the existing models use three datasets: OTCBVS, KAIST, and FLIR for pedestrian detection, but it is difficult to analyze pedestrian behavior. The main blame is that these images are not continuous sequences, and their movement time span is large, which makes it difficult to conduct correlation analysis. The InfAR dataset [18] is currently a benchmark dataset publicly available in the field of behavior recognition, including 12 daily behaviors such as walking, fighting, clapping, shaking hands, jogging, and hugging. Each behavior type has 50 video sequences. It is completed by a single person or multiple people interactively, but the amount of data is limited. Most algorithms perform transfer learning on visible datasets to improve the recognition effect of behavior in sequences.

The model proposed in this paper mainly analyzes the behavior characteristics of a single pedestrian in the surveillance area. Therefore, the project team collected a large number of pedestrian motion videos to help improve the performance of the model. In order to facilitate performance comparison, this paper also established a self-built dataset, and these pedestrians and their actions have been annotated in the image, including standing, squatting, lying, running, and other action categories. The number of all categories is relatively balanced with a total of 3115 video clips. Table 1 shows the number of sequences in different categories. The first 12 categories are single-person actions, and the last four categories are multiperson interactive actions.

4.2. Parameter Setting. The resolution of all the images in this paper is 6400×512, and the performance of the proposed model is verified by 5-fold cross validation. The networks selected in this paper are all based on TensorFlow framework. The random gradient descent method is used to learn the network parameters. The batch size is 128, and the momentum value and weight attenuation are set to 0.9 and 0.0005, respectively. The initial value of learning rate is 0.01. The learning rate remains unchanged at 0.001 during the first 50 rounds in process of training, and then, the learning rate will be reduced by 10% every 10 rounds so as to prevent overfitting. The value range of λ is $\{1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3\}$.

In order to objectively analyze the effectiveness, the precision rate (PR), miss rate (MR), and recall rate (RR) are selected to quantify the detection performance. All the indicators can be calculated by true positive (TP), false positive (FP), false negative (FN), and true negative (TN). In addition, the confusion matrix is also used to analyze the effectiveness.

4.3. Ablative Analysis. In this paper, an improved pedestrian action recognition model based on deep differential dual-stream network is proposed. In this model, deep differential network is used instead of temporal network to obtain the difference map of feature propagation based on the key frame, which can reduce the computational complexity and enhance the representation ability of feature propagation map of human motion. At the same time, the strong classifier based on improved Softmax logistic regression is used to make pedestrian behavior category decision to improve the ability of category recognition. In order to analyze the effects of different improvement measures, an ablation analysis will be performed in this section. Table 2 shows the recognition effect of deep differential network (DDN), improved Softmax (IS), and decision fusion (DF) on pedestrian motion sequence. The box ☑ indicates the substituted modules in the benchmark network.

Table 1: Categories and number of datasets.

No.	Categories	Total
1	Walking	152
2	Standing	203
3	Climbing	186
4	Jogging	265
5	Jumping	174
5	Punching	128
7	Lying	295
8	Waving 1	168
9	Waving 2	177
10	Crouching	312
11	Sitting	268
12	Handclapping	208
13	Pushing	158
14	Fighting	119
15	Handshaking	134
16	Hugging	168

Table 2: Performance analysis of different modules.

DDN	IS	DF	Pr (%)	FPS
			77.12	13.9
✓			77.83	18.1
	✓		79.91	13.8
		✓	79.78	12.7
✓	✓		81.79	17.8
✓		✓	82.09	18.5
	✓	✓	81.83	11.6
✓	✓	✓	83.01	17.7

The first line in Table 2 does not replace any modules, which is the original dual-stream network. The recognition accuracy and frame rate are 78.12% and 13.9%, respectively. In ablation analysis, if different modules are replaced, their performance will be changed accordingly. Especially, the feature propagation differential map can replace the complex optical-flow calculation, which can greatly improve the processing efficiency. The frame rate is increased by 17.1%, and the accuracy is also improved. The improvement of IS and DF can also improve the performance and increase the recognition accuracy by 3% and 1.1%, respectively. If two modules are replaced in the original dual-stream network at the same time, it can be seen that the performance of any module is better than that of only one module. It is worth noting that the three improved modules designed in this paper are mainly to optimize the efficiency and accuracy. As long as they are replaced by the deep differential network, the final recognition efficiency will be greatly improved. The main reason is that the feature propagation differential map has the advantages of fast generation speed and small computational complexity. Finally, the three modules improve the dual-stream network from different angles and achieve 82.01% recognition accuracy and 17.7 processing frame rate for the video sequences.

4.4. Qualitative and Quantitative Analysis. In order to analyze the performance of the proposed pedestrian behavior recognition algorithm, we select the common behavior recognition algorithms for performance comparison, which are IDT (improved density trajectories) [19], C3D (continuous 3D) [20], SCNN-3G [21] (spatiotemporal continuous neural network based on 3D gradients), L-LSTM [22] (lattice long short-term memory), Ts-3D [23] (two-stream expanded 3D convolutional), and OFGF [24] (optical flow guided feature), where IDT is a very classic traditional algorithm in the field of behavior recognition. By introducing the background optical-flow elimination method and extracting features along the trajectory, the obtained features are more suitable for the description of human motion. C3D is to construct three-dimensional multichannel convolution features for continuous frames, extract multidimensional features through prior knowledge, and enhance the training speed and feature representation ability of back propagation. L-LSTM is a behavior recognition model based on rasterized long-term and short-term memory, which acts on video sequence by convolution and assumes that the motion in the video is stationary in different spatial positions. Ts-3D is a behavior recognition algorithm based on improved dual-stream network, which is extended from 2DCNN Inception-V1, and can use pretraining parameters to enhance the training efficiency. OFGF is a fast and robust motion representation method for video motion recognition. It can obtain the human motion trend by calculating the spatiotemporal gradient. All the comparison algorithms are tested using the author's source code. Because part of the original code is mainly for 3D natural image analysis, the research object of this paper is the 2D gray image. For the consistency of the algorithm model, all input images are expanded into three channel images. In addition, all experiments in this paper use the same test set and training set for comparison.

Because the digital video output of the camera reaches 100 frames, the content of adjacent frames changes very slowly. In order to make the input sequence fully extract temporal information, this paper uses multiscale frame extraction strategy to obtain the input dataset, which ensures that the input sequence can obtain more abundant temporal information on the premise of fixed dimension. Therefore, the data between some key frames is very redundant, and only a small amount of information is needed to represent the human movement trend. That is to say, the trend information independent of the duration can be obtained by using the differential key frame to ensure that the obtained feature information is evenly distributed along the time dimension. It can be seen that the improved strategy in this paper has the similar performance of optical-flow map and can fully represent the temporal action information of human behavior, but the computational complexity is smaller.

Figure 4 shows the differential key frame and the corresponding optical-flow map. It is obvious that most of the background noise in the sequence has been deleted, and the human action has been successfully retained. In addition, the differential information obtained in this paper is similar to the optical-flow information of the original image. This operation can not only reduce the computational complexity but also make the model more robust.

Figures 5 and 6 show the loss value and recognition accuracy in the training process for the proposed model. In the training process, the learning rate of the loss function is updated dynamically with the change of the number of training rounds to prevent the training process from overfitting. It can be seen from the results in Figure 5 that the loss convergence is faster and more stable after using the decision fusion mechanism, while the training loss without fusion is more jittery. In Figure 6, the training accuracy of the fused dual-stream network can rise rapidly and approach 99%, which indicates that the fusion mechanism can effectively fuse the spatial-temporal information, and improve the representation ability of human action through complementary feature information.

It is very difficult to subdivide all the categories because of the diversity of human behavior. This paper mainly verifies the performance of the proposed algorithm. Therefore, 16 kinds of behaviors, such as standing, walking, running, and jumping, were selected for recognition. Yolo-v3 is used for the human detection model. The differential map obtained in this paper is refined on the basis of pedestrian detection results, and the range of processing is reduced, which helps to improve the representation ability of human movement trend. Table 3 shows the quantitative results of all comparison algorithms under the same test set. It can be seen that although the results of IDT in some datasets are not as good as the behavior recognition algorithm based on deep learning, and the performance can be improved by integrating the results of IDT in general, especially in the standing sequence. C3d and Ts-3D (two-stream 3D) are the two mainstream methods for behavior recognition, and their recognition accuracy reaches 75.2%. However, these two methods rely heavily on the variation of adjacent time sequences. Once there are fewer frame sequences between the key frames, their performance will be greatly reduced. For example, large amplitude motion leads to large variation of adjacent frames, and the final recognition accuracy is insufficient. For example, the result of sequence 2 is only 57%. L-LSTM often relies on the last layer feature of convolution network as input, which cannot capture low-level motion features, and it is difficult to train for traversing the whole video. In order to improve the long-time behavior recognition, dense sampling is a common method to improve the performance, but it requires huge computational overhead. The accuracy rate, missed detection rate, and recall rate of OFGF are 73.8%, 19.2%, and 78.4%, respectively. Although this is the best algorithm of the contrast algorithm, it can be embedded in any existing deep network framework with only a small time cost, and its processing frame rate reaches 69.7. The model proposed in this paper inputs it into the deep differential network to extract the time dimension features, which can ensure the accuracy and greatly reduce the operation time, and finally get 78% recognition accuracy. It can be seen that the recognition accuracy of this model is 6.7% higher than that of L-LSTM, and 1.8% higher than that of C3D with 68 layers,

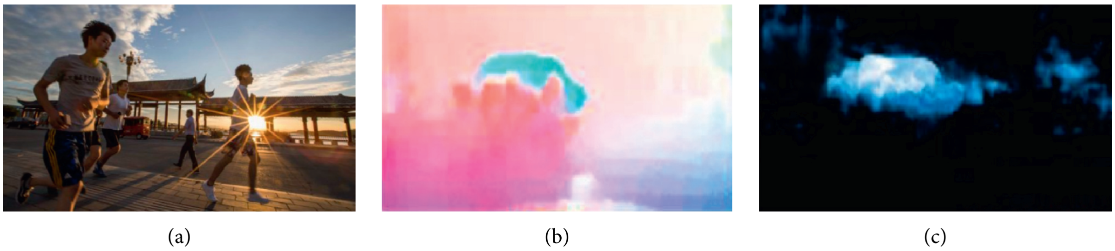

Figure 4: Differential key frame and corresponding optical flow. (a) Key frame; (b) optical flow; (c) differential key frame.

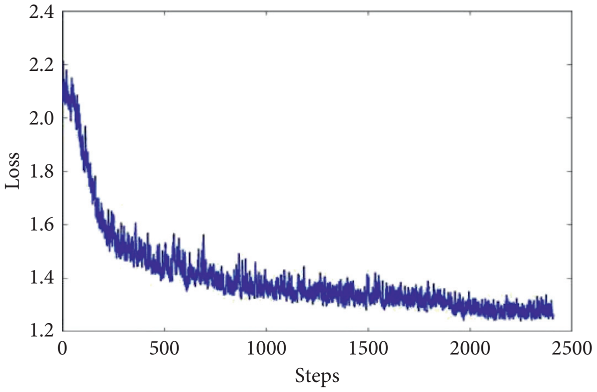

Figure 5: Change trend of loss value during training.

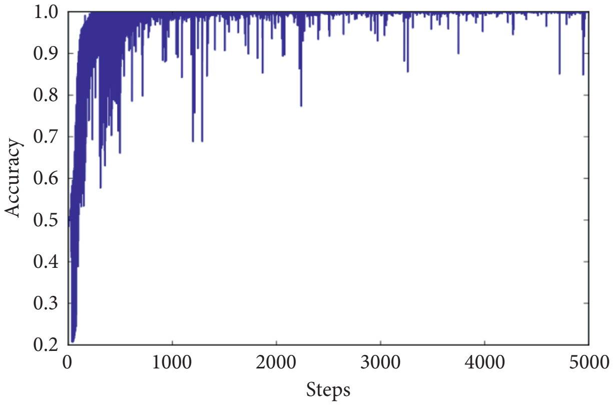

Figure 6: Change trend of precision during training.

which fully shows that the proposed model can recognize human actions more effectively. Figure 7 shows the histogram of the recognition results of comparison algorithms on the mixed data set, so as to intuitively analyze the performance of the proposed model, where the ordinate is the percentage of different index results. From the histogram trend, it can be seen that the proposed model has better results in the quantitative evaluation index precision rate (PR), miss rate (MR), and recall rate (RR).

To analyze the recognition performance of the proposed method for different types of behaviors, we also analyze the confusion matrix before and after the fusion strategy. According to the experimental results, walking and standing can achieve a high recognition rate whether or not using this method. When using the decision fusion method, the recognition rate of hugging, shaking hands, fighting, and other behaviors is slightly low, and there are many wrong points. Using our proposed network can improve the wrong points between these behaviors

Table 3: Performance analysis of different comparison models

Categories	IDT			C3D			SCNN-3G			L-LSTM			Ts-3D			OFGF			Our		
	Pr	Mr	Rr	Pr	Mr	Rr	Pr	Mr	Rr	Pr	Mr	Rr	Pr	Mr	Rr	Pr	Mr	Rr	Pr	Mr	Rr
Walk	65	28	71	67	22	73	69	24	73	75	20	78	77	28	75	**80**	17	**81**	79	**11**	**81**
Stand	73	21	76	77	20	78	77	20	75	83	20	88	85	21	76	85	**17**	86	**86**	21	**87**
Climb	51	37	62	54	32	64	62	35	67	67	26	68	72	37	62	77	25	**82**	**79**	**17**	82
Jog	67	29	71	69	24	76	71	24	71	68	29	77	72	29	71	77	20	79	**87**	**9**	**91**
Jump	61	33	66	62	32	69	68	35	68	61	33	75	73	33	66	**73**	23	78	72	**17**	**81**
Punch	42	51	45	42	41	44	47	52	49	52	41	59	61	51	65	62	31	65	**68**	**23**	**70**
Lying	57	37	61	58	32	67	60	34	66	57	37	68	71	31	68	67	23	70	**68**	**17**	**71**
Wave 1	66	32	66	69	30	69	69	31	69	66	32	77	73	24	76	76	12	81	**83**	**12**	**86**
Wave 2	69	29	70	71	31	72	72	24	77	69	29	88	79	29	80	82	18	87	**89**	**9**	**89**
Crouch	42	30	42	44	35	46	45	24	47	42	30	59	54	21	51	61	23	62	69	27	**72**
Sitting	71	25	79	74	29	81	73	29	80	72	25	82	79	20	82	81	16	**89**	**83**	**15**	88
Handclap	38	34	39	39	35	43	39	31	34	38	34	51	46	24	59	68	23	69	**73**	24	**77**
Push	42	47	45	45	48	47	43	43	48	42	47	58	67	31	65	72	24	75	72	17	**80**
Fight	54	36	58	59	31	59	57	32	59	54	36	68	68	30	68	64	16	78	**81**	**14**	**81**
Handshake	63	30	68	66	32	71	67	27	71	63	30	77	72	21	78	76	20	**88**	**77**	**23**	82
Hug	68	27	70	67	28	73	62	29	75	77	29	75	75	27	79	79	26	80	**82**	**15**	**86**
Mixed	58	32	61	60	31	64	61	30	64	61	31	71	70	28	70	73	19	78	**78**	**16**	**81**

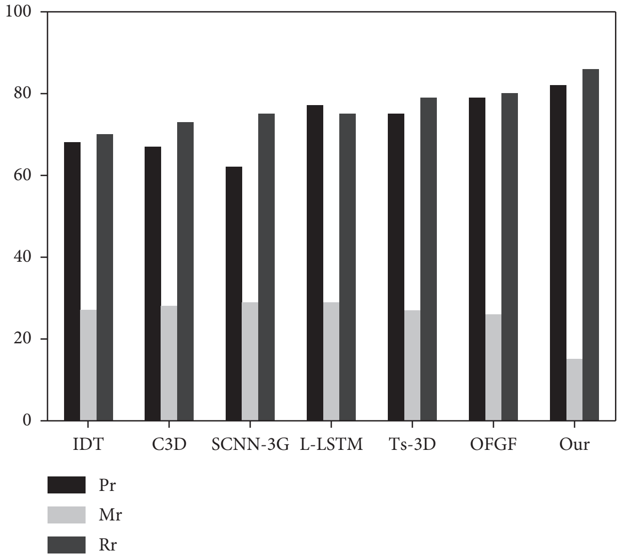

Figure 7: Qualitative analysis of different models under mixed datasets.

and improve the correct recognition rate. The simulation results show that the proposed improved network achieves 87% recognition accuracy on the self-built infrared dataset, and the computational efficiency is improved by 15.1%.

5. Conclusion

Aiming at the problem of low accuracy of pedestrian behavior recognition in image sequences, this paper proposes a pedestrian action recognition model based on improved spatial-temporal dual-stream network. The model uses differential key frames instead of optical flow sequence for temporal feature extraction, which ensures the accuracy and greatly reduces the computational complexity. At the same time, this paper also uses the cost function based on the decision-level feature fusion mechanism to train the model, which can retain the spatio-temporal characteristics of images between different network frames to a greater extent, and reflect the action category of pedestrians more realistically. Simulation experiments also verify the effectiveness of the model from different angles. The next step is to refine the behavior categories, establish a larger and richer training sample set, improve the recognition accuracy and generalization ability of the model, and transplant the model on the basis of AI embedded platform to realize behavior recognition in complex monitoring environment.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This paper was supported by the construct program of applied characteristic discipline in Hunan University of Science and Engineering.

References

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. F. Fei, "Large- scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, Columbus, OH, USA, June 2014.

[2] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotem-poral features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, Santiago, Chile, December 2015.

[3] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2718–2726, Las Vegas, NV, USA, June 2016.

[4] J. C. Niebles, C. W. Chen, and L. F. Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the European Conference on Computer Vision ECCV*, pp. 392–405, Heraklion, Crete, Greece, September 2010.

[5] P. Tumas, A. Nowosielski, and A. Serackis, "Pedestrian detection in severe weather conditions," *IEEE Access*, vol. 8, Article ID 62775, 2020.

[6] W. Li, M. Ding, and L. Zeng, "Pedestrian detection based on objectness and sparse coding in a single infrared image," *Infrared Technology*, vol. 38, no. 9, pp. 752–757, 2016.

[7] B. Fernando, E. Gavves, O. M. Jose, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5378–5387, Boston, MA, USA, June 2015.

[8] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, Article ID 04494, 2018.

[9] J. Donahue, L. H. Anne, and S. Guadarrama, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2634, Boston, MA, USA, June 2015.

[10] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: a dataset of 101 human actions classes from videos in the wild," *CoRR abs*, vol. 1212, 2012.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the 13th International Conference on Computer Vision*, pp. 2556–2563, Barcelona, Spain, November 2011.

[12] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the The 32nd International Conference on Machine Learning (ICML)*, pp. 448–456, Lille, France, July 2015.

[13] L. Wang, Y. Qiao, and X. Tang, "Video action detection with relational dynamic-p," in *Proceedings of the 13th European Conference on Computer Vision ECCV*, pp. 565–580, Zurich, Switzerland, September 2014.

[14] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, "You lead, we exceed: labor-free video concept learning by jointly exploiting web videos and images," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 923–932, Las Vegas, NV, USA, June 2016.

[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 150, pp. 109–125, 2014.

[16] R. Peng, L. Wang, L. Xin, and P. Liu, "Deep convolution neural network based on improved softmax classifier and its application in face recognition," *Journal of Shanghai University (Social Sciences Edition)*, vol. 24, no. 3, pp. 352–366, 2018.

[17] H. Yasin, M. Hussain, and A. Weber, "Keys for action: an efficient keyframe-based approach for 3D action recognition using a deep neural network," *Sensors*, vol. 20, no. 8, p. 2226, 2020.

[18] A. G. Hauptmann, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng et al., "InfAR dataset: infrared action recognition at different times," *Neurcomputing*, vol. 212, pp. 36–47, 2016.

[19] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the 2013 IEEE Lnternational Conference on Computer Vision*, pp. 3551–3558, IEEE, Sydney, Australia, December 2013.

[20] D. Tran, L. Bourdev, and R. Fergus, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the 2015 IEEE, International Conference on Computer Vision*, pp. 4489–4497, IEEE, Boston MA, USA, June 2015.

[21] T. Yang, Z. Chen, and W. Yue, "Spatio-temporal two-stream human action recognition model based on video deep-learning," *Journal of Computer Applications*, vol. 38, no. 3, pp. 895–899, 2018.

[22] L. Su, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Silvio, "Lattice long short-term memory for human action recognition," in *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pp. 2166–2175, IEEE, Venice, Italy, October 2017.

[23] J. Carrlira and A. Gisslrman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of the 2017 IEEE, Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, IEEE, Honolulu, HI, USA, July 2017.

[24] S. Sun, Z. Kuang, L. Sheng et al., "Optical flow guided feature: a fast and robust motion representation for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20118–20132, IEEE, Salt Lake City, Utah, USA, June 2018.