Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 9761842, 1 page
https://doi.org/10.1155/2023/9761842

Retraction

Retracted: Intelligent Transport Surveillance Memory Enhanced Method for Detection of Abnormal Behavior in Video

Journal of Advanced Transportation

Received 8 August 2023; Accepted 8 August 2023; Published 9 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] D. Zhang, "Intelligent Transport Surveillance Memory Enhanced Method for Detection of Abnormal Behavior in Video," *Journal of Advanced Transportation*, vol. 2022, Article ID 5631281, 12 pages, 2022.

Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 5631281, 12 pages
https://doi.org/10.1155/2022/5631281

Research Article

Intelligent Transport Surveillance Memory Enhanced Method for Detection of Abnormal Behavior in Video

Deng-Hui Zhang

College of Information Science, Zhejiang Shuren University, Hangzhou, Zhejiang 310015, China

Correspondence should be addressed to Deng-Hui Zhang; dhzhang@zjsru.edu.cn

Received 7 December 2021; Revised 4 January 2022; Accepted 7 January 2022; Published 2 March 2022

Academic Editor: Muhammad Arif

The purpose is to build a better intelligent transport platform and improve the performance of surveillance video abnormal behavior detection systems under rapid progress of science and technology, to process large-scale traffic surveillance video data. Autoencoder (AE) can detect abnormal behavior by using reconstruction error information. However, it cannot decode some abnormal codes well, so an AE based on memory needs improvement. The objective of this research is to propose a model where abnormal surveillance video can be handled. Therefore, a self-coding method based on memory enhancement is proposed. The steps are as follows: different abnormal behavior detection system algorithms are analyzed at first. The characteristics of three different methods, namely, the original autoencoder (AE), recurrent neural network, and convolutional neural network, are compared. Then, a memory module is proposed to enhance the automatic encoder to reduce the reconstruction error of normal samples and increase the reconstruction error of abnormal samples. The effect image is obtained by Laplace transform and convolution for the image with low definition, and the image with noise is processed by guided filtering. Finally, different methods are used for experimental comparison. Experiments show that, on the dataset Avenue, the frame-level result of the method proposed is about 2% higher than that of the optimal ConvLSTM in the comparison method; on the Ped1 and Ped2 datasets, it is also about 3% higher than ConvLSTM. The comparison of different methods shows that the effect of the method proposed is the best. The self-coding traffic surveillance video abnormal behavior detection system based on memory enhancement is designed with a modular structure and it uses the self-coding method based on memory enhancement. The effectiveness of the proposed method in the real scene is verified by comparing the performance of different methods in the same data set (Xia and Li, 2021).

1. Introduction

Intelligent transport was founded on September 2, 2014. It is a transportation-oriented service system that fully combines modern electronic information technologies such as Internet of things, cloud computing, artificial intelligence, automatic control, and mobile Internet in the transportation field [1]. Through various high and new technologies, it controls and supports all aspects of transportation fields such as traffic management, transportation, and public travel, as well as the whole process of traffic construction management. With the increase of monitoring equipment, the detection of abnormal behavior in surveillance video plays a crucial role in the construction of intelligent transport. The increasing number of vehicles on the road leads to more frequent traffic accidents. The intelligent traffic monitoring system can help deal with traffic accidents and improve the processing efficiency of traffic accidents.

2. Related Work

For a long time, researchers have done a lot of research in video abnormal behavior detection. Chen proposed a general framework for analyzing people based on extracting set features from surveillance video foreground. This method could flexibly integrate different foreground detection technologies to adapt to different monitoring environments. Moreover, the representative features that could be extracted depended on heterogeneous foreground data. Finally, a classification algorithm was applied to these features to automatically model

crowd behavior and distinguish abnormal events from normal patterns [2]. Fan et al. proposed a crowd abnormal behavior detection method based on improved statistical global optical flow entropy. This method could better describe the chaotic degree of the crowd, extract the optical flow field from the video sequence, and obtain the two-dimensional optical flow histogram. Then, combined with information theory and statistical physics, the improved optical flow entropy was calculated from the two-dimensional optical flow histogram [3]. Shen and Wu proposed an abnormal crowd behavior detection algorithm based on image processing, which was mainly to determine the region of interest through the rapid flow of people on the bus; the moving target was extracted by improving the Vi Be algorithm, and the multiscale sliding window algorithm was introduced to determine the recognition area; combined with the continuous multiframe recognition area, the abnormal behavior recognition of the improved convolutional neural network (CNN) algorithm was carried out, and the recognition results were used to judge whether the crowd in the bus was abnormal [4]. Xia and Li proposed an accurate and effective abnormal behavior detection method, introduced a new time attention mechanism to learn the contribution of different historical appearance features at the same location to the current features, so as to solve the representation problem of dynamic motion features. The Long Short-Term Memory (LSTM) network was used to decode the time attention of the historical feature sequence and predict the characteristics of the current time [5]. Zhou et al. proposed a new behavior recognition framework. In this framework, a target depth estimation algorithm was proposed to calculate the three-dimensional spatial position information of the target, and the information was used as the input of the behavior recognition model. Meanwhile, a skeleton behavior recognition model based on spatiotemporal convolution and attention-based LSTM was proposed to obtain more spatiotemporal information and better process long-term video [6]. Harrou et al. proposed an automatic monitoring scheme based on Vision, which was especially used for atypical event detection and location in crowded areas [7].

In video abnormal behavior detection, the deep learning method has made a lot of contributions. For example, recurrent neural network (RNN) [8], convolutional neural networks (CNN) [9], and LSTM network [10] are all applied to video abnormal behavior detection. Autoencoder (AE) can detect abnormal behavior by using reconstruction error information. However, it cannot decode some abnormal codes well, so an AE based on memory enhancement is proposed. The memory retrieval step is added in the process of encoding and decoding, and the memory module is updated by the encoder and decoder simultaneously, which expands the abnormal reconstruction error information and improves the detection performance of the system. Several public datasets are used to verify the detection effect of memory enhanced self-coding.

3. Methodology

The steps of methodology are from Sections 3.1 to 3.6. Figure 1 is framework of an abnormal behavior detection process based on memory enhancement self-coding: Figure 1 is an abnormal behavior detection process based on memory enhancement self-coding: The method flow is as follows. First, the obtained original monitoring image is preprocessed into an image more suitable for analysis. Then, the relevant information is extracted into memory enhancement AE for information reconstruction. The reconstructed image is compared with the original image, and the error obtained is compared with the set threshold to see if there is abnormal behavior. The specific process of feature extraction is standardized space, image gradient, gradient histogram, and feature collection. The specific process of pedestrian behavior detection is training sample set, sample processing, feature extraction, labeling, and training. The trained model can detect pedestrians. Figure 2 reveals that the model can well detect passer-by targets.

3.1. AE Principle. As one of the most crucial methods in deep learning technology, AE is generally used in compressed video, image reconstruction, and video codec. AE consists of two parts, encoding and decoding. The function of AE is to reconstruct the original image data after a series of network layer processing and reconstruct the reconstructed image similar to the original image using the decoded data. The coding process is

$$y = f(Wx + b). \tag{1}$$

AE also needs to cooperate with the activation function in the coding process to change the simple linear structure of the network, so that it has higher learning ability and learns more feature information. The decoding process is

$$x' = f(W'x + b'). \tag{2}$$

AE needs to use loss function to reduce error in the process of encoding and decoding:

$$L = -\log P(x \mid x'). \tag{3}$$

To improve the feature expression of the hidden layer and better represent the structural information of the input signal, researchers propose noise reduction AE. The main process is to add noise to the original input information and then input the noise information into the traditional AE to reconstruct the signal. The process is as follows:

$$\begin{aligned} \widehat{x} &\sim q_D(\widehat{x} \mid x), \\ h_1 &= \sigma_e(W_1 x + b_1), \\ y &= \sigma_d(W_2 h_1 + b_2). \end{aligned} \tag{4}$$

$q_D(\widehat{x} \mid x)$ is noise distribution; W_1 is coding weight; b_1 is coding offset; W_2 is decoding weight; b_2 is decoding offset; $\widehat{x}$ is input of mixed noise information. The loss function is

$$J_{DAE}(W) = \sum E_{x \sim q_D(\widetilde{x}|x)}^{\sim}[L(x, y)]. \tag{5}$$

Noise reduction AE is the use of artificial noise in the input to obtain better feature expression. The advantage is good robustness, and the disadvantage is that the time for adding noise needs to be increased before training, which increases the training time compared with AE [11].

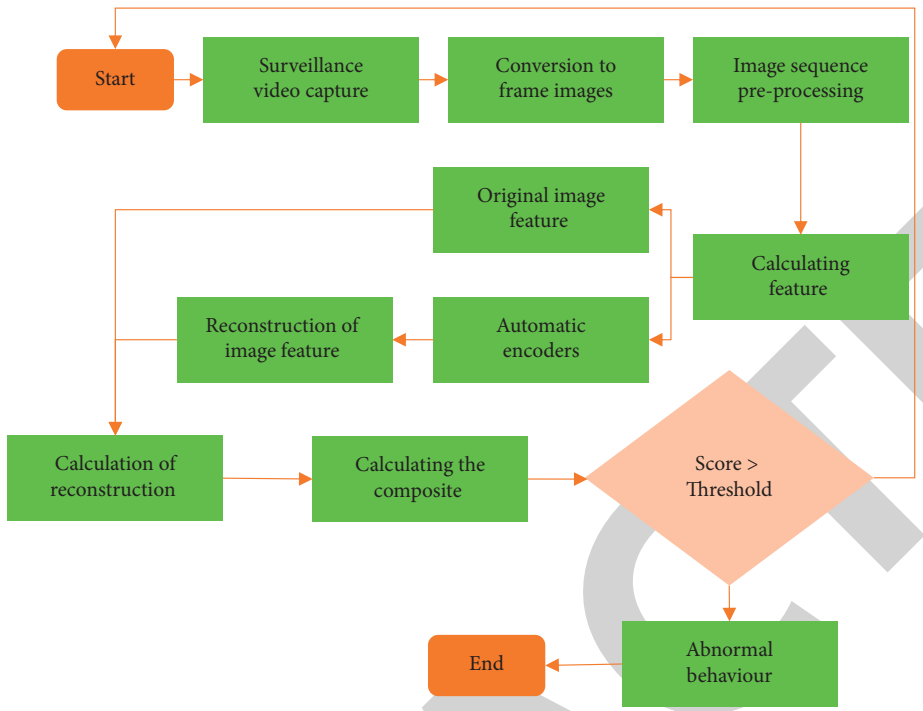

Figure 1: Flowchart of abnormal behavior detection.

Figure 2: Pedestrian detection renderings.

Variational AE adds a generation part to the coding part and decoding part, and the coding part and decoding part are trained simultaneously [12]. The objective function is

$$P_{(x)} = \int P(x \mid z; \theta) P(z) \mathrm{d}z. \tag{6}$$

x is input data; z is implicit variable; $P_{(x)}$ is maximized generation probability; θ is model parameters. For the loss function, the similarity of two variables is evaluated:

$$\mathcal{D}[Q(z \mid x) \| P(z \mid x)] = E_{z \sim Q}[\log Q(z \mid x) - \log P(z \mid x)]. \tag{7}$$

The objective function is obtained by transformation:

$$\begin{aligned}&\log P(x) - D[Q(z \mid x)PP(z \mid x)]\\&= E_{z\sim Q}[\log(x(z \mid x))] - D[Q(z \mid x)PP(z)].\end{aligned} \tag{8}$$

Convolution AE is an optimized multilayer neural network model, which transforms the input expression into a new expression and then decodes it [13]. Training AE equation is

$$\min_{\theta,\gamma} \sum_{x} L_{rec}(x, \widehat{x}), \text{ with } \widehat{x} = \text{dec}_{\gamma}(enc_{\theta}(x)), \tag{9}$$

where enc θ is the encoder, is the decoder, θ and γ are the parameters, and x is the input. Convolution AE is an improved structure using convolution layer and pooling layer on the original AE:

$$\begin{aligned}h^k &= \sigma\left(x * w^k + b^k\right),\\ y &= \sigma\left(h^k * \widehat{w}^k + c\right).\end{aligned} \tag{10}$$

h^k and b^k are convolution kernel parameters; k is number of convolution kernels. The input and output are compared to obtain the complete convolution AE:

$$E = \frac{1}{2n}\sum (x_i - y_i)^2. \tag{11}$$

Convolution AE has the advantage of better image data processing and a better reconstruction effect.

3.2. RNN and LSTM Network. Traditional neural networks are prone to problems when dealing with some sequence data, especially when these sequence data have an up-down relationship, so RNN appears. The advantage of RNN is that it has a memory mechanism, which can fully analyze the relationship among these data when dealing with the problems related to these sequence data with up-down connection, and it is more optimized on the whole [14]. Figure 3 shows the structure of RNN.

x is input of current time h; s is hide node status at present; o is output (RNN processing). The specific equation reads

$$\begin{aligned}s_t &= f\left(U_{x_h} + W_{s_{h-1}} + b\right),\\ o_t &= \text{soft max}\left(V_{s_h} + c\right).\end{aligned} \tag{12}$$

Activate function sigmoid is f; U and W are weight matrix between layers; b and c are offset value. RNN has the advantage of sharing model parameters at different times and can deal with long-term dependence problems. However, its disadvantages are obvious, such as the unstable update of model parameters, the existence of gradient explosion or disappearance, and only short-term memory.

LSTM is an improvement of the RNN model. A "gate" structure is added, which can solve the problem caused by too long distance, even when the length of the data sequence is different [15]. The neurons of the LSTM model are composed of unit state, output gate, input gate, and forget gate. The operation mode of the forget gate: sigmoid function allocates the weighted calculation value of input p_t at current time t and output $n_{t\text{-}1}$ at time $t-1$ and uses the above to control the influence of sequence information of past output on input streams. The equation is

$$g_t = \sigma\left(W_f \cdot [n_{t-1}, p_t]\right) + b_g. \tag{13}$$

The value s of input p_t at time t and output $n_{t\text{-}1}$ at time $t-1$ is weighted by the sigmoid function, expressed as (14). The new state candidate value $\widetilde{A}_t$ of the unit is generated by the nonlinear tanh function. The new unit state A_t can be obtained only by adding the two and then passing through the forget gate and the input gate.

$$\begin{aligned}s_t &= \sigma\left(W_i \cdot [n_{t-1}, p_t]\right) + b_s,\\ \widetilde{A}_t &= \tan n\left(W_c \cdot [n_{t-1}, p_t]\right) + b_A,\\ A_t &= g_t \cdot A_{t-1} + s_t \cdot \widetilde{A}_t.\end{aligned} \tag{14}$$

The output gate outputs q_t value. The sigmoid function needs to be used to weight the input p_t at current time t and output $n_{t\text{-}1}$ at time $t-1$, expressed as (15). Next, the output of the LSTM unit is calculated and controlled by the nonlinear tanh function, and finally the output value n_t is obtained. The advantage of LSTM is to solve the data problem due to long distance.

$$\begin{aligned}q_t &= \sigma\left(W_q \cdot [n_{t-1}, p_t]\right) + b_q,\\ n_t &= q_t \cdot \tan n(A_t).\end{aligned} \tag{15}$$

3.3. CNN. CNN is a feedforward neural network, which has excellent performance for large-scale image processing [16]. Figure 4 shows the structure of the CNN model. The picture is convoluted by the convolution layer first and then pooled. After several convolution and pooling operations, the obtained characteristic information is sent to the fully connected layer and finally sent to the output layer, and the size of the output layer is determined by the task of CNN.

The detail of Figure 5 is here; this figure shows the schematic connection of the CNN model. The size of the feature map after convolution operation is calculated by

$$\text{Size}_{\text{out}} = \frac{(\text{Size}_{\text{in}} - F + 1)}{\text{stride}}. \tag{16}$$

Through filling, the size of the feature map remains unchanged after convolution. The same filling means that the size of the characteristic image remains unchanged after the convolution operation, and its equation is

$$\begin{aligned}\text{Size}_{\text{out}} &= \frac{(\text{Size}_{\text{in}} + 2 \times \text{padding} - F + 1)}{\text{stride}},\\ \text{padding} &= \frac{(F-1)}{2}.\end{aligned} \tag{17}$$

Size_{out} is the output size of the feature map; Size_{in} is the input size of the feature map; F is the size of the convolution kernel; stride is the step size; padding is the number of circles filled around the feature graph. Convolution layer

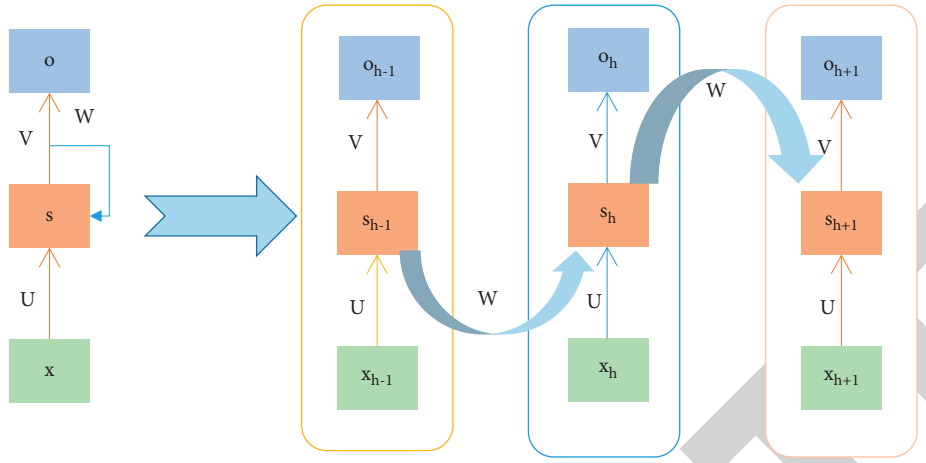

Figure 3: Structure diagram of RNN.

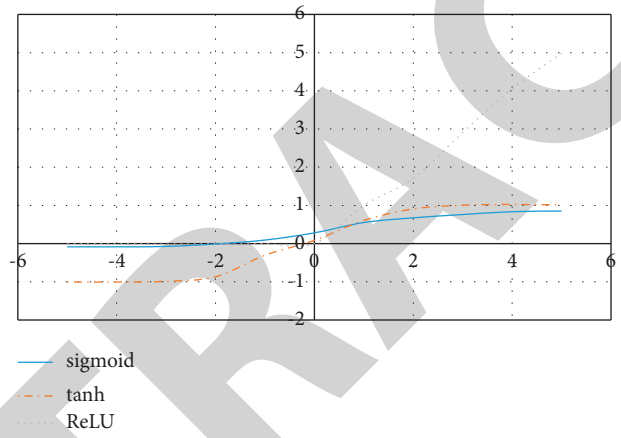

Figure 4: Curves of sigmoid, tanh, and ReLU activation functions.

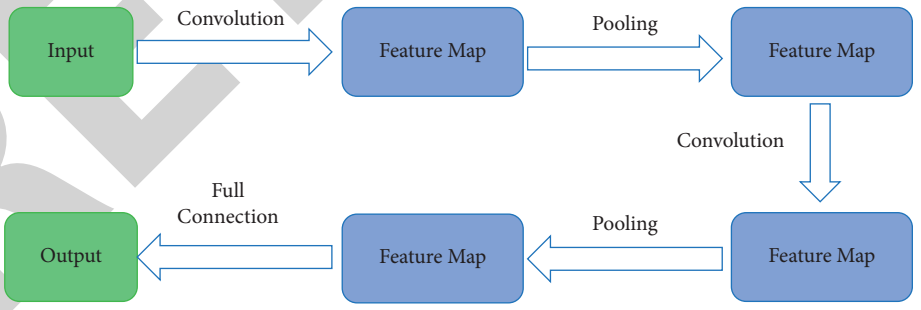

Figure 5: Schematic diagram of CNN model.

characteristics are local connection and convolution kernel sharing. Local connection is that the nodes of the convolution layer only connect some nodes of the previous layer, so as to reduce the number of parameters, improve the calculation speed, and effectively reduce the overfitting probability. Convolution kernel sharing means that when extracting a feature map, the same convolution kernel is shared between positions to reduce the number of parameters and further improve the calculation speed. The purpose of the pooling layer is to compress data, reduce parameters, and improve calculation speed. Fully connected layer: it is the hidden layer of a traditional neural network. Each neuron in this layer is connected with the previous neuron, so the number is the largest. In CNN, the convolution layer, pooling layer, and fully connected layer need to add activation functions. The common activation functions are sigmoid, tanh, and ReLU. The central value of the output value of the sigmoid function is not 0, and gradient

dispersion will occur during the backpropagation of the deep neural network. Figure 4 is a comparison diagram of sigmoid, tanh, and ReLU function curves.

tanh function: the output value of the function is centered on 0. Although the convergence speed is faster, there is gradient dispersion. ReLU function: it is the commonly used activation function at present, which solves the gradient dispersion, has fast convergence speed, and can reduce the possibility of overfitting. LeakyReLU activation function is an improved version of the ReLU function. The equation is as follows. Generally, 0.01 is taken as the value of a, which is a constant with a constant value.

$$f(x) = \begin{cases} x, & x \geq 0, \\ ax, & x < 0. \end{cases} \tag{18}$$

.

CNN propagates forward layer i output. Weight is W; offset is b. The input of each layer in the network is the output of the previous layer.

$$\begin{aligned} x^i &= f\left(u^i\right), \\ u^i &= W^i x^{i-1} + b^i. \end{aligned} \tag{19}$$

Overall loss function:

$$E^N = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{c} \left(t_k^n - y_k^n\right)^2. \tag{20}$$

.

N is total number of samples; c is number of sample categories; t_k^n is k-dimension of the n-th sample label; y_k^n is the n-th sample output of k-dimension. Each layer of CNN uses the gradient descent method to update the weight. The equations of weight update and offset update are as follows:

$$\begin{aligned} W_n^i &= W_o^i - \eta \frac{\partial E}{\partial W_o^i}, \\ b_n^i &= n_o^i - \eta \frac{\partial E}{\partial b_o^i}. \end{aligned} \tag{21}$$

W_o^i is weight and offset before updating; W_n^i and b_n^i are updated weight and offset; η is learning rate in gradient descent method. The convolution layer in CNN is forward propagation. The convolution output characteristic diagram of each i layer is as follows:

$$x_j^i = f\left(\sum_{l \in M_j} \cdot x_l^{i-1} \cdot k_{lj}^i + b_j\right). \tag{22}$$

M_j is input characteristic graph; k_{lj}^i is convolution kernel; f is activation function. The equations of CNN backpropagation, pooled layer error backpropagation, convolution layer error direction propagation, convolution layer weight update, and offset update are as follows:

$$\begin{aligned} \delta^{i-1} &= \text{up sample}\left(\delta^i\right) \odot \sigma'\left(u^{i-1}\right), \\ \delta^{i-1} &= \delta^i \left(\partial \delta^i / \partial \delta^{i-1}\right) \\ &= \delta^i * \text{rot180}\left(W^i\right) \odot \sigma'\left(u^{i-1}\right), \\ \partial E / \partial W^i &= \left(\partial E / \partial u^i\right)\left(\partial u^i / \partial W^i\right) \\ &= x^{i-1} * \delta^i, \\ \partial E / \partial b^i &= \sum u, v\left(\delta^i\right) u, v. \end{aligned} \tag{23}$$

up sample (δ^i) is sampling operation; is Hadamard product; δ is error; i is layer i; E is loss function.

3.4. Memory Enhancement AE. The generalization of AE itself is too high, resulting in the decoder being unable to decode abnormal coding well during reconstruction. The memory module is introduced to enhance AE. When new test information is input, the memory enhancement AE will not directly encode and input to the decoder but find relevant contents in the memory module and send all contents to the decoder. During training, the encoder and decoder update the memory module simultaneously to reduce the reconstruction error of normal samples and increase the reconstruction error of abnormal samples. Figure 6 presents a structure diagram of memory enhancement AE.

Figure 6 shows the specific process of memory enhancement AE and it is as follows. The encoder first obtains the encoded value of the input information, queries the relevant content according to the encoded value in the memory module, and then sends the content to the decoder for reconstruction. The output of the memory enhancement module is

$$\widehat{z} = \mathbf{w}M = \sum_{i=1}^{N} w_i m_i, \tag{24}$$

$\mathbf{w}$ is nonnegative row vector with sum 1; w_i is one of $\mathbf{w}$; M is memory storage unit coefficient; $\widehat{z}$ is reconstructed input features; N is memory storage unit capacity. Calculation during training and prediction:

$$\begin{aligned} w_i &= \frac{\exp\left(d\left(z, m_i\right)\right)}{\sum_{j=1}^{N} \exp\left(d\left(z, m_i\right)\right)}, \\ d\left(z, m_i\right) &= \frac{z m_i^T}{\|z\| \|m_i\|}. \end{aligned} \tag{25}$$

$d(z, m_i)$ is similarity variables of input features and memory storage units; $\mathbf{z}$ is input feature. To restrict the reconstruction of abnormal features, the correlation coefficient is constrained, and the small weight coefficient is set to 0 during hard compression.

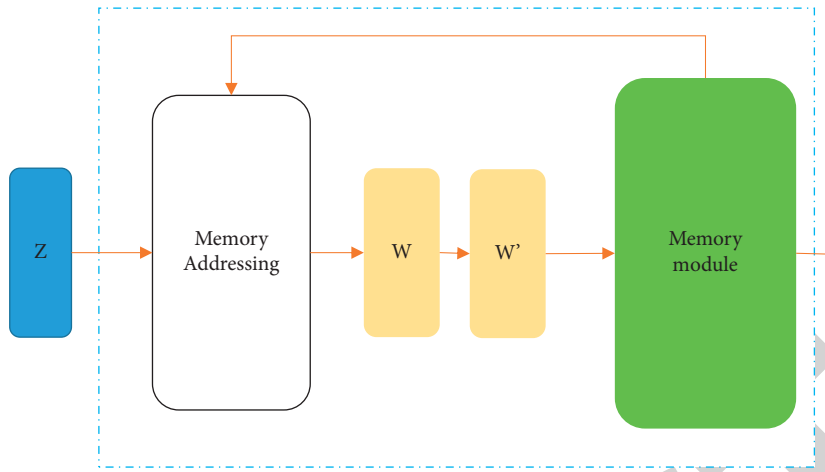

Figure 6: Memory enhancement structure.

$$w_i = h(w_i; \lambda) = \begin{cases} w_i, \text{ if } (w_i > \lambda), \\ 0, \text{ otherwise}, \end{cases} \quad w_i = \frac{\max(w_i - \lambda, 0) \cdot w_i}{|w_i - \lambda| + \varepsilon}. \tag{26}$$

λ is sparse threshold, value $1/N$; ε is very small positive scalar.

3.5. Surveillance Video Image Processing. There are multiple problems in traffic monitoring equipment, such as low image resolution, low video definition, and poor lighting conditions, resulting in image blur and so on. Before behavior analysis, image processing and interference elimination must be carried out on the surveillance video screen, that is, preprocessing some images. Image enhancement technology is to highlight crucial information and eliminate miscellaneous information to achieve the effect of image enhancement. Image enhancement includes image sharpening, smoothing, and histogram processing [17]. Image sharpening is to enhance the edge information of the image through operation, improve the clarity of the blurred image, facilitate observation and recognition, and extract the edge information of the target. For an image $F(x, y)$, the gradient vector and gradient amplitude are

$$\nabla F(x, y) = \left[\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}\right], \quad |\nabla F(x, y)| = \sqrt{\left[\left(\frac{\partial F}{\partial x}\right)^2 + \left(\frac{\partial F}{\partial y}\right)^2\right]}. \tag{27}$$

The equation of digital images is

$$\nabla F(x, y) = \sqrt{|[F(x, y) - F(x+1, y)]^2 + [F(x, y) - F(x, y+1)]^2|},$$
$$G(x, y) = \nabla F(x, y),$$
$$G(x, y) = \begin{cases} \nabla F(x, y), \nabla F(x, y) > T, \\ F(x, y), \text{ otherwise}, \end{cases} \tag{28}$$
$$G(x, y) = \begin{cases} L_\alpha, \nabla F(x, y) > T, \\ F(x, y), \text{ otherwise}. \end{cases}$$

The above three are methods of outputting sharpening results. Gradient replaces sharpening output, and the overall brightness of the image becomes lower, affecting recognition [18]. Output threshold judgment can be sharpened without affecting the background. Laplace operator is suitable for images with low contrast and brightness. This method is used to enhance video images here. For binary images, the expression is

$$\nabla^2 \mathrm{F} = \frac{\partial^2 \mathrm{F}}{\partial \mathrm{x}^2} + \frac{\partial^2 \mathrm{F}}{\partial \mathrm{y}^2},$$

$$\frac{\partial^2 F}{\partial x^2} = F(x+1, y) + F(x-1, y) - 2F(x, y),$$

$$\frac{\partial^2 F}{\partial y^2} = F(x, y+1) + F(x, y-1) - 2F(x, y), \tag{29}$$

$$\nabla^2 \mathrm{F} = F(x+1, y) + F(x-1, y) + F(x, y+1) + F(x, y-1) - 4F(x, y).$$

The above equations can be understood as follows. The Laplace operator of a point in the image is the difference between the gray value of its surrounding adjacent pixels and its own gray value. If the operator is rotated in a certain direction and added to the original operator, it is an eight-neighborhood operator:

$$g(x, y) = F(x, y) + c\left[\nabla^2 F(x, y)\right]. \tag{30}$$

$g(x, y)$ is output image; c is coefficient; $F(x, y)$ is original image. The enhanced image can be obtained by convoluting the obtained operator with the original image (Figure 6):

In Figure 7, the left is the original image of the surveillance video, and the right is the enhanced image. It is obvious that after the convolution operation on the image with low brightness, the image effect is significantly enhanced compared with the original image, which is very beneficial for the next analysis. Image smoothing is to filter out the high-frequency information in the image and retain the effective low-frequency information. Generally, a low-pass filter is used to remove the noise of the image [19]. Gaussian filtering is a common method. The binary Gaussian function and single element are calculated as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\left(x^2+y^2/2\sigma^2\right)},$$
$$H_{i,j} = \frac{1}{2\pi\sigma^2} e^{-\left((i-k-1)^2+(j-k-1)^2/2\sigma^2\right)}. \tag{31}$$

σ is variance; k is matrix dimension. The larger the σ is, the better the smoothing effect is. For the discrete image, discrete points are used as weights to weigh each pixel and the surrounding area to eliminate Gaussian noise. When the amount of calculation is too large, the filtering should be realized by Fourier transform. The advantage of bilateral filtering is that it considers not only the space near the image, but also the pixel similarity. It can eliminate image noise while retaining edge information, and the effect is better than the original denoising method [20]. The definition domain core and value domain core of output pixels are

$$g(i, j) = \frac{\sum_{k,l} f(k, l) w(i, j, k, l)}{\sum_{k,l} w(i, j, k, l)},$$
$$d(i, j, k, l) = \exp\left(\frac{(i-k)^2 + (j-l)^2}{-2\sigma_d^2}\right), \tag{32}$$
$$r(i, j, k, l) = \exp\left(-\frac{\|f(i, j) - f(k, l)\|^2}{2\sigma_r^2}\right).$$

The weight function of bilateral filtering can be obtained by multiplying the above two equations:

$$w(i, j, k, l) = \exp\left(-\frac{(i-k)^2 + (j-l)^2}{2\sigma_d^2} - \frac{\|f(i, j) - f(k, l)\|^2}{2\sigma_r^2}\right). \tag{33}$$

The advantage of bilateral filter is that it can protect the pixel value of the edge, but it is not good enough in color image processing. It can only suppress low-frequency noise and cannot deal with impulse noise well. The guided filter can filter out the noise and protect the edge information as much as possible [21] and it adopts a local linear process. The linear relationship between the filtered output of pixel i and the output image q and the guide image I is as follows:

$$q_i = \sum_j W_{ij}(I) p_j,$$
$$q_i = a_k I_i + b_k, \ \forall i \in \omega_k. \tag{34}$$

i and j are pixel subscripts; W_{ij} is filter core; (a_k, b_k) is constant coefficient; ω_k is filter window with radius r. Least squares optimization and window loss function are as follows:

Figure 7: Effect drawing of Laplace operator.

$$E(a_k, b_k) = \sum_{i\in\omega_k} \left((a_k I_i + b_k - p_i)^2 + \epsilon a_k^2 \right),$$

$$a_k = \frac{\sum_{i\in\omega_k}(p_i I_i) - b_k \sum_{i\in\omega_k}(I_i)}{\sum_{i\in\omega_k}(I_I + \epsilon)},$$

$$b_k = \sum_{i\in\omega_k} p_i - a_k \sum_{i\in\omega_k} I_i, \tag{35}$$

$$a_k = \frac{1/|\omega| \sum_{i\in\omega_k} I_i p_i - \mu_k \overline{p}_k}{\sigma_k^2 + \epsilon},$$

$$b_k = \overline{p}_k - a_k \mu_k.$$

3.6. The Abnormal Behavior Detection Process. In Figure 8, there is a thick fog on the left. After guided filtering, some fog is filtered out, and the image edge remains good. Guided filtering can remove the excess noise of the image while ensuring the original edge information. Therefore, it is used to remove the noise of the image in this experiment and Figure 2 is the Pedestrian detection.

4. Results and Discussion

Experimental results are explained here and different methods are already mentioned in Methodology section but they are analyzed here. Experimental data and evaluation index can be seen in the following discussion.

Memory enhancement AE is used. The network training samples are mainly from the following datasets. Avenue dataset: the videos in the dataset are real videos taken directly. Walking on the sidewalk belongs to normal behavior, while running and walking in the wrong direction and other behaviors are regarded as abnormal behavior. Ped1 and Ped2 datasets in UCSD include normal behavior and abnormal behavior. The main background is the pedestrian road, so the appearance of cars on the sidewalk is abnormal behavior.

The evaluation indexes of model performance are Area Under Curve (AUC) and Equal Error Rate (EER). AUC is defined as the area under the ROC curve. Its value is often used as the evaluation standard of the model because the ROC curve cannot clearly explain which classifier is better. However, as a value, the classifier with a larger AUC is better.

To prove the effect of the method proposed in traffic surveillance video abnormal behavior detection, memory enhanced self-coding is compared with the other six methods. Figure 9 is a diagram of the experimental results of four methods.

Figure 9(a) shows the comparison of detection results between MPPCA + SF and MDT. MDT has a better detection effect on the Avenue dataset than MPPCA + SF based on manual features, and its frame-level AUC is 15% higher. On Ped1 and Ped2 datasets, MDT still performs well, with an average frame-level AUC of about 10% higher than MPPCA + SF. Figure 9(b) presents the detection results of Conv and Conv3D based on AE, which basically reach about 75% of the frame-level AUC. In the Avenue dataset, the overall abnormal behavior detection performance on Ped1 and Ped2 datasets is good, and there is little difference between them. Figure 10 shows the comparison of the experimental results of the other two methods: RNN and ConvLSTM.

Figure 10 shows the comparison of detection results of AE-based ConvLSTM and Stacked RNN on each dataset. ConvLSTM performs better on the Avenue dataset. Compared with Stacked RNN, the former is better at obtaining time information. In the whole image, it can also be observed that the method based on self-coding achieves better frame-level AUC than the method based on handmade features.

4.1. Comparative Analysis of Experimental Results of Memory Enhancement Self-Coding. Figure 11 shows the comparison of experimental results of all methods.

Figure 12 shows the comparison of this method with other methods. The self-coding based on memory enhancement proposed is about 2% higher than the best

Figure 8: Denoising effect.

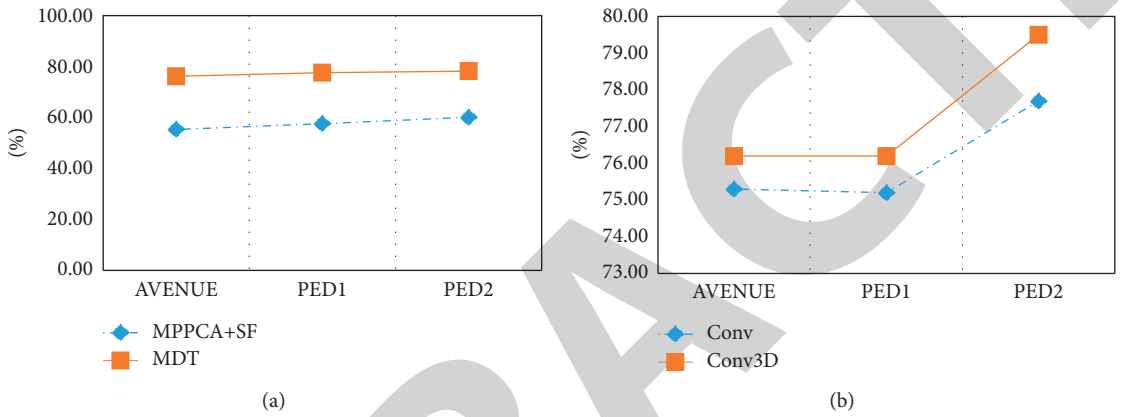

Figure 9: Comparison of experimental results of different methods. (a) Comparison of experimental results of MPPCA + SF and MDT. (b) Comparison of experimental results of Conv and Conv3D.

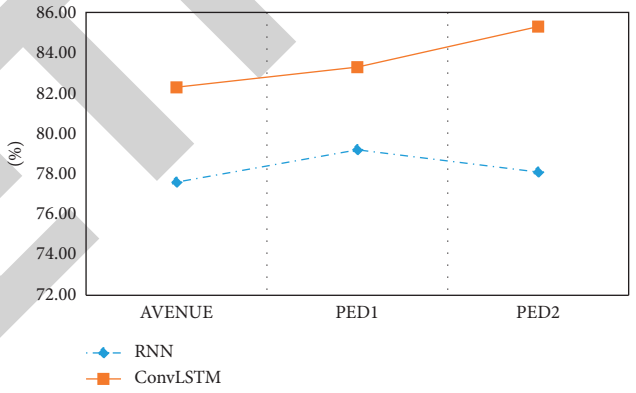

Figure 10: Comparison of experimental results of RNN and ConvLSTM.

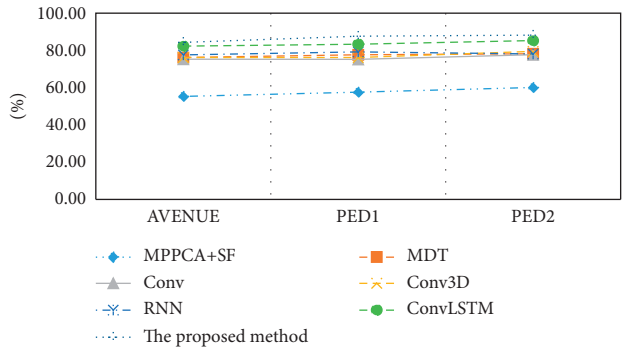

Figure 11: A comparison of experimental results between all methods and the proposed method.

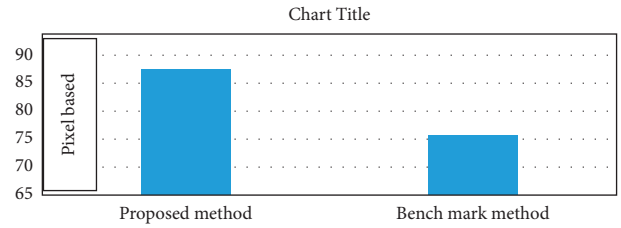

Figure 12: It shows the comparison of experimental results of all methods.

ConvLSTM in the above figure on the dataset Avenue and about 3% higher on the Ped1 and Ped2 datasets. This is mainly due to the introduction of the memory module. AE can well reconstruct abnormal information when reconstructing information. On the Ped1 and Ped2 datasets, there are more abnormal behaviors, which are more complex. The self-coding method based on memory enhancement can detect abnormal behaviors more accurately. In conclusion, self-coding based on memory enhancement can flexibly deal with the detection of different abnormal behaviors in different scenarios and use the reconstruction error information. The method proposed can obtain better results.

5. Recapitulation

The average accuracy from the proposed models is 87.5 percent as compared to the state of the art where average accuracy is 75.5 percent. The results are compared after being analyzed at microlevel like pixel based.

6. Conclusion

In this paper, the aim is mainly to study the abnormal behavior detection system of self-coding surveillance video based on memory enhancement. First, the principle of AE is analyzed. The generalization ability of AE can reconstruct exception information well, so it cannot be detected. Noise reduction AE is the use of artificial noise in the input to obtain better feature expression. The advantage is good robustness, while the disadvantage is that the time for adding noise before the training needs to be increased, and the training time increases. Convolution AE has the advantage of better image data processing and a better reconstruction effect. Then, the images with low definition are sharpened by Laplace transform and convolution, and the noisy images are processed by guided filtering. Finally, the experimental comparison of different methods is carried out. The self-coding based on memory enhancement proposed is about 2% higher than the best ConvLSTM on the dataset Avenue and about 3% higher on the Ped1 and Ped2 datasets, which verifies the effectiveness of the proposed method. The disadvantage is that the occlusion of characters will be more serious in the real scene, so it is necessary to consider adding algorithms to reduce the impact of occlusion. Moreover, the complexity of the real scene will be higher, so more situation and more complex scene data should be used to train the system.

Data Availability

The data and the MATLAB program used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no potential conflicts of interest.

Acknowledgments

This research was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant no. LGF21F020024 "Youth Academic Team Project of Zhejiang Shuren University."

References

[1] C. Lin, J. Pan, Z. Lian, and X. Shen, "Networked electric vehicles for green intelligent transportation," *IEEE Communications Standards Magazine*, vol. 1, no. 2, pp. 77–83, 2017.

[2] S. Y. Chen, "Analysis on the problems and improvement of informatization construction of basic Education in China under the epidemic crisis," *Advances in Education*, vol. 10, no. 6, pp. 1158–1163, 2020.

[3] Z. Fan, W. Li, Z. He, and Z. Liu, "Abnormal crowd behavior detection based on the entropy of optical flow," *Journal of Beijing Institute of Technology (Social Sciences Edition)*, vol. 28, no. 4, pp. 85–92, 2019.

[4] Z. Shen and W. Wu, "Video-based abnormal crowd behavior detection on bus," *Nanjing Li Gong Daxue Xuebao/Journal of Nanjing University of Science and Technology*, vol. 41, no. 1, pp. 65–79, 2017.

[5] L. Xia and Z. Li, "A new method of abnormal behavior detection using LSTM network with temporal attention mechanism," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3223–3241, 2021.

[6] K. Zhou, B. Hui, and J. Wang, "A study on attention-based LSTM for abnormal behavior recognition with variable pooling," *Image and Vision Computing*, vol. 108, no. 08, Article ID 104120, 2021.

[7] F. Harrou, M. M. Hittawe, Y. Sun, and O. Beya, "Malicious attacks detection in crowded areas using deep learning-based approach," *IEEE Instrumentation and Measurement Magazine*, vol. 23, no. 5, pp. 57–62, 2020.

[8] M. Geravanchizadeh and H. Roushan, "Dynamic selective auditory attention detection using RNN and reinforcement learning," *Scientific Reports*, vol. 11, no. 1, Article ID 15497, 2021.

[9] H. Pang, Y. Bu, C. Wang, and X. Hui, "Automatic detection of breast nodule in the ultrasound images using CNN," *The Journal of China Universities of Posts and Telecommunications*, vol. 26, no. 02, pp. 13–20, 2019.

[10] G. Cruz and A. Bernardino, "Learning temporal features for detection on maritime airborne video sequences using convolutional LSTM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6565–6576, 2019.

[11] M. Li, W. Zhe, Q. Xu et al., "A study on noise reduction for dual-energy CT material decomposition with autoencoder," *Radiation Detection Technology & Methods*, vol. 3, no. 3, pp. 44.1–44.13, 2019.

[12] T. Osa and S. Ikemoto, "Goal-conditioned variational autoencoder trajectory primitives with continuous and discrete latent codes," *SN Computer Science*, vol. 1, no. 5, 303 pages, 2020.

[13] W. Shin, S. J. Bu, and S. B. Cho, "3D-Convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance," *International Journal of Neural Systems*, vol. 30, no. 1, Article ID 2050034, 2020.

[14] K. Kumar and M. Haider, "Enhanced prediction of intra-day stock market using metaheuristic optimization on RNN–LSTM network," *New Generation Computing*, vol. 39, no. 10, pp. 1–42, 2020.

[15] T. Verma and S. Dubey, "Prediction of diseased rice plant using video processing and LSTM-simple recurrent neural network with comparative study," *Multimedia Tools and Applications*, vol. 80, no. 19, pp. 29267–29298, 2021.

[16] A. Haj-Manouchehri and H. M. Mohammadi, "Polyp detection using CNNs in colonoscopy video," *IET Computer Vision*, vol. 14, no. 5, pp. 241–247, 2020.

[17] Z. A. Mustafa, B. A. Abrahim, A. Omara, A. A. Mohammed, I. A. Hassan, and E. A. Mustafa, "Reduction of speckle noise and image enhancement in ultrasound image using filtering technique and edge detection," *Journal of Clinical Engineering*, vol. 45, no. 1, pp. 51–65, 2020.

[18] A. Yang, X. Tian, and B. Yang, "Single underwater image sharpening based on multi-feature fusion," *Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/Journal of Tianjin University Science and Technology*, vol. 51, no. 10, pp. 1031–1041, 2018.

[19] Y. Liu, F. Zhang, Y. Zhang, and X. Li, "Image smoothing based on histogram equalized content-aware patches and direction-constrained sparse gradients," *Signal Processing*, vol. 183, no. 4, Article ID 108037, 2020.

[20] X. L. Zhang and L. Dai, "Bilateral filtering," *Electronics Letters*, vol. 55, no. 5, pp. 258–260, 2019.

[21] C. Zhu and Y. Z. Chang, "Stereo matching for infrared images using guided filtering weighted by exponential moving average," *IET Image Processing*, vol. 14, no. 5, pp. 830–837, 2020.