Research on Character Behavior Recognition Based on Local Spatio-temporal Relationship in Surveillance Video

Weisong Dong*
School of Electric, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou, Fujian, 350118, China

*Corresponding author: 1171800479@qq.com

Abstract. Because of its wide application value, behavior recognition has long been one of the research hot spots in the field of computer vision and pattern recognition. At present, the method based on local features and word packet model has been widely used in the field of behavior recognition. However, this method does not consider the temporal and spatial relationship between features, and the local temporal and spatial relationship between features is very important for behavior representation and behavior recognition. In view of the above problems, this paper proposes a modeling method of character behavior recognition based on local spatio-temporal relationship in surveillance video. Firstly, each part of the proposed network model is introduced in detail, and then the proposed model is compared with the advanced skeleton action recognition methods in recent years on several skeleton data sets. Finally, the effectiveness of the proposed method is verified. The experimental results show that, compared with the recognition results of related literatures, the features extracted by choosing the starting point of trajectory have better recognition performance under the fusion framework.

Keywords: Surveillance video; Local space-time relationship; Character behavior recognition

1. Introduction

The current video surveillance system only uses cameras to acquire images and record them in real time. In this way, video data exists as a legal evidence only after the fact. Therefore, in order to truly achieve the purpose of video surveillance, what we need is a surveillance system that can process the monitored content in real time, so as to have enough time to prevent illegal activities.

Human behavior recognition is a hot and difficult point in computer vision research, and its key is to detect moving objects in video sequences, classify and track moving human bodies, and understand and describe their behaviors [1]. Human behavior recognition has important academic research significance and great practical application value in intelligent monitoring, motion analysis, video retrieval and analysis, human-computer interaction, medical diagnosis and monitoring.

Motion direction information is an important feature of describing actions in the field of behavior recognition. As for the local features of images, reference [2] puts forward a local directional pattern in which the directional optical flow histogram describes the motion features. Direction mode is
defined as dividing the direction ring into left and right parts, and the direction angles of the left and right parts are symmetrical. Literature [3] uses the relative maximum motion direction and second-order statistics of direction distribution of cascaded optical flow characteristics to describe motion characteristics. Document [4] uses Markov chain to transform the trajectory shape into stationary vector representation. Different from reference [4], we only study the direction information of track displacement. For trajectory shape descriptors, the response values of directions on each predefined direction template are counted respectively. Here, the predefined direction template considers the direction starting point configuration and the distribution information of the thickness horizontal direction.

The main content of this paper is the detection of abnormal behavior of video target under local spatio-temporal relationship. Based on the low efficiency, time-consuming and labor-consuming problems caused by manual recognition of abnormal behavior in video under the current video big data, this paper proposes a new target abnormal behavior detection model, skeleton action recognition model based on local time-space relationship. The proposed local spatio-temporal convolution network is used to obtain the spatio-temporal features of skeleton sequences in a short period of time, and the short-temporal memory network based on frame segment selection is used to obtain the spatio-temporal features of skeleton sequences in a long period of time, thus solving the problem of skeleton sequence recognition in a long period of time.

2. Character behavior recognition algorithm
In recent years, the method of establishing behavior model of local features has achieved good recognition performance [5], especially the behavior model of trajectory features has become a hot research direction [6]. Trajectory features can capture the process of moving objects changing with time, and obtain rich geometric structure information of moving objects. In the recognition stage, the multi-class descriptors of trajectory features have achieved better recognition performance under the framework of multi-channel fusion and SVM classification recognition.

2.1. Extract relative motion sequence diagram
Moving edge is a method to restrain the uniform motion of camera. In fact, edges and their uniform regions contain abundant motion features. Compared with reference [10], this paper uses super pixels and detected moving edges to extract important moving objects. After passing the method of reference [11], the optical flow gradient is calculated according to equation (1).

\[
v_x = \frac{\partial v(x, y)}{\partial x}, v_y = \frac{\partial v(x, y)}{\partial y}
\]

In which \(v_x, v_y\) represent gradient maps and optical flow maps in horizontal and vertical directions, respectively.

In a restricted environment, changes to adjacent frames only occur in local areas of moving objects, and most areas are relatively static. Therefore, according to the distribution statistics of the moving edge map, we use adaptive threshold to select the moving edge of the object of interest, and use superpixels containing important moving edges as the changing region of the moving object. Firstly, we define the number \(a\) of changing pixels in adjacent frames, which is proportional to the resolution of video frames, as shown in formula (2).

\[
a = WHx
\]

Where \(W, H\) is the width and height of the video frame, \(x\) is an empirical constant, which is set to 0.002 in the experiment. \(W, H\) is the width and height of the video frame, and \(x\) is an empirical constant set to 0.002 in the experiment. Calculate the distribution of moving edge map using \(Bin\) quantization interval, sort the index of each quantization \(Bin\) according to the ascending number of pixels in the quantization...
interval, then accumulate the number of pixels in the quantization interval according to the formula, select a rising index, and select a quantization level whose accumulated number is not greater than $\alpha$ index. Finally, the last index number $i_{\text{Ind \_max}}$ is selected from the index set, and the adaptive threshold $\beta$ is calculated according to equation (3).

$$
\beta = V_{\text{min}} + i_{\text{Ind \_max}}(V_{\text{max}} - V_{\text{min}})/\text{Bin}
$$

Where $V_{\text{max}}, V_{\text{min}}$ is the pixel extremum of the edge image. The edge coordinates moving with pixel values larger than $\beta$ are selected by thresholding, and a $\{ \} \in = \{ z_k : z_k \in R^2 \}$ set is constructed. In which $z_k$ represents the $k$th coordinate.

2.2. Learning process

The purpose of this paper is to learn the behavior model $\{ \tilde{T}_M \}_{k=1}^M$ from the training samples. Because $\{ \tilde{T}_M \}_{k=1}^M$ can be regarded as a parameter of classifier (4), this paper needs to construct classifier (4) for classification.

$$
y = \max_{h \in H}(x^T Wh)
$$

Where $W$ is the sample weight.

However, it is difficult to construct only a strong classifier (mainly because the constructed classifier needs to search a large space composed of the symbiotic relationship between local features), but it is relatively simple to construct some weaker classifiers $\{ g_i(F) \}_{j=1}^f$. Therefore, this paper uses Ada Boost framework to combine some weak classifiers to get a strong classifier, that is, to get a weak classifier by sampling; Update distribution; Combining weak classifiers into strong classifiers. Firstly, we need to sample from two polynomial distributions to get the weak classifier. These two polynomial distributions are defined as follows:

$$
SR_1 = \left\{ \frac{\text{hist}_i^{\text{pos}}}{\text{hist}_i^{\text{pos}} + \text{hist}_i^{\text{neg}}} \right\}_{i=1}^K
SR_2 = \left\{ \frac{\text{hist}_{ij}^{\text{local}}}{\text{hist}_{ij}^{\text{local}}} \right\}_{i=1}^K
$$

Among them, $\text{hist}_i^{\text{pos}}, \text{hist}_i^{\text{neg}}$ is the number of the $i$th local feature in the positive sample and the negative sample, so sampling from $SR_1$ can get the local feature with discriminant ability; $\text{hist}_{ij}^{\text{local}}$ is the frequency of the $i,j$th symbiotic mode. Every weak classifier $g_i(F)$ determines whether there are patterns $\{ S_{i,j}^{\text{pos}} \}_{i=1}^\nu$ and $\{ S_{i,j}^{\text{neg}} \}_{i=1}^\nu$.

In this paper, $\{ S_{i,j}^{\text{pos}} \}_{i=1}^\nu$ and $\{ S_{i,j}^{\text{neg}} \}_{i=1}^\nu$ are obtained from visual dictionary $C$ by sampling. Sampling process is based on the distribution information provided by $SR_1$ and $SR_2$. It should be pointed out that $SR_1$ and $SR_2$ are not static, and the subsequent process needs to be updated according to the classification error of the weak classifier obtained by this sampling.

2.3. Skeleton action recognition network model based on local spatio-temporal relationship

The network architecture is divided into three parts: local spatio-temporal convolution network (LST-CNN), fragment selection long short-term memory (FS-LSTM) and full convolution network.
2.3.1 Local spatio-temporal convolution network. According to the structural characteristics of three-dimensional skeleton sequence, this paper proposes LST-CNN, and the schematic diagram of this network is shown in Figure 1.

First, define the input data. The skeleton sequence segment is divided into continuous \( n \) segments, and the continuous skeleton sequence of these \( n \) segments is taken as the input of the network.

Assume that the input sequence is \( x = (x_1, x_2, \ldots, x_n) \), \( n \) represents a total of \( n \) segments of continuous skeleton sequences, \( x_i \) represents the joint data of the human body in the \( i \) period, and \( x_i \) is a three-dimensional matrix of \( s \times l \times t \). In which \( l \) represents the dimension of human joint data of each frame, \( s \) represents the number of skeleton joints, and \( t \) represents the number of frames of this skeleton sequence, that is, the size of time period. Then, space-time convolution operation is performed according to the input data. The format of the input data is \( s \times l \times t \). When the input is 3D skeleton data, \( l \) is 4, \( s \) represents the number of joints of the extracted skeleton data, and \( t \) represents the sequence length of the input skeleton data. The three-dimensional data represents all skeleton information in time period \( t \), and each time point represents a frame of skeleton data. Set the filter size to the size \( l \) of skeleton data dimension, and the filter dimension is \( d \), where \( d < t \). Therefore, every convolution operation will include the convolution operation of \( L \) skeleton joints in continuous \( D \) time period. This convolution operation can not only obtain the temporal feature representation of skeleton data, but also obtain the spatial feature representation of skeleton data.

2.3.2 Fragment Selection Long Short-Term Memory. According to the different degree of attention in behavior discrimination in each time period, FS-LSTM is proposed in this section.

The \( n \) feature sets \( F = (f_1, f_2, \ldots, f_n) \) obtained by local spatio-temporal convolution neural network are used as the input of the second part of the network, FS-LSTM, and each time FS-LSTM receives a spatio-temporal feature map.

According to the input of the formula at each time, a high-level feature map can be obtained, so \( n \) high-level feature maps can be obtained after the \( n \)-segment skeleton sequence feature map passes through LSTM network. Combine and output all time feature maps output by LSTM, which can be expressed by formula (6) and formula (7):

\[
h_i = g(f_i), i \in (1, n) \quad (6)
\]

\[
Q = (h_1, h_2, \ldots, h_n) \quad (7)
\]

The \( g \) function in formula (6) represents the feature extraction process of LSTM, and formula (7) is the result of combining the feature images at all times output by LSTM.

The gate of LSTM represents the time period selection process in skeleton sequence, which preserves useful time period features and discards useless time period features, that is, obtains the most useful time series features in skeleton sequence. Because of the continuous activation of the function, the obtained useful information will increase with time, which makes the information retained in the output \( h_n \) at the last moment less and less. Therefore, in order to make the network get all useful information evenly, FS-LSTM design averages all the output results to get the global feature graph as
the second half of the network. The formula is shown in (8).

$$S_j = \frac{1}{n} \sum_{i=1}^{n} h_{ij}, i \in (1, n), j \in (1, k)$$  \hspace{1cm} (8)

$$S = [S_1, S_2, \cdots, S_k]$$  \hspace{1cm} (9)

In formula (9), $S$ is the global feature representation of the whole skeleton sequence.

Because the local spatio-temporal feature map of skeleton extracted by convolution network is input into LSTM network, the spatio-temporal features of all skeletons are obtained and the computation of LSTM is greatly reduced. Compared with directly inputting the whole skeleton sequence into LSTM network for training, it improves the efficiency of network training and operation.

3. Experimental analysis and prospect

3.1. Recognition rate analysis

In order to verify the effectiveness of the algorithm proposed in this paper, this section has carried out experiments on the public data set KTH database, which contains 600 videos, "boxing, run, jogging, walk, wave, applaud". Each type of behavior was repeated several times by 25 people. Specifically, most video clips are repeated 4 times by one person, and a few video clips are repeated 3 times by one person, and one video clip is missing.

In this section, all video clips are divided into several small videos according to the repetition times of behaviors, and these small video clips only contain the single behavior of a single person. Therefore, the total number of small video clips is 2276. All videos were shot in 4 different scenes: Outdoor 1 (without scale change), outdoor 2 (with scale change), outdoor 3 (wearing different clothes) and indoor. The background of most videos is simple and there is no significant change. The method proposed in this section can improve the recognition rate by increasing the size of the visual dictionary. The change of the recognition rate of six kinds of behaviors with the size of the dictionary is shown in Figure 2.

![Figure 2](image)

Figure 2 Changes of recognition rate of six kinds of behaviors with dictionary size

It can be seen from fig. 4 that with the increase of visual dictionary, the average recognition rate tends to increase. Because these local descriptors are easy to be confused, the model obtained from running has more discriminant features, while the model obtained from jogging has fewer discriminant features, so the recognition rates of running and jogging have opposite growth trends.

3.2. Analysis and comparison of operation time

Analysis of time expenditure, taking the "riding" behavior sequence of UCF-sports behavior database in experimental environment and parameter setting as an example. Each video of this motion category has 60 frames with a resolution of 710x418, and the number of tracks is expressed as the effective average number of tracks. The time cost of trajectory extraction without selecting interest points is composed of trajectory extraction and static trajectory removal. The time required by this algorithm is
shown in Table 1.

**Table 1** Time comparison of different trajectory extraction

| Way                              | Number of effective tracks | Time consuming (s/seq) |
|----------------------------------|----------------------------|-----------------------|
| No valid feature points are      | 524                        | 94.83                 |
| selected                         |                            |                       |
| Select feature points            | 368                        | 77.21                 |

In Table 1, it can be seen that the time required for trajectory extraction under the condition of selecting the relative working point is less than the time when no tracking point is selected. The method in this paper reduces the number of tracking tracks and shortens the calculation time by about 17 seconds. Compared with reference [12], when the average number of tracks is 368, the average tracking time per frame is 1.22 ms. The main time of the track extraction method based on the above analysis is to extract the optical flow and select the relatively moving superpixel part, and the total time required is larger than that of the method in reference [12]. However, our method can extract effective trajectory features and obtain better recognition performance.

3.3. **Model analysis**

The quality of the model network depends not only on the number of iterations in the training process, but also on the optimization of the network model. In order to verify whether the number of hidden layer nodes will affect the experiment, this paper takes the number of hidden layer nodes as 100, 200, 300, 400 and 500, respectively. The experimental results are shown in Figure 3, and the specific experimental results are shown in Table 2.

![Figure 3](image)

**Figure 3** Experimental comparison of different numbers of hidden layer nodes

**Table 2** Experimental comparison table of the number of hidden layer nodes

| Number of hidden layer nodes | Learning rate | d size of the first convolution layer | Iterations | Highest recognition rate(%) |
|------------------------------|---------------|-------------------------------------|------------|----------------------------|
| 100                          | 0.02          | 6                                   | 1500       | 82.35                      |
| 200                          | 0.02          | 6                                   | 1500       | 83.66                      |
| 300                          | 0.02          | 6                                   | 1500       | 88.69                      |
| 400                          | 0.02          | 6                                   | 1500       | 85.01                      |
| 500                          | 0.02          | 6                                   | 1500       | 87.11                      |

According to the chart analysis, when the number of hidden layer nodes is 300, the highest recognition rate of this model can reach 88.69%.

4. **Conclusion**

In this paper, the character behavior recognition based on local spatio-temporal relationship in surveillance video is modeled, and skeleton action recognition model based on local spatio-temporal relationship is adopted. This method extracts rich feature tracks and eliminates static tracks. For the trajectory shape descriptor, the predefined multiple direction modes and corresponding quantization
levels are used to describe the trajectory shape. At the same time, the parameters of the model are optimized and the selection of the model parameters is verified. Experimental results show that the extracted feature track can describe the motion changes of the object of interest well. Compared with the direction amplitude descriptor, the multi-direction pattern statistical descriptor improves the recognition performance. Compared with the recognition results of other methods, the extracted effective trajectory features have better recognition performance by multi-core learning fusion.

References
[1] Liu Y, Yang J, Wang C. Research on Behavior Recognition Algorithm Based on Compact Representation of Low-level Feature. Journal of Physics Conference Series, 2019, 1176:032018.
[2] Kai Z, Tw D, Cw D, et al. Skeleton Based Abnormal Behavior Recognition Using Spatio-Temporal Convolution and Attention-Based LSTM. Procedia Computer Science, 2020, 174:424-432.
[3] Fakhar B, Kanan H R, Behrad A. Event detection in soccer videos using unsupervised learning of Spatio-temporal features based on pooled spatial pyramid model. Multimedia Tools and Applications, vol. 78, no. 12, pp. 16995-17025, 2019.
[4] Wu H, Yang M, Yang S, et al. A novel DAS signal recognition method based on spatiotemporal information extraction with 1DCNNs-BiLSTM network. IEEE Access, no. 99, pp. 1-1, 2020.
[5] Qu J, Liu F, Ma Y, et al. Temporal-spatial Collaborative Prediction for LTE-R Communication Quality Based on Deep Learning. IEEE Access, no. 99, pp. 1-1, 2020.
[6] MEI-JIAO WANG, LI-ZHEN WANG, LI-HONG ZHAO. Spatial Co-location Pattern Mining Based on Fuzzy Neighbor Relationship. Journal of Information Science and Engineering, vol. 35, no. 6, pp. 1343-1363, 2019.
[7] Pan hui, Jiang zhenlan, Zhang Baoyu. geographically weighted regression-based research on the spatial relationship of the influence of urban land use on land surface temperature. journal of Fujian forestry university, vol. 039, no. 002, pp. 165-173, 2019.
[8] Hadian F, Jafari R, Bashari H, et al. Estimation of spatial and temporal changes in net primary production based on Carnegie Ames Stanford Approach (CASA) model in semi-arid rangelands of Semirom County, Iran. Journal of Arid Land, vol. 11, no. 4, pp. 477-494, 2019.
[9] Wang Y, Dai E. Spatial-temporal changes in ecosystem services and the trade-off relationship in mountain regions: A case study of Hengduan Mountain region in Southwest China. Journal of Cleaner Production, 2020, 264:121573.
[10] MR Rahnama, R Wyatt, Shaddel L. A spatial-temporal analysis of urban growth in melbourne; Were local government areas moving toward compact or sprawl from 2001–2016. Applied Geography, vol. 124, no. 2, pp. 102318, 2020.
[11] Han F, Zhu J, Liu B, et al. Fish Shools Behavior Detection Based on Convolutional Neural Network and Spatiotemporal Information. IEEE Access, no. 99, pp. 1-1, 2020.
[12] Qs A, Yu Z A, Lu S B, et al. Spatial–temporal differences in operational performance of urban trunk roads based on TPI data: The case of Qingdao. Physica A: Statistical Mechanics and its Applications, 2020, 568.