Hindawi
Scientific Programming
Volume 2021, Article ID 2934943, 8 pages
https://doi.org/10.1155/2021/2934943

Research Article

Group Behavior Pattern Recognition Algorithm Based on Spatio-Temporal Graph Convolutional Networks

Xinfang Chen [1] **and Venkata Dinavahi** [2]

[1]*College of Information Engineering, Institute of Disaster Prevention, Sanhe 065201, Hebei, China*
[2]*Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6G 1H9, Alberta, Canada*

Correspondence should be addressed to Xinfang Chen; chenxinfang@cidp.edu.cn

Received 25 April 2021; Revised 29 May 2021; Accepted 5 June 2021; Published 8 July 2021

Academic Editor: Shah Nazir

With the rapid growth of population, more diverse crowd activities, and the rapid development of socialization process, group scenes are becoming more common, so the demand for modeling, analyzing, and understanding group behavior data in video is increasing. Compared with the previous work on video content analysis, factors such as the increasing number of people in the group video and the more complex scene make the analysis of group behavior in video face great challenges. Therefore, a group behavior pattern recognition algorithm based on spatio-temporal graph convolutional network is proposed in this paper, aiming at group density analysis and group behavior recognition in the video. A crowd detection and location method based on density map regression-guided classification was designed. Finally, a crowd behavior analysis method based on density grade division was designed to complete crowd density analysis and video group behavior detection. In addition, this paper also proposes to extract spatio-temporal features of crowd posture and density by using the double-flow spatio-temporal map network model, so as to effectively capture the differentiated movement information among different groups. Experimental results on public datasets show that the proposed method has high accuracy and can effectively predict group behavior.

1. Introduction

The growth of population and the diversity of crowd activities make group scenes become common. Group behavior [1–3] contains many important clues in interdisciplinary fields. Understanding the formation mechanism of group behavior has long been one of the important research topics in sociology and natural science. When the number of people in the video increases and the crowd scene becomes more complex [4], how to automatically and effectively model, analyze, and understand the group behavior data so as to better serve human beings becomes an important challenge. Research on group behavior analysis can provide support and corresponding solutions for many key engineering applications, such as intelligent video surveillance, crowd anomaly monitoring, and public facility planning. From the perspective of the cognitive mechanism of group behavior, this paper studies the effective computational framework and algorithm model of group behavior, trying to mine the dynamic group pattern and behavior in the real scene video data [5, 6], so as to solve the practical problems in the field of computer vision.

At present, the problem of group behavior analysis [7, 8] in video is based on ordinary surveillance video, and it uses computer vision technology to understand and analyze group behavior and events in the monitored scene. This changes the problem that traditional video surveillance relies too much on manpower. It can automatically realize the analysis and description of group behavior and realize the intelligent monitoring of large-scale crowd scenes. Group behavior analysis and understanding has become an important research branch of video surveillance, which has been widely applied in many fields such as public security, transportation, and facility planning. At the same time, the vigorous development of artificial intelligence [9–12], machine vision [13–15], cognitive science, and other cutting-

edge technologies also provides a guarantee for intelligent understanding of video content. Previous behavior understanding work [16] in video content analysis mainly focused on understanding individual behaviors, such as motion detection, target tracking, and object recognition, while ignoring the understanding of large-scale group behaviors. Compared with the analysis and recognition of individual behavior, group behavior is more real and complicated.

In the detection of crowd density [17, 18] and crowd behavior [19] by a computer, the movement of the crowd is complex and the scene is changeable. Due to the change of illumination, the blocking of the crowd, the perspective effect, the different shooting angles, and other factors, it will bring difficulties to the detection by the computer. Crowd behaviors have different semantics in different scenes. It is of great significance to quickly and effectively understand and distinguish the semantics of normal and abnormal behaviors of crowds and realize effective judgment, which is an urgent problem to be solved in the field of computer vision [20].

The processing of video images through computer vision can further replace manual monitoring to perform real-time and efficient monitoring of crowd density and crowd behavior. Recently, many scholars have used deep learning-based methods to conduct research on multiple tasks such as pedestrian detection, face recognition, and group behavior recognition and have made major breakthroughs. At present, when computer vision performs crowd detection, there are problems such as large crowd, poor detection accuracy, variable scenes, and high complexity. The existing technology can effectively overcome the abovementioned difficulties on the basis of deep learning and affect the distribution of the population. As well as by real-time monitoring of behavior, it provides solutions for crowd supervision, which has great practical significance and application value.

The main innovations and contributing points of this paper is to propose a group behavior pattern recognition algorithm [21, 22] based on spatio-temporal graph convolutional network, which can effectively recognize group behavior. The paper also proposed to use the dual-stream spatio-temporal map network model to extract spatio-temporal features of the crowd posture and density to effectively capture the differentiated movement information between different crowds.

The paper is organized as follows. Section 2 represents briefly the related work to the proposed research. Section 3 elaborates the methodology of the paper with details in sections. Experiments and results of the paper are given in Section 4. The paper is concluded in Section 5.

2. Related Work

The initial population research is mainly based on the detection of the crowd. The image is segmented before the target detection of the crowd using a sliding window, and finally, the crowd is counted based on the classifier. Detection-based methods include detection based on the whole [23] and detection based on parts of the human body [24]. The typical traditional method uses random forest matrix, SVM detector, and other methods to train the classifier and extracts various features such as pedestrian direction gradient histogram, edge, texture, and whole body wavelet. In scenes with highly dense crowds, crowds are severely occluded, and the method of detecting parts of the human body such as the head and shoulders is used instead of the method based on overall detection. The effect is improved, but the robustness of human detection is still not high.

Crowd density analysis and crowd counting based on regression are mainly used to learn the mapping relationship between image features and number of people [25]. Image segmentation is based on the regression method first, the image, texture, edge, and the prospect of gradient low-level features such as extraction and then the linear regression, Gaussian regression, ridge regression, and regression function are studied, such as learning exists in the mapping function of the number of low-level features and the image, generating a static background model, which is sensitive to illumination changes. The model needs to be retrained each time the scene is transformed, which is costly in terms of time and computation. Regression-based methods usually believe that the relationship between the number of people in the image and the foreground area can be approximately linear. However, such linear relationship is difficult to be established because of the problems of occlusion, overlap, and perspective of the crowd in the real scene.

In densely crowded images, deep learning usually uses convolutional neural networks to generate end-to-end models to extract features of different scales of pedestrians in the image, so as to generate crowd density maps through Gaussian kernel functions to achieve the effect of crowd counting [26]. The crowd density map can not only realize crowd counting but also provide rich spatial information, detect crowd density distribution, and further analyze crowd behavior through crowd density detection. Zhang et al. [27] proposed a multicolumn deep convolutional neural network MCNN model, which used different subnetworks with different convolutional kernel sizes to realize crowd count in the scene of serious crowd occlusion and height transformation. In the latest research, Sam et al. [28] proposed a switching network based on MCNN, which has multiple CNN subnetworks with different depths and different convolution kernel sizes of each subnetwork, thus improving the accuracy and robustness of crowd density analysis and crowd count results of high occlusion and multiscale scene transformation. Sindagi and Patel [29] proposed a context pyramid model CP-CNN. In order to extract global and local context information, the network learns the MCNN network of multicolumn architecture, designs two subnetworks to map the input image or video frame data to a high-dimensional feature map, and uses the CNN network to estimate the context at all levels. To reduce technical errors and generate higher quality density maps, Li et al. [30] proposed the deep neural network [31–35] model CSRNET, which abandoned the multicolumn framework, and believed that the multicolumn framework had no obvious advantages compared with the single-column framework. The front end of the model was the VGG-16 model, which abandoned the full connection layer and only retained the convolutional

layer and pooling layer, followed by the void convolution to expand the receptor field and obtain the features of different levels of images. Generate the population density distribution map, and obtain better detection results.

3. Methodology

3.1. Overall Framework. Crowd flow in video has the characteristics of time dynamic, space correlation, and uncertainty. Aiming at these characteristics, this paper proposes a kind of spatio-temporal dynamic graph convolutional network [36, 37] to study and predict crowd flow. Figure 1 shows the framework of spatiotemporal dynamic graph convolutional network (STDGCN) proposed in this paper. The STDGCN model consists of an input transformation layer, an STDGCN layer, and an output layer composed of a full connection layer.

The model uses the spatio-temporal data collected by the crowd flow sensor in the video and external factors to predict the crowd flow and other parameters in the future and comprehensively obtain the spatio-temporal network [38] prediction output. The input conversion layer embeds and converts crowd flow attribute data and exogenous factor data, among which three types of data are used for exogenous factors. The STDGCN layer contains a graph convolution module and a time-dimensional encoder-decoder structure. The output layer generates the prediction result of each node through a fully connected layer.

The core ideas of the STDGCN model can be summarized in the following two points. First, regard the sensor data at the same time as a graph data, connect the nodes and neighbor nodes to represent the spatial correlation of the crowd flow, and use the graph convolutional network to capture traffic. Second, treat the data at different moments of the same node as a time series and use the gated recurrent unit and attention mechanism to deal with the time dynamics of the traffic flow. The STDGCN layer structure is shown in Figure 2.

The spatio-temporal dynamic graph convolution module consists of two parts: graph convolutional network (GCN) and attention encoder network (AEN). Graph convolutional network is used to deal with the spatial dependence of crowd data, and attention encoder network is used for capture time dimension dynamics.

3.2. Spatial Feature Extraction of Crowd. Compared with the use of two-dimensional image convolution to obtain the patterns and characteristics of the crowd, the pedestrian sensor data with the characteristics of map data can obtain more primitive and real spatial attributes. In the proposed model STDGCN, graph convolution is directly applied to graph structure, and highly meaningful patterns and features are extracted in the spatial domain. Traditional convolutional neural networks can effectively extract local features of data, but they are not suitable for general graph structures. There are two types of methods to generalize convolutional neural networks to graph structures. One method is to expand the spatial definition of convolution, and the other is to use the Fourier transform of the graph to operate in the spectral domain.

The spectrogram method of graph convolution is to use the diagonalized linear operator defined in the Fourier domain to convolve the graph signal and use the convolution kernel g_θ. The convolution operation on the graph signal G_v can be expressed as

$$g_\theta \times G_v = g_\theta\left(U \wedge U^T\right)v = U g_\theta(\wedge)U^T v, \tag{1}$$

where U is the Fourier basis composed of eigenvectors and $g_\theta(\wedge)$ is the diagonal matrix composed of eigenvalues of L. Because the scale of the graph becomes larger, that is, when the crowd is large, the computational complexity of eigendecomposition of the Laplace matrix in equation (1) is very high, which can be approximated by Chebychev polynomial:

$$g_\theta \times G_v = g_\theta(L)v \approx \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})v. \tag{2}$$

3.3. Time Feature Extraction of Crowd Flow. As shown in Figure 2, after the crowd flow data is extracted through the graph convolutional network for spatial feature extraction, the spatial feature sequence and the embedding representation of exogenous factors are used as the input of time dimension modeling. The AEN module is composed of two GRU networks with independent parameters. The GRU network on the left is the encoder module, and the GRU network on the right is the decoder module. The encoder encodes the input time sequence and initializes the decoder through the last moment of the encoder. Module, the decoder generates prediction output from the context vector in time steps:

$$h_{i,t} = \mathrm{GRU}\left(\left[y_{i,t}; e_{i,t}\right], h_{i,t-1}\right), \tag{3}$$

where $h_{i,t}$ is the output representation of sensor node i at time t, $y_{i,t}$ is the feature sequence obtained by the graph convolution operation, and $e_{i,t}$ is the exogenous factor.

A potential problem of the encoder-decoder model is that the model needs to be able to compress the context information of the source sentence into a fixed-length vector. This makes it difficult for the model to handle long sequences, especially those longer than the feature sequences in the training data. In other words, the encoder-decoder model may be difficult to grasp the longer periodic features in the crowd flow, such as weekly regularity.

3.4. Semantic Relevance of Group Behavior. In this paper, a spatio-temporal correlation model of video is designed to infer the behavioral semantics of group figures in video sequences. The model is composed of two layers of GRU. The first layer of GRU predicts feature mask sequence $E = \{E_1, E_2, \ldots, E_N\}$ to encode. The second layer GRU decodes the hidden codes of the first layer output one by one and outputs the action semantics of the characters and the

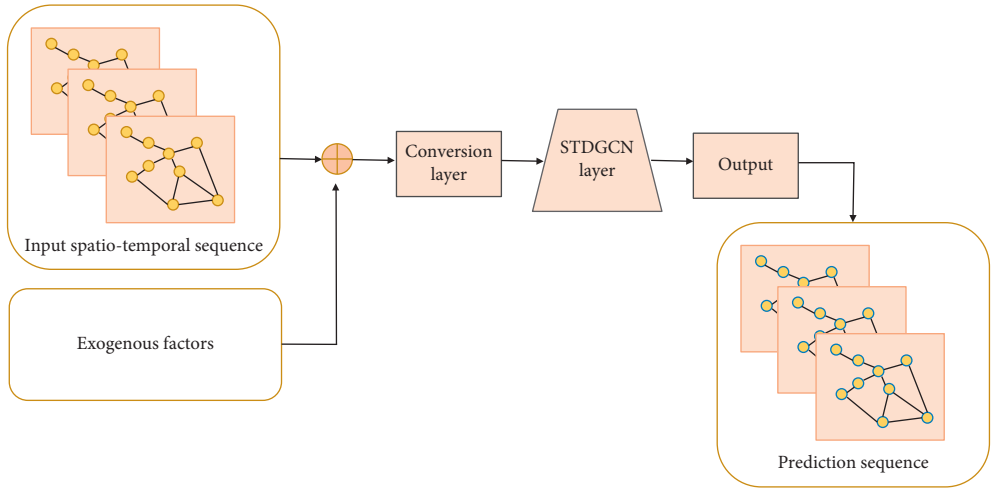


Figure 1: Schematic diagram of spatio-temporal dynamic graph convolutional networks.

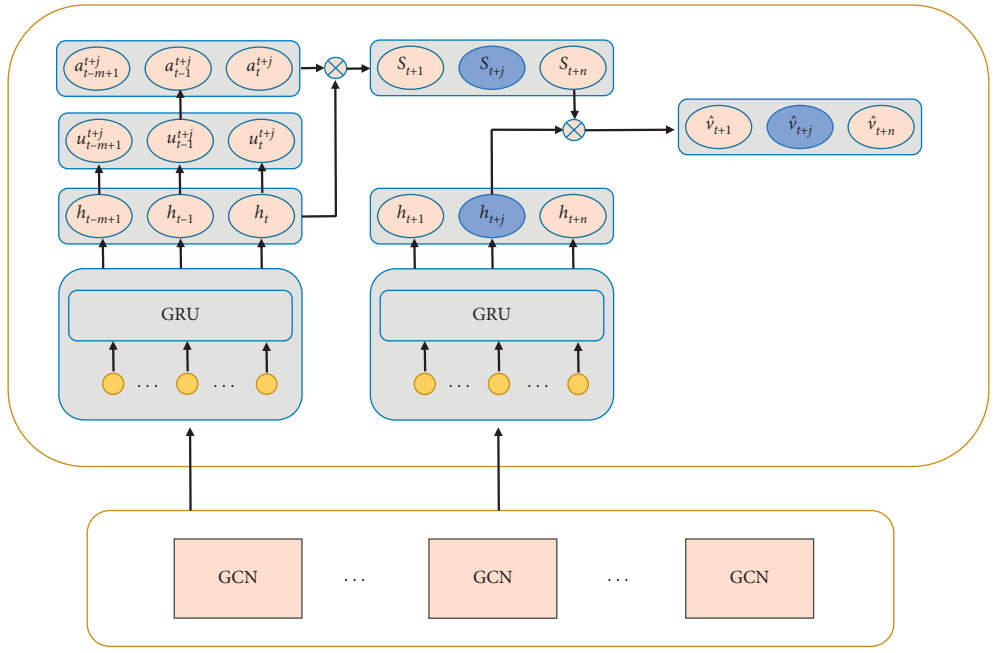


Figure 2: STDGCN layer structure.

behavior semantics of the group characters in the corresponding time sequence after spatio-temporal correlation. The model can be divided into two stages: encoding stage and decoding stage in the process of spatio-temporal correlation of group character behavior.

In the coding stage, the first part of the GRU structure of the first layer has a value of code_i^t, and the hidden layer information H_t is calculated. H_t includes the hidden layer information $h_t = \{h_1^t, h_2^t, \ldots, h_n^t\}$ of a single person and the hidden layer output g_t of the group of people. The equation for calculating the output of each character in the tth video frame is as follows:

$$p_i^t = x_i^t \oplus h_i^t, \tag{4}$$

where x_i^t is the predicted feature mask input of the person i in the tth frame of the first-layer encoding stage, h_i^t is the output result of the hidden layer of the person i in the tth frame of the video frame in the first-layer encoding stage, and $\oplus$ is the fusion function. The calculation equation of the hidden layer output structure of group character behavior is as follows:

$$g_t = p_1^t \vee p_2^t \vee \cdots p_N^t, \tag{5}$$

where p_i^t represents the character feature after fusion and $\vee$ is the maximum pooling operation.

In the decoding stage, H_t is output according to the semantic description of the action of the previous character and the hidden layer of the previous moment. Analyze the behavioral semantics of the group characters so that the group characters' behaviors after the GRU structure have temporal sequence information.

The obtained group character behavior semantic prediction probability set is calculated by a maximization equation, and the group character behavior semantics with the largest prediction probability is taken as the video group character behavior semantics:

$$\text{Group} = \arg\max\left(pt_{\text{group}}\right), \tag{6}$$

where pt_{group} is the set of semantic prediction probabilities of group character behavior we obtained.

4. Experiments and Results

4.1. Experimental Setup. The Volleyball data set is selected to verify the semantic extraction method of sparse group behavior based on the spatio-temporal trajectory of video. The Volleyball data set contains 55 real volleyball match videos and 4380 frame labels. The image size of each frame is 720×108, and each frame label contains the number of the current video frame. So, the position information of the player is composed of the coordinates of the upper left corner of the character's bounding box and the height and width of the bounding box. 3493 frame labels of the first 39 videos were used as the IJ} L training set, and 1137 video frames of the last 16 frames were used as the test set.

In the experiment, the length of the input video sequence is T, the individual action and group behavior semantics of N players are extracted, and $T = 10$ is defined. The first 4 frames and the next 5 frames, including the labeled video frames, are, respectively, taken as a video sequence fragment, and $N = 12$ is defined according to the characteristics of the volleyball match in the data set. All experiments in this section were developed using TensorFlow and run on Linux platform.

4.2. Evaluation Standard. In order to extract the semantics of sparse group behavior based on the spatial-temporal trajectory of video, the test is conducted with Volleyball data set and the mask position matching feature F_ code-B is used to complete the matching of people. The experimental results are compared with Inception and HDTM. It includes the comparison of the semantics of human action and the semantics of group action and takes the accuracy of the extracted semantic of group action and the semantic of individual action as the evaluation standard.

4.3. Experimental Results. Table 1 shows the accuracy comparison results of Inception, HDTM, and our algorithm, including two parts: group behavior semantics and individual action semantics. It can be seen from Table 1 that the algorithm in this paper is superior to the above two algorithms in terms of semantics of people's actions and group behaviors. Compared with the above two algorithms, the semantic accuracy of individual actions increases by 4.5% and 2.1%, and that of group actions increases by 8.3% and 1.8%. After integrating the related movement track of the group figures, the complete movement clues of the figures in the video sequence can be grasped by the accurate tracking of the group figures. Figure 3 is an example of successful semantic extraction of character actions and group behaviors in some videos of the data set. In Figure 3, the bounding box information and individual action semantics of each player in this video frame are specifically drawn, and the current group behavior semantics are marked.

4.4. Group Anomaly Recognition Experiments. Aiming at the evaluation of the detection effect of abnormal motion behavior in dense groups, this section uses the PETS 2009 data set containing sequence activities of different groups of people. The data set is divided into five parts: calibration, training, counting, density estimation, and crowd tracking. The video frame image has a resolution of 576×768, contains 9 videos, and has 152 abnormal data. The first 1134 frames of the experiment in this section are used as the training set, and the last 378 frames are used as the test set.

For the detection of abnormal gathering behaviors of dense groups, the experiments in this section are divided into image-level detection and pixel-level detection. Image-level detection gives abnormal aggregation detection results, and pixel-level detection can locate the abnormal gathering place and calculate the number of gatherings. Adopt the same data input and processing methods as the abnormal dispersion behavior.

The detection results of the abnormal dispersion behavior of dense groups on the PETS 2009 data set were quantitatively detected with the DBM algorithm based on optical flow and the D-IncSFA method based on deep learning. The results are shown in Table 2.

Experiments show that this paper has a good detection effect based on the abnormal dispersion behavior of crowd density distribution images and can detect abnormal video frames more accurately. Only when the movement speed exceeds a certain threshold can it be judged as abnormal. The specified movement speed is not less than 0.5 meters per second and more than 1.2 meters is abnormal dispersion behavior. The qualitative evaluation of abnormal dispersion behavior detection in dense groups is shown in Figures 4 and 5.

Table 1: Comparison of the accuracy of semantic extraction.

Methods	Personal action (%)	Group behavior (%)
Inception	78.1	75.5
HDTM	80.2	81.9
Ours	**85.6**	**86.2**

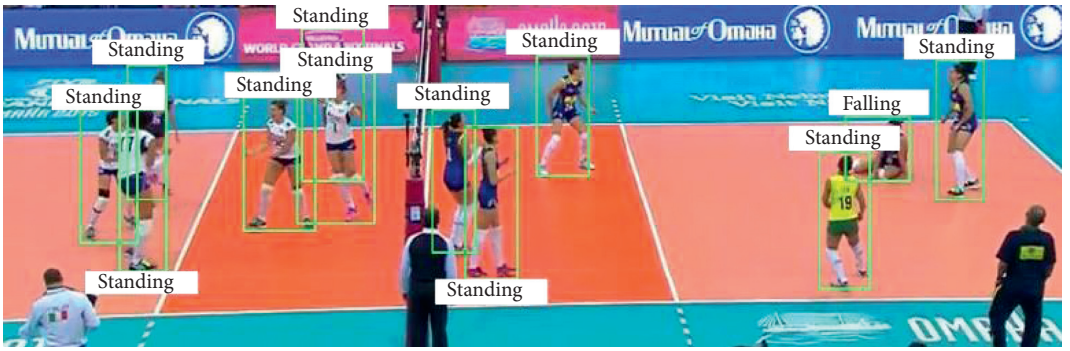
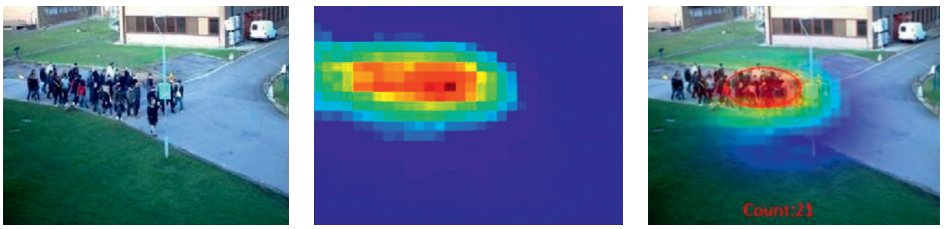


Figure 3: Successful example of group behavior semantic extraction.

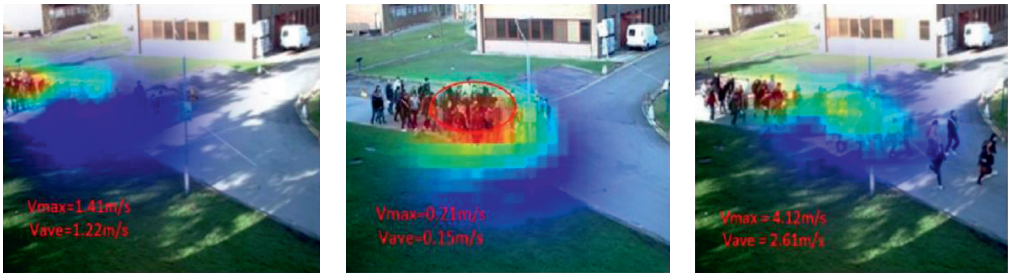
Table 2: Comparison of detection results of abnormal scattered behaviors of dense groups.

Methods	AUC
DBM	0.8770
D-IncSFA	0.9797
Ours	**0.9899**



(a) (b) (c)

Figure 4: Detection results of abnormal crowd behavior. (a) Test image. (b) Population density distribution map. (c) Density analysis and positioning.



(a) (b) (c)

Figure 5: Results of detection of abnormal scattered behaviors of dense groups. (a) Normal behavior, (b) abnormal aggregation behavior, and (c) abnormal dispersion behavior.

5. Conclusion

With the speedy growing population, various crowd activities, and the rapid development of the socialization process, group scenes are becoming more common. Due to this, the demand for analyzing, modeling, and understanding group behavior data in video is increasing. In this paper, we take group density analysis and group behavior recognition in video as the goal and propose a group behavior pattern recognition algorithm based on spatio-temporal graph convolutional network. We designed a crowd detection and positioning method based on density map regression guided classification and, finally, a crowd behavior analysis method based on density level division to complete crowd density analysis and video group behavior detection. In addition, this paper also proposes to use the dual-stream spatio-temporal map network model to extract spatio-temporal features of the crowd posture and density to effectively capture the differentiated movement information between different crowds. We have conducted experiments on public data sets, and the experimental results show that the method has high recognition accuracy and can effectively predict group behavior. The experimental results of the study have shown the effectiveness of the proposed research.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this paper.

Acknowledgments

This work was supported by Special Funds for Basic Scientific Research in Central Universities (ZY20215126) and China Scholarship Fund.

References

[1] Y. Yuan, Y. Lu, and Q. Wang, "Tracking as a whole: multi-target tracking by modeling group behavior with sequential detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3339–3349, 2017.

[2] P. Ramdya, J. Schneider, and J. D. Levine, "The neurogenetics of group behavior in *Drosophila melanogaster*," *Journal of Experimental Biology*, vol. 220, no. 1, pp. 35–41, 2017.

[3] W. Yustisia, I. E. Putra, C. Kavanagh, H. Whitehouse, and A. Rufaedah, "The role of religious fundamentalism and tightness-looseness in promoting collective narcissism and extreme group behavior," *Psychology of Religion and Spirituality*, vol. 12, no. 2, pp. 231–240, 2020.

[4] Z. Pei, X. Qi, Y. Zhang, M. Ma, and Y.-H. Yang, "Human trajectory prediction in crowded scene using social-affinity long short-term memory," *Pattern Recognition*, vol. 93, pp. 273–282, 2019.

[5] U. Singh and M. K. Choubey, "Motion pattern recognition from crowded video," in *Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 431–435, IEEE, Noida, India, June 2020.

[6] W. Wang, J. Shen, X. Lu, S. C. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, In press.

[7] H. Yao, A. Cavallaro, T. Bouwmans, and Z. Zhang, "Guest editorial introduction to the special issue on group and crowd behavior analysis for intelligent multicamera video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 405–408, 2017.

[8] H. Y. Swathi, G. Shivakumar, and H. S. Mohana, "Crowd behavior analysis: a survey," in *Proceedings of the 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, pp. 169–178, IEEE, Bangalore, India, March 2017.

[9] R. Liu, X. Ning, W. Cai, and G. Li, "Multiscale dense cross-attention mechanism with covariance pooling for hyperspectral image scene classification," *Mobile Information Systems*, vol. 2021, Article ID 9962057, 15 pages, 2021.

[10] Q. Liu, L. Cheng, A. L. Jia, and C. Liu, "Deep reinforcement learning for communication flow control in wireless mesh networks," *IEEE Network*, vol. 35, no. 2, pp. 112–119, 2021.

[11] X. Ning, X. Wang, S. Xu et al., "A review of research on co-training," in *Concurrency and Computation: Practice and Experience*John Wiley & Sons, Hoboken, NJ, USA, 2021.

[12] W. Cai, Z. Wei, R. Liu, Y. Zhuang, Y. Wang, and X. Ning, "Remote sensing image recognition based on multi-attention residual fusion networks," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 1–8, 2021.

[13] Q. Luo, "Research on the teaching mode of physical education in colleges and universities," in *Proceedings of the 2017 2nd International Conference on Education, Sports, Arts and Management Engineering (ICESAME 2017)*, pp. 716–719, Atlantis Press, Beijing, China, June 2017.

[14] F. Muñoz-Bullón, M. J. Sanchez-Bueno, and A. Vos-Saz, "The influence of sports participation on academic performance among students in higher education," *Sport Management Review*, vol. 20, no. 4, pp. 365–378, 2017.

[15] Y. Cheng, G. Pang, B. Xiao et al., "Beyond triplet loss: person Re-identification with fine-grained difference-aware pairwise loss," *IEEE Transactions on Multimedia*, 2021, In press.

[16] M. G. Li, B. Jiang, Z. Che et al., "DBUS: human driving behavior understanding system," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2436–2444, IEEE, Seoul, Republic of Korea, October 2019.

[17] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResnetCrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–7, IEEE, Lecce, Italy, August 2017.

[18] A. Almagbile, "Estimation of crowd density from UAVs images based on corner detection procedures and clustering analysis," *Geo-Spatial Information Science*, vol. 22, no. 1, pp. 23–34, 2019.

[19] W. G. Aguilar, M. A. Luna, J. F. Moya et al., "Real-time detection and simulation of abnormal crowd behavior," in *Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pp. 420–428, Lecce, Italy, 2017.

[20] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person Re-identification," in *IEEE Transactions on Circuits and Systems for Video Technology*IEEE, Piscataway, NJ, USA, 2020.

[21] L. Zhou, X. Bai, X. Liu, J. Zhou, and E. R. Hancock, "Learning binary code for fast nearest subspace search," *Pattern Recognition*, vol. 98, Article ID 107040, 2020.

[22] C. Wang, X. Wang, X. Bai, Y. Liu, and J. Zhou, "Self-supervised deep homography estimation with invertibility constraints," *Pattern Recognition Letters*, vol. 128, pp. 355–360, 2019.

[23] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2547–2554, Portland, OR, USA, June 2013.

[24] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1324–1332, 2010.

[25] L. Gao, Y. Wang, X. Ye, and J. Wang, "Crowd counting considering network flow constraints in videos," *IET Image Processing*, vol. 12, no. 1, pp. 11–19, 2017.

[26] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[27] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, Las Vegas, NV, USA, June 2016.

[28] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4031–4039, IEEE, Honolulu, HI, USA, July 2017.

[29] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNS," in *Proceedings of the IEEE international conference on computer vision*, pp. 1861–1870, Venice, Italy, October 2017.

[30] Y. Li, X. Zhang, and D. Chen, "CSRNet: dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, Salt Lake City, UT, USA, June 2018.

[31] Y. Tong, L. Yu, S. Li, J. Liu, H. Qin, and W. Li, "Polynomial fitting algorithm based on neural network," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 32–39, 2021.

[32] Z. Chu, M. Hu, and X. Chen, "Robotic grasp detection using a novel two-stage approach," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 19–29, 2021.

[33] X. Ning, Y. Wang, W. Tian, L. Liu, and W. Cai, "A biomimetic covering learning method based on principle of homology continuity," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 9–16, 2021.

[34] Z. Huang, P. Zhang, R. Liu, and D. Li, "Immature apple detection method based on improved Yolov3," *ASP Transactions on Internet of Things*, vol. 1, no. 1, pp. 9–13, 2021.

[35] Y. Ding, X. Zhao, Z. Zhang, W. Cai, and N. Yang, "Multiscale graph sample and aggregate network with context-aware learning for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4561–4572, 2021.

[36] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, 2020, In press.

[37] A. Feng, Z. Gao, X. Song, K. Ke, T. Xu, and X. Zhang, "Modeling multi-targets sentiment classification via graph convolutional networks and auxiliary relation," *Computers, Materials & Continua*, vol. 64, no. 2, pp. 909–923, 2020.

[38] S. Wang, X. Yu, L. Liu, J. Huang, and T. Jiang, "An approach for radar quantitative precipitation estimation based on spatiotemporal network," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 459–479, 2020.