H-041

Skeleton Motion Based AGCN for Shoplifting Event detection

肇中 汪　　清一郎 鎌田
ZhaoZhong Wang　Seiichiro Kamata

1. Abstract

Identification of crime events occurred in public area (e.g. shoplifting, robbery and vandalism) is a crucial point to prevent the crime economy losses and the property damage. Graph convolution networks(GCN) is one of most common used in action recognition in recent years. Besides, attention with GCN achieve better action representation. In this paper, we make a new dataset in terms of lack of shoplifting events(SE) datasets, then we propose an extended partitioning intent graph adjacent matrix group suitable for suspicious action during SE, we also combine different attention mechanism and propose a new mechanism for spatial and temporal domain. The experiments carried out our model SAT-GACN can enhance suspicious behaviors recognition in our dataset.

2. Introduction

In recent years, according to the National Retail Security Survey, Retail crime, violence and theft continue to impact the retail industry at unprecedented levels, the report represent $112.1 billion in losses in 2022. The identification of SE behaviors has become more essential from the viewpoint of reduce retail economy loss in shop, grocery and etc. Suspicious behavior before a criminal committing theft can provide clues about their impending crime. Therefore, Suspicious behavior detection can effectively prevent and detect theft in public areas, thereby reducing economic losses [1].

Skeleton or joint information has become increasingly prevalent in the field of human action recognition (HAR) due to its view-invariant representation of pose structures. Compared to RGB video data, skeleton data represented as graphs provides a more robust and compact depiction of human movements [19]. This graph-based representation is less affected by viewpoint variations, occlusions, background clutter, inter-class pose variations, lighting conditions, and clothing. Human skeleton data can be acquired using devices such as Kinect cameras and motion sensors or through advanced human pose estimation algorithms [9]. Early skeleton-based action recognition methods primarily use hand-crafted features or use recurrent neural network (RNN), long short-term memory (LSTM) networks, and convolutional neural networks (CNN) to classify vectors or pseudo images from joint data. However, above networks or methods cannot remain the skeleton sequences from natural spatial graph structure and deal with temporal features which input is videos. Thus, these models are weak for suspicious behaviors' spatial and temporal features [6, 9, 12, 15].

In contrast, graph convolution network(GCN) maintain the spatial-temporal skeleton graph structure by representing body joints as vertices and their intra/inter-frame connections as edges. Due to their strong spatial-temporal feature learning capabilities, Suspicious behavior contains so much semantic information than simple action such as walk, run and etc. GCN have been widely adopted for skeleton-based HAR and have emerged as the predominant method in the field. Attention mechanism can also help with the disadvantage of learning temporal information of GCN [7, 11, 18].

Zhaozhong Wang and Sei-ichiro Kamata are with the Image Media Laboratory, Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 8080135, Japan.

3. Related Works

3.1 Shoplifting Dataset

In former researches, shoplifting events (SE) detection based on the human activity problems which mostly use video-label data. These tasks commonly use CNN + LSTM, CNN + GRU and 3DCNN. Additionally, most existing works focus on the whole video, judging whether SE occur or not. And SE datasets are insufficient, researcher's exploration involved two main small datasets: one with 155 shoplifting videos, and another with 175 video clips, comprising 88 clips of normal behavior and 87 clips of shoplifting incidents. Additionally, the UCF-Crime dataset, including a subset of 28 videos recorded from surveillance cameras in retail stores, was also examined [2, 9, 10].

3.2 Graph Convolution Network

Skeleton-based human action recognition has been widely researched using graph convolutional networks. Recent and notable advancements have focused on spatial-temporal graph convolutional networks (ST-GCNs) to effectively capture motion and temporal dependencies in video sequences. The traditional ST-GCN architecture (Shown in Figure 1) consists of a series of ST-GCN blocks, which apply spatial graph convolutions and temporal graph convolutions alternately on a skeleton graph. Ultimately, fully connected dense layers followed by a SoftMax classifier are used to predict the action class [16].

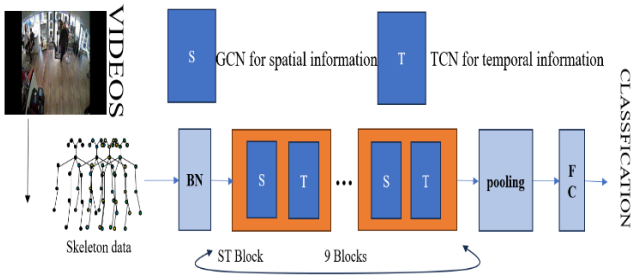

Figure 1. ST-GCN baseline

3.3 Attention based GCN

Chen et al. [3] introduced CTR-GCN, which dynamically refines a shared prior topology for each feature channel. This method creates multi-channel attention maps to enhance the correlation between joints in each skeleton graph, resulting in a more robust representation through effective modeling across different channels. Zhang et al. [20] proposed a Spatial-Temporal Specialized Transformer Encoder (STST), designed to model the skeleton posture of each frame and capture temporal changes in posture, thereby providing a strong framework for action sequence modeling. Similarly, various approaches [13, 14] utilize the transformer architecture to extract spatial and temporal dependencies (ST-TR). Recently, Chi et al. [4] employed the information bottleneck approach to define the objective and corresponding loss, aiming for maximum informative latent representation of skeleton-based actions.

In this paper, for the classification of suspicious behaviors during SE, we made a new six-category dataset, and use ST-GCN as our baseline. We propose a model SAT-GACN for our dataset, the proposed method for partitioning strategy is adaptive to suspicious behaviors and we also made novel attention mechanism to help our model better understand spatial temporal information.

4. Proposed Method

4.1 Adaptive Partitioning Strategy (APS)

Most existing state-of-the-art skeleton-based ST-GCNs focus on the design of mapping functions that categorize joints into three distinct groups. The scheme performs effectively but solely takes into account the information from adjacent joints, neglecting the information from distant joints and those that are physically disconnected. For motion change during SE, not only adjacent joints, but also hands joints and even further joints will make sense to classification. Therefore, we propose an extended adaptive partitioning strategy for suspicious behaviors occurred during SE [2, 17].

There are three types of partitioning mapping functions: uni-labeling, distance, and spatial. Spatial Strategy works better than the other two methods. Most researches will divide human joints into three subgroups: 1) root node/center node: every joint itself 2) centripetal subgroup: the subsets of neighboring skeleton joints that are closer to neck joint than central joint3) centrifugal subgroup: the subsets of neighboring skeleton joints that are further from neck joint than central joint. Conventional partitioning schemes is shown in Figure. The ST-GCN designates the neck node as the skeleton's center of gravity. However, this partitioning scheme focuses only on a limited number of nodes closest to the central point, neglecting information from more distant nodes.

We propose a better partitioning strategy APS by roughly dividing motion of skeleton joints into the intention to neck joint, right hand joint and left hand joint. Suppose we have a graph G = (V, E) which is shown in Figure 1(a), where V is the set of nodes and E is the set of edges. We also have a center node C, a neck joint node N, a right-hand joint node R, and a left-hand joint node L. Using these symbols, we can define the seven groups as follows:1. This group consists of only the center node C shown in Figure 1(b) 2. neighbor nodes who closer to neck joint 3. neighbor nodes who further from neck joint shown in Figure 1(c) 4. neighbor nodes who closer to right hand joint 5. neighbor nodes who further from right hand joint 6. neighbor nodes who closer to left hand joint 7. neighbor nodes who further from left hand joint Figure 1(d).

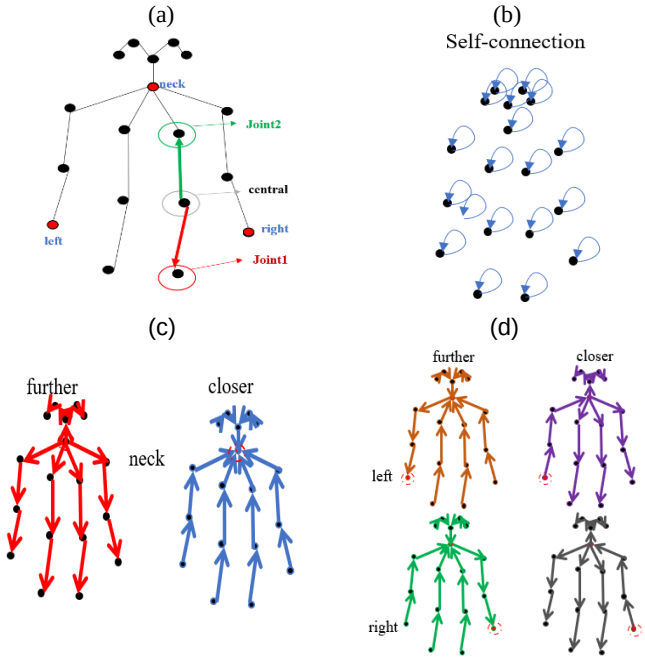

Figure 2. Adaptive partitioning strategy (APS)

4.2 SAT-GACN

After using adaptive partitioning strategy between GCN and TCN, our model is turned into ST-GACN. Graph Convolutional Networks (GCNs) help us learn the local features of joints that are spatially adjacent. Building on this, we need to learn the local features of joint movements over time. Integrating temporal features into the graph is one of the challenges faced by graph networks. Popular attention mechanism for GCN like important edges, Squeeze-and-Excitation (SE) [5] and Cosine Similarity. To better understand spatial and temporal features and enhance the weak point of temporal features learning. We modify the ST-GCN block as SAT-GACN, it is shown in Figure 3. We add a TSC attention block, the attention module contains a temporal attention module, a spatial attention module and a channel attention module. The temporal attention module focusing on attention among continue frames and let the model capture long and

complex semantic action time features, spatial attention module assigns single frame and help the model with spatial information and channel attention module is used to enhance the discriminative ability of channel features for different input samples. It achieves this by assigning different attention weights to each channel, thereby highlighting the most important channel features.

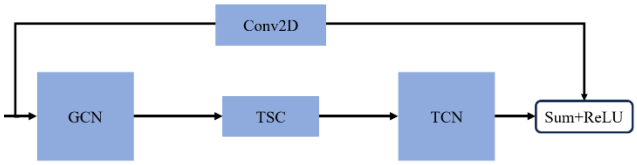

Figure 3. TSC block

After GCN model, we get output vertex as TSC input.

$$X_{gcn} \in R^{C_{gcn} \times T \times N} \quad (1)$$

Our TSC block first process with Temporal attention module. Performing average pooling across the temporal dimension T to reduce the dimensionality, then calculating temporal attention score A_t applied to X_{gcn}.

$$A_t = \sigma(F_t(AvgPool(X_{gcn}))) \quad (2)$$

$$X_{ta} = X_{gcn} \odot A_t + X_{gcn} \quad (3)$$

Then we calculate the spatial attention score A_s and apply it to X_{ta}.

$$A_s = \sigma(F_s(AvgPool(X_{ta}, \mathrm{dim} = 1))) \quad (4)$$

$$X_{sa} = X_{ta} \odot A_s + X_{ta} \quad (5)$$

Finally, we calculate channel attention score A_c apply to X_{ca}. W_{c2} and W_{c1} are learnable parameters.

$$A_s = \sigma(W_{c2}(\delta\left(W_{c1}(AvgPool(X_{ta}))\right))) \quad (6)$$

$$X_{ca} = X_{sa} \odot A_c + X_{sa} \quad (7)$$

5. Experiments

5.1 Dataset and evaluation matrix

A self-made multi-class dataset, medium size dataset which contain six types of suspicious behaviors drained from shoplifting events mostly occurred in shop, mall or grocery which details are shown in Table 1, each type of suspicious actions contains about 50 videos, frame rate is 30 and video size is 320*240, video source is from UCF-Crime dataset and Youtube videos.

TYPE	Description
Pacing back and forth	In special region or before special goods and items, stand for a long time or pace back and forth.
Scanning the store	Watching around, other customers and stuff.
Spotting CCTVs	Watching cameras.
Hiding items	Hiding goods in store.
Standing for long time	Stay somewhere for a long time.
Leaving the store fast	Quickly leaving the store.

Table 1. Description of each suspicious behavior

Because we only have six types of suspicious behaviors , evaluating recognition performance judging which kind of suspicious behaviors we use Top-1 accuracy, Accuracy is calculated by calculating the proportion of model predictions matching actual labels. During the testing phase, the model outputs and actual labels are used to calculate accuracy, and loss function uses mean loss function.

5.2 Training details

Hardware platform parameters is shown as follow, GPU – single Nvidia GeForce RTX 3090, python 3.7 with pytorch 1.11.2, CPU- i7-7700, 32GB of RAM. Our model is trained using the Adaptive Partitioning Strategy and weight decay of 0.0001. The batch size is 32, and the number of iterations is set to 50. The initial learning rate is 0.1, and as number of every decade epoch, the learning rate is reduced by a factor 0.1 at each step, continuing the iterations.

5.3 Comparison experiment with ST-GCN

To validate that our model structure exhibits superior spatiotemporal information modeling capabilities compared to ST-GCN, a comparative experiment was conducted on our six-category dataset using the APS partitioning method. Results are shown in Figure 4.

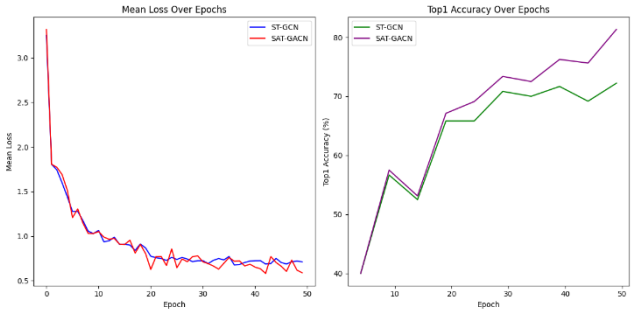

Figure 4. ST-GCN vs. SAT-GACN

5.4 Ablation study

To indicate the recognition accuracy and generalization performance of our model, every contrast model is trained and test on the six category suspicious behaviors dataset.

Four different partitioning strategy (uniform, distance, spatial, adaptive) are used in experiments, and the experiment results are compared without partitioning strategy. According to the experiment results, ST-GCN with adaptive partitioning strategy exhibits the best performance. The experiment results are presented in Table 2, validating that the introduction of partitioning mechanisms ensures the model better feature learning.

Strategy	Mean_loss	Acc
None	0.712	70.29%
Uniform	0.682	71.17%
Distance	0.622	73.44%
Spatial	0.609	76.24%
APS	0.597	79.81%

Table 2. Different Strategies contrast

From Table 2, Comparing to ST-GCN, our model's accuracy has been improved. the accuracy of adaptive strategy is 9.52% higher than original, 8.64% higher than uniform strategy, 6.37% higher than distance strategy, 3.57% higher than spatial strategy.

To verify our attention-based model advantage on time series, we compared the performance with four different attention mechanism on our own dataset and the experiment results are shown in Table 3.

Attention mechanism	Mean_loss	Acc
None	0.731	69.24%
Important edges	0.632	72.17%
SE	0.617	74.28%
LSTM	0.619	73.24%
TSC	0.601	77.17%

Table 3. Different attention modules

Without any attention mechanism, the model has the highest mean loss (0.731) and the lowest accuracy (69.24%). Important edges: The attention mechanism focusing on important edges significantly reduces the mean loss (0.632) and improves the accuracy (72.17%). SE (Squeeze-and-Excitation): Using the SE attention mechanism further reduces the mean loss (0.617), achieving an accuracy of 74.28%. LSTM: When using the LSTM attention mechanism, the mean loss is 0.619, and the accuracy is (73.24%), which is slightly lower than SE. TSC (Temporal Spatial Convolution): The TSC attention mechanism performs the best, with the lowest mean loss (0.601) and the highest accuracy (77.17%).

6. Conclusion

In this paper, we introduce a SAT-GACN model for suspicious behaviors recognition during shoplifting events. Firstly, we use existing datasets and online source edit suspicious behaviors clips, then we introduce our adaptive partitioning strategy which can indicate intention of conducting thief during shoplifters' action. Finally, we made our own TSP attention mechanism to help our model to deal with such long and complex semantic information. From the result of experiment, it indicates that our model and algorithm can achieve a better performance in this specific dataset.

Acknowledgement

The authors declare no conflict of interest. Chenxi Yu: Conceptualization, Methodology, Research, Experiment, Validation. Sei-ichiro Kamata: Conceptualization, Research, Validation. I would like to express my great appreciation to my supervisor Sei-ichiro Kamata for his consecutive suggestions.

Reference

[1] A. Ben Mabrouk, E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: A review, Expert Syst. Appl. 91 (2018) 480–491, https://doi. org/10.1016/j.eswa.2017.09.029
[2] Ansari, M.A.; Singh, D.K. ESAR, An Expert Shoplifting Activity Recognition System. Cybern. Inf. Technol. 2022, 22, 190–200.
[3] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13359–13368 (2021)
[4] Chi, H.G., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: INFOGCN: representation learning for human skeleton-based action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20186–20196 (2022)
[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141, DOI: 10.1109/CVPR.2018.00745.
[6] Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1623–1631 (2017)
[7] Kirichenko, L.; Radivilova, T.; Sydorenko, B.; Yakovlev, S. Detection of Shoplifting on Video Using a Hybrid Network. Computation 2022, 10, 199
[8] Koniusz, P., Cherian, A., Porikli, F.: Tensor representations via kernel linearization for action recognition from 3D skeletons. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 37–53. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_3
[9] K. Xu, F. Ye, Q. Zhong and D. Xie, "Topology-aware convolutional neural network for efficient skeleton-based action recognition", Proc. AAAI Conf. Artif. Intell. (AAAI), vol. 36, pp. 2866-2874, 2022.
[10] K. Yang, X. Ding and W. Chen, "Attention-based generative graph convolutional network for skeleton-based human action recognition", Proc. 3rd Int. Conf. Video Image Process., pp. 1-6, Dec. 2019.
[11] Martínez-Mascorro, G.A.; Abreu-Pederzini, J.R.; Ortiz-Bayliss, J.C.; Garcia-Collantes, A.; Terashima-Marín, H. Criminal intention detection at early stages of shoplifting cases by using 3D convolutional neural networks. Computation 2021, 9, 24.
[12] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3590-3598, Jun. 2019.
[13] Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, Part III, pp. 694–701 (2021)
[14] Qin, X., Cai, R., Yu, J., He, C., Zhang, X.: An efficient self-attention network for skeleton-based action recognition. Sci. Rep. 12, 2045–2322 (2022)
[15] S.-E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, "Convolutional pose machines", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4724-4732, Jun. 2016.
[16] S. Yan, Y. Xiong and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition", Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, pp. 1-10, 2018, [online] Available: https://ojs.aaai.org/index.php/AAAI/article/view/12328.
[17] W. Zheng, P. Jing and Q. Xu, "Action recognition based on spatial temporal graph convolutional networks", Proc. 3rd Int. Conf. Comput. Sci. Appl. Eng. (CSAE), vol. 118, pp. 1-5, 2019.
[18] X. Gao, W. Hu, J. Tang, J. Liu and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression", Proc. 27th ACM Int. Conf. Multimedia, pp. 601-610, Oct. 2019.
[19] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun, "Cascaded pyramid network for multi-person pose estimation", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 7103-7112, Jun. 2018.
[20] Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In: ACM International Conference on Multimedia, pp. 3229–3237 (2021)