Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Abnormal Behavior Recognition Method Based on Fine-Grained Human Dynamic Skeleton Features

Chenyue Xu, Rong Wang *

Department of Information and Network Security, People's Public Security University of China, Beijing, China
*** Corresponding author:** Rong Wang (Email: dbdxwangrong@163.com)

Abstract: To address the insufficiency of feature extraction in existing video anomaly recognition methods, this paper proposes an improved graph-embedded pose clustering-based abnormal behavior recognition method. By employing fine-grained feature extraction, it enhances information representation capabilities and improves algorithm robustness. The framework is structured as follows: First, spatial and channel reconstruction convolutions are integrated into a Deep Residual Network backbone, effectively eliminating spatial-channel redundancy in extracted features while reducing computational complexity, thereby enhancing operational efficiency. Second, a hierarchical decomposed graph convolutional network (modified from spatio-temporal graph convolutional networks) is implemented for dynamic skeleton feature extraction, coupled with an attention-guided hierarchical aggregation module for multi-level feature fusion. Evaluations on ShanghaiTech and NTU-RGB+D datasets demonstrate detection accuracies of 76.4% and 74.6% respectively, validating the method's effectiveness.

Keywords: Abnormal Behavior Recognition; Dynamic Skeletal Features; Fine-Grained Feature Extraction.

1. Introduction

Abnormal behavior recognition [1-4], a prominent research direction in computer vision, involves identifying key frames in videos and determining the presence of anomalous activities. Practical challenges arise from massive video volumes and diverse anomaly categories, while existing frame-level annotated datasets remain limited in scale—insufficient for training supervised video anomaly recognition models. Sato et al. [5] leveraged pre-trained skeletal feature extractors, aligning skeleton features with user prompts through common spatial domain alignment to establish anomaly scores. Cho et al. [6] proposed an implicit dual-path autoencoder to enhance feature informativeness through differentiated learning strategies.

However, unsupervised anomaly detection methods [7,8] often suffer from weak feature expressiveness and redundant representations. To address these limitations, we enhance the Graph-Embedded Pose Clustering (GEPC) framework by introducing dynamic human skeleton sequences as model inputs. Our proposed solution combines:

2. Organization of the Text

2.1. Model Framework

Building upon the GEPC framework, we have optimized both the feature extraction backbone network and the spatio-temporal graph convolution module. The enhanced model demonstrates superior capability in extracting comprehensive global and spatio-temporal features from human pose graph sequences, significantly improving abnormal behavior recognition accuracy.

The improved model can be divided into three modules: the pose estimation feature enhancement module, the encoding module, and the clustering module. The overall structure is shown in Figure 1. In the pose estimation module, the model employs the AlphaPose module to extract dynamic human skeleton graphs from video frames. To extend the human skeleton graphs from the spatial dimension to the temporal dimension, spatio-temporal feature vectors of the dynamic skeleton graphs are extracted. In the encoding module, an improved deep temporal graph autoencoder structure is proposed to embed the dynamic human skeleton graphs. The encoder is designed based on spatio-temporal graph convolution, and the original model is optimized using an attention-guided hierarchical method for extracting features from dynamic skeleton graphs, constructing a hierarchically decomposed spatio-temporal graph convolutional autoencoder. In the clustering module, the training set is first jointly embedded into a latent space and subjected to clustering operations to build an underlying action dictionary. Then, each sample is represented by a probability distribution over the clustered underlying actions. The clustering module consists of three parts: an encoder, a decoder, and a soft clustering layer. The encoder preserves the input graph structure and compresses the dynamic skeleton graph sequence into a latent vector through convolution. The decoder then uses temporal upsampling layers and additional graph convolution blocks to gradually restore the original channel count and temporal dimensions. Finally, the initial reconstruction-based embeddings are fine-tuned during the clustering optimization stage to achieve the final optimized clustering embeddings.

2.2. Feature Preprocessing Network

The Spatial and Channel Reconstruction Convolution (SCConv) [9] is introduced into the ResNet residual network structure, replacing the original 3×3 convolutional network. By utilizing two sub-modules, the Spatial Reconstruction Unit (SRU) and the Channel Reconstruction Unit (CRU), global feature extraction is performed, effectively reducing redundancy in intermediate feature maps. This approach decreases the number of parameters and computational load while enhancing the overall performance of the model.

In the spatial feature refinement stage, the SRU sub-module employs a separation and reconstruction joint operation to eliminate spatial redundancy in the features. In the channel feature refinement stage, to address the channel

redundancy in feature maps caused by k×k standard convolution, the CRU sub-module uses a split-transform-merge method to remove redundant channel features. The k×k standard convolution can be expressed as $Y = M^k X$, where $M^k \in \mathbb{R}^{c\times k\times k}$ represents the convolution kernel of the k×k standard convolution, and $X, Y \in \mathbb{R}^{c\times h\times w}$ denote the input and output features of the convolution operation, respectively.

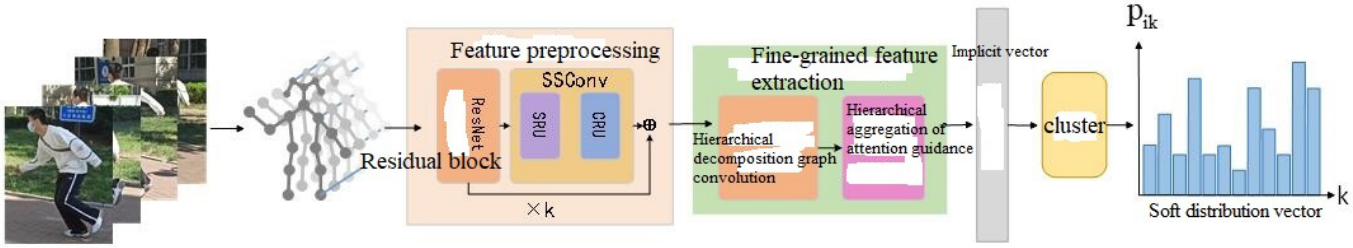

Fig 1. Overall Structure

In the split operation, first, the spatially refined features output from the SRU sub-module are divided into two parts, αC and $(1-\alpha)C$, along the channel dimension according to the split ratio $\alpha \in (0,1)$. Second, a 1×1 convolution is used to compress the channel dimensions of the feature maps to improve computational efficiency. After the splitting and compression operations, the spatially refined features can be divided into upper and lower parts. In the merge operation stage, a simplified selective kernel network is utilized to adaptively combine the output features from the upper and lower transformation stages.

2.3. Fine-Grained Human Dynamic Skeleton Feature Extraction

The proposed method introduces Hierarchically Decomposed Graph Convolutional Networks (HDGCN) [10] based on spatio-temporal graph convolutional networks, focusing on the relationships between distant joints and the features of key edges, effectively improving the overall performance of the model. HD-GCN consists of two modules: a graph module that establishes hierarchical decomposition levels and an attention-guided hierarchical aggregation module. The improved model extracts fine-grained human dynamic skeleton features, enhancing the expressive capability of the extracted features and significantly boosting the overall performance of the model.

2.3.1. Fine-Grained Feature Extraction

The spatial graph convolution module of the hierarchically decomposed spatio-temporal graph convolution introduces a hierarchically decomposed graph convolution module. This module divides the human skeleton graph into three levels and extracts spatial features of the human skeleton for each level separately. During the construction of the hierarchically decomposed graph, a tree-like graph with a root node is first established based on the joints and physically connected edges of the human dynamic skeleton graph.

The hierarchically decomposed graph convolution module has four parallel computational branches, including three graph convolution branches and an additional edge convolution branch. To reduce computational complexity, all four branches undergo linear transformation. In the three graph convolution branches, the same graph convolution operation is performed on each hierarchical edge set containing three edge subsets, and the output values are concatenated along the channel dimension. To extract sample-level edge association information reflecting the similarity between nodes in the feature space, the module employs Edge Convolution (EdgeConv) [11], which extracts features of the human dynamic skeleton graph through local neighborhood graphs in the feature space. First, the edge convolution branch uses average pooling to improve computational efficiency. Then, the k-nearest neighbor algorithm is applied based on Euclidean distance to obtain the adjacency edge set.

2.3.2. Attention-Guided Feature Fusion

After the hierarchically decomposed graph convolution module, human skeleton features extracted from different levels are obtained. For the fusion of features at different levels, this paper uses an attention-guided hierarchical aggregation module. This feature fusion module focuses on the relationships and importance of key edges, applying an attention-weighted strategy to the combined outputs of different levels, effectively enhancing the expressive capability of the extracted features.

The input to the attention-guided hierarchical aggregation module is the output of the graph convolution operation. For this output, the frame with the highest score is extracted. First, considering that each level has representative key points, a Representative Spatial Average Pooling (RSAP) layer Ψ is introduced after the multi-level feature extraction output to avoid scaling bias during computation, assigning higher weights to representative key points. Then, hierarchical edge convolution is applied after the representative spatial pooling layer, fusing the human skeleton graph features optimized by the attention module at each level into an overall feature.

2.4. Clustering

During the clustering phase, the hierarchically decomposed spatio-temporal graph convolutional model obtains spatio-temporal features of human dynamic skeletons containing rich information from training samples and constructs an underlying action dictionary. In the clustering phase, the model maps the learned features to a latent space and embeds them into clusters, calculating the probability of video sequences containing abnormal behaviors through clustering.

3. Experiments

3.1. Experiments Setting

The experimental design is divided into two parts. First, the model conducts fine-grained abnormal behavior recognition experiments on the ShanghaiTech dataset. The experiment is divided into training and testing phases, where the training dataset includes only normal samples, while the testing dataset includes both normal and abnormal samples. For each input video sequence, the model uses a sliding window approach to crop the sequence into fixed-length segments and assigns a score to each video frame. In cases where multiple individuals appear in a video frame, the model scores each individual separately and takes the highest score as the

anomaly score for that frame. Then, the model performs coarse-grained abnormal behavior recognition experiments on the NTU-RGB+D dataset. During the training phase, the model is trained on an unsupervised dataset containing only normal samples. In the testing phase, the model distinguishes between normal and abnormal samples by measuring the difference between each test sample and the training samples.

3.2. Ablation Study

Table 1. Ablation Experiment Results

Method	ShanghaiTech	NTU-RGB+D
GEPC	0.752	0.730
GEPC+SSConv	0.760	0.741
GEPC+HDGCN	0.757	0.738
GEPC+SSCOnv+HDGCN	0.764	0.746

From the table above, we can observe that when training with the ShanghaiTech and NTU-RGB+D datasets, the addition of the human dynamic skeleton feature preprocessing network enhances the expressive capability of feature information, leading to an overall improvement in model performance. After incorporating the fine-grained human dynamic skeleton feature extraction module, the model's ability to extract information from human dynamic skeleton features is enhanced compared to the pure spatio-temporal graph convolutional network model, resulting in an overall performance improvement. Therefore, this ablation experiment demonstrates the effectiveness of both individual modules and the fused modules for the improved GEPC network model.

3.3. Comparative Experiments

Table 2. Comparative Experiment Results

Method	ShanghaiTech	NTU-RGB+D
PGM [12]	0.612	-
ASTNet[13]	0.736	-
Proposed Method	0.764	0.746

From the table above, compared to PGM, which proposes a video anomaly detection method based on target bounding box probability analysis, improving anonymity and computational efficiency by simplifying data representation and making it suitable for edge devices, and ASTNet, which introduces a spatio-temporal dual-branch residual autoencoder network achieving unsupervised video anomaly detection through channel attention modules and temporal shift methods, the hierarchically decomposed graph convolutional network used in this paper can extract dynamic human skeleton features in a fine-grained manner, leading to improved accuracy in abnormal behavior recognition on both the ShanghaiTech and NTU-RGB+D datasets.

3.4. Visualization

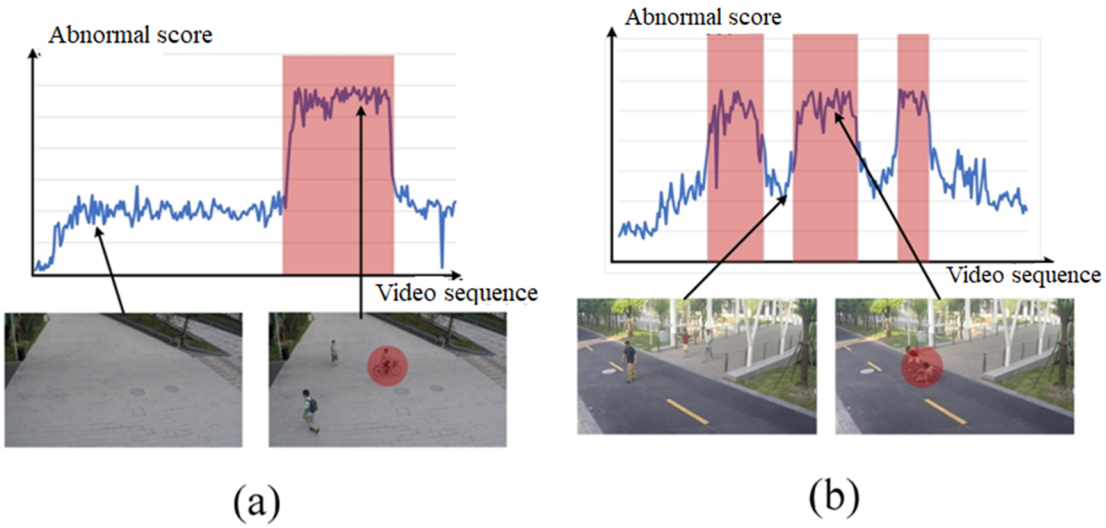

Fig 2. Visualization

The visualization results of the model on the ShanghaiTech dataset are shown in Figure 2, where the red parts represent the originally annotated abnormal frames in the dataset, and the blue line represents the predicted anomaly scores of the video frames. The visualization results demonstrate that the model assigns higher anomaly scores to video frames containing abnormal behaviors and lower anomaly scores to normal video frames, proving the effectiveness of the model.

4. Summary

The proposed method improves upon the GEPC model by first enhancing the residual modules of the backbone network ResNet through the introduction of the SSConv module, thereby improving the model's ability to express the extracted

human dynamic skeleton features. Then, by incorporating a hierarchically decomposed graph convolutional network, multi-granularity extraction of human dynamic skeleton features is achieved, effectively enhancing the expressive capability of the final human dynamic skeleton features. Experiments conducted on the ShanghaiTech and NTU-RGB+D datasets demonstrate that the proposed method significantly improves the detection accuracy of abnormal behavior recognition methods.

References

[1] Verma, Kamal Kant, Brij Mohan Singh, and Amit Dixit. "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system." International Journal of Information Technology 14.1 (2022): 397-410.

[2] Pogadadanda, Vikas, et al. "Abnormal activity recognition on surveillance: a review." 2023 third international conference on artificial intelligence and smart energy (ICAIS). IEEE, 2023.

[3] Qiu, Yuting, James Meng, and B. J. J. I. Li. "Automated falls detection using visual anomaly detection and pose-based approaches: experimental review and evaluation." J ISSN 2766 (2024): 2276.

[4] Duong, Huu-Thanh, Viet-Tuan Le, and Vinh Truong Hoang. "Deep learning-based anomaly detection in video surveillance: A survey." Sensors 23.11 (2023): 5024.

[5] Sato, Fumiaki, Ryo Hachiuma, and Taiki Sekii. "Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[6] Cho, MyeongAh, et al. "Unsupervised video anomaly detection via normalizing flows with implicit latent features." Pattern Recognition 129 (2022): 108703.

[7] Coşar, Serhan, et al. "Toward abnormal trajectory and event detection in video surveillance." IEEE Transactions on Circuits and Systems for Video Technology 27.3 (2016): 683-695.

[8] Chen, Chunyu, Yu Shao, and Xiaojun Bi. "Detection of anomalous crowd behavior based on the acceleration feature." IEEE sensors journal 15.12 (2015): 7252-7261.

[9] Li, Jiafeng, Ying Wen, and Lianghua He. "Scconv: Spatial and channel reconstruction convolution for feature redundancy." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[10] Lee, Jungho, et al. "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[11] Lee, Minhyeok, et al. "Edgeconv with attention module for monocular depth estimation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[12] Siemon, Mia, et al. "Bounding boxes and probabilistic graphical models: video anomaly detection simplified." arXiv preprint arXiv:2407.06000 (2024).

[13] Le, Viet-Tuan, and Yong-Guk Kim. "Attention-based residual autoencoder for video anomaly detection." Applied Intelligence 53.3 (2023): 3240-3254.