Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 2, 2025

Deep Learning-Based Video GIS Behavior Recognition Method Using Cesium

Youwei Jia

Henan Polytechnic University, Jiaozuo, Henan, China

Abstract: The Cesium Video GIS Behavior Recognition Method based on the SlowFast network combines video analysis with Geographic Information Systems (GIS), aiming to enhance behavior recognition in smart city and public safety applications. By utilizing the Cesium platform for video data visualization and spatial data management, and leveraging the dual-path characteristics of the SlowFast model, dynamic behavior features in videos are extracted. The SlowFast model effectively captures video information at different speeds, recognizing behavior patterns with both rapid and slow changes. This method improves the accuracy and robustness of behavior recognition, making it suitable for long-duration video streams and high-resolution data, providing more precise decision support for intelligent monitoring and traffic management.

Keywords: SlowFast model; Cesium; Video GIS.

1. Introduction

Behavior recognition technology, as a key research direction in computer vision and deep learning, holds significant importance in solving critical issues in social management, public safety, and intelligent development. Through accurate recognition and analysis of individual or group behaviors, it not only enhances public safety, such as in traffic management and urban surveillance for timely detection of abnormal behavior, but also optimizes social resource allocation and improves urban governance efficiency. In addition, behavior recognition has broad application prospects in the healthcare field. By monitoring the daily activities of elderly people or patients, it enables remote healthcare and intelligent early warning systems, providing more efficient protection and support for vulnerable groups. Behavior recognition is widely applied in public safety, traffic management, smart cities, healthcare, industrial monitoring, and virtual reality. For example, in intelligent transportation systems, behavior recognition can be used to analyze drivers' driving status and pedestrians' crossing behaviors, reducing the occurrence of accidents. In smart cities, recognizing crowd movement patterns can optimize urban infrastructure layout. In virtual reality, behavior recognition technology enhances user interaction and promotes the naturalization and intelligence of human-computer interaction.

Video Geographic Information Systems (Video GIS) is an emerging research direction combining geographic information system (GIS) technology with video data, enabling traditional GIS to have dynamic perception and real-time analysis capabilities. With the rapid development of video capture devices, drones, and satellite technology, massive video data can be deeply integrated with spatial information, providing strong support for research across multiple fields. Unlike traditional static GIS, Video GIS not only presents the distribution and features of geographic entities but also captures dynamic events, behavior trajectories, and change processes, showing great potential in emergency management, environmental monitoring, and smart city construction.

In terms of application potential in behavior recognition, Video GIS provides richer semantic information by integrating temporal, spatial, and behavioral features from video data. For example, by analyzing video surveillance data in geographical spaces, real-time monitoring of crowd gathering, traffic flow, and abnormal behaviors can be achieved. In smart cities, Video GIS combined with behavior recognition technology can enable fine management, such as monitoring behavior patterns in crowded places and optimizing resource allocation. In disaster emergency response, behavior analysis based on Video GIS can quickly locate the distribution and escape routes of affected populations, thus improving rescue efficiency. With the development of deep learning and computer vision technologies, Video GIS provides broader application space for behavior recognition, making it an important direction for future technological integration and innovation.

SlowFast is a deep learning-based model specifically designed for video behavior recognition tasks. It is an extension of convolutional neural networks (CNNs) and employs a unique dual-path structure that combines input streams at different frame rates to effectively process video data. The design inspiration comes from the hierarchical perception mechanism of human visual systems for fast and slow motion. The model works by coordinating two branches: the "Slow" branch captures long-term semantic information at a lower frame rate, such as the overall motion process; the "Fast" branch focuses on short-term dynamic details, such as transient features of fast movements. These two branches extract features at different time scales, and their information is fused through cross-path connections, allowing the model to efficiently capture temporal dynamic features while retaining rich contextual information.

Cesium is a web-based open-source 3D geospatial visualization platform, and its high-performance 3D rendering engine and broad compatibility have established its importance in geospatial data visualization. As a framework supporting large-scale geospatial data interaction, Cesium efficiently loads and renders various types of geospatial data, including vector data, raster data, 3D models, and time-varying data. Its support for open standards (such as 3D Tiles and GLTF) and cross-platform capabilities have made it widely applied in smart cities, drone remote sensing, and dynamic event monitoring.

Cesium demonstrates unique advantages when integrated with video data. By mapping video data onto geospatial environments, Cesium can present video content within real-world geographic contexts. For example, in drone aerial photography or satellite monitoring scenarios, videos can be embedded into 3D terrain models for visual analysis. This fusion not only enhances the intuitiveness of geographic information expression but also strengthens the interaction between video data and spatial information. In behavior recognition applications, Cesium provides an efficient visualization platform that can dynamically display individual or group behavior patterns in videos, and through time and space dynamic analysis, help users detect abnormal behaviors, monitor crowd activity paths, or conduct real-time disaster response evaluations.

The integration of video data with the Cesium platform drives the evolution of GIS towards dynamic and intelligent systems, providing innovative solutions for smart city management, public safety monitoring, and emergency response. As deep learning technologies, especially behavior recognition models like SlowFast, are integrated, Cesium's potential in the Video GIS domain will be further explored, offering broader development space for interdisciplinary research.

2. Related Work

2.1. SlowFast Model

2.1.1. SlowFast Model and Its Features

The SlowFast model is a deep learning architecture designed for video understanding, particularly suited for processing video data with multi-scale spatial-temporal information. The core idea of the model is to divide the processing of video frames into two different-speed streams, namely the "slow" and "fast" streams, to effectively capture both long-term and short-term temporal information in the video, enabling precise understanding of video content. The slow pathway network is responsible for extracting long-term temporal information from low-frame-rate videos, focusing on global dynamic changes, while the fast pathway network processes the video at a higher frame rate to capture local fast changes, able to detect rapid motions and detailed variations.

The slow pathway network uses low-frame-rate input, typically processing only a few frames per second, which reduces computational load and focuses on capturing global, long-term spatial-temporal structural features in the video. This design allows the network to effectively analyze large-scale motion changes, such as the overall motion of the human body or the trajectory of objects. In contrast, the fast pathway network uses high-frame-rate video input, processing more frames per second to capture the finer details and fast changes, such as quick movements or instantaneous changes. The high-frame-rate design of the fast pathway enables the network to handle fast motion or fine spatial-temporal changes, compensating for the slow pathway's lack of information on short time scales.

In terms of network architecture, the SlowFast model uses a parallel design for the slow and fast pathways, processing low-frame-rate and high-frame-rate inputs respectively, and then fusing the features of both. Specifically, the slow and fast networks typically share some lower-layer convolutional layers, but the fast pathway's convolutional layers use a higher time resolution, while the slow pathway operates at a lower time resolution. This design allows the SlowFast model to efficiently compute while maximizing the extraction of spatial-temporal information from the video, enabling flexible capture of various movement patterns.

The parallel design of the slow and fast streams allows the SlowFast model to capture spatial-temporal features at different scales, making it adaptable to a variety of complex video understanding tasks, such as action recognition and video classification. In practical applications, the combination of the slow and fast streams not only improves the model's ability to capture long-term sequence information but also enhances its responsiveness to short-term rapid events, demonstrating excellent performance across multiple video analysis tasks. The SlowFast network architecture is shown in the figure below:

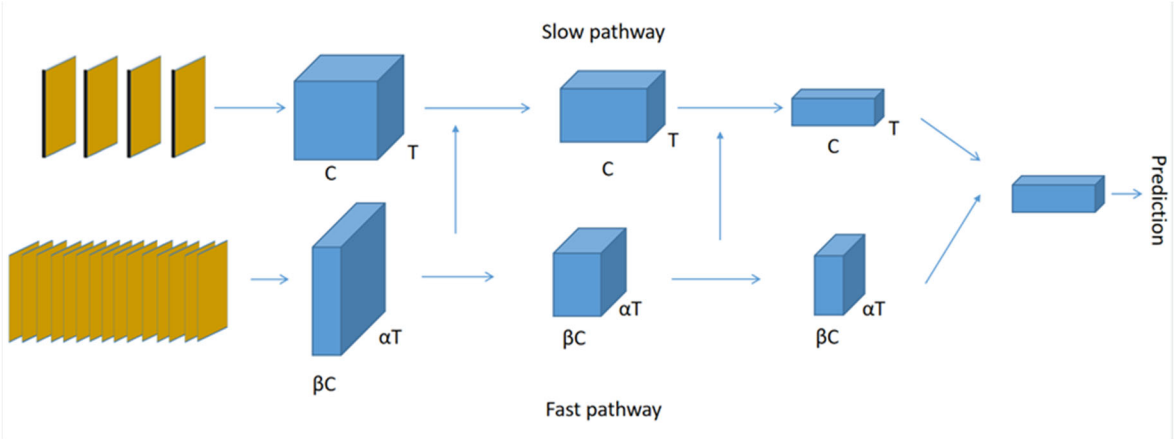

Figure 2-1. Schematic Diagram of the SlowFast Model Structure

2.1.2. SlowFast Model Training Process

The training process of the SlowFast model is designed to leverage the multi-scale spatial-temporal information in videos through the parallel slow pathway and fast pathway networks, significantly improving the model's performance in video understanding tasks. The training involves key technical details such as data preprocessing, network architecture configuration, loss function design, and optimization strategies. Each component is carefully designed to ensure the model can efficiently capture long-term temporal dependencies and short-term temporal information in videos while maintaining good computational efficiency to meet the demands of large-scale video data processing.

First, the training of the SlowFast model begins with data

preprocessing. Video data usually have high frame rates and rich spatial-temporal information, so efficiently extracting this information is crucial in training. Input videos are typically split into several frames and fed into the slow and fast networks at specific time resolutions. The slow pathway network receives low frame rate video inputs, usually sampling only a few frames per second, which effectively capture long-term temporal dependencies and global features in the video, such as motion trends and background changes. On the other hand, the fast pathway network receives high frame rate video inputs, sampling more frames per second, enabling it to capture detailed changes and fast motions, which is crucial for detecting fine movements or high-frequency variations in the video. To ensure that the model can process video data at different frame rates, spatial-temporal downsampling and time alignment are performed during training to ensure that both slow and fast networks can process the same time period at different resolutions. This process prevents time misalignment between video frames, ensuring the model can extract valid features at different resolutions.

In terms of network architecture design, the SlowFast model relies on two parallel flow paths—the slow and fast paths—interactively trained. Each path contains independent convolutional layers, pooling layers, and other processing units, but some lower-layer network structures are shared. The slow pathway focuses on extracting low-time resolution and global features, effectively capturing long-term information in the video. In contrast, the fast pathway operates at high time resolution, focusing on capturing fast motions and detailed changes, which is particularly important for high-frequency movements. Through this parallel structure design, the network can extract rich feature information at different time scales, improving the model's performance. During training, by sharing some lower-layer network layers, the model can reduce the number of parameters and lighten the computational burden, while still effectively extracting features at different time scales. To allow each flow to independently and effectively handle its specific task, the SlowFast network design considers the independent training process of both the slow and fast pathways, optimizing the parameters of each flow to adapt to their specific spatial-temporal information capture requirements. This design allows the network to optimize the parameters of both flows separately during training, enabling it to capture key information at different time scales in the video and significantly improving overall performance.

In terms of loss function design, the SlowFast model usually adopts a multi-task learning framework, simultaneously optimizing the performance of both flows with a weighted loss function. In behavior recognition tasks, the loss function typically consists of classification loss and temporal loss. The classification loss ensures that the model can accurately predict the current action category at each moment in the video, while the temporal loss improves the model's accuracy in capturing temporal information, ensuring the model can correctly understand the evolution of actions in the video. This multi-task learning framework allows SlowFast to handle not only static information in videos (such as image content) but also the dynamic changes of actions in the video. To further enhance the model's generalization ability and stability, data augmentation techniques such as temporal cropping, spatial cropping, and image flipping are also employed during training. These augmentation methods improve the model's adaptability in various environments, especially when the training data is small or the environmental variations are large. Data augmentation helps to enhance the robustness of the model.

In terms of optimization strategies, the SlowFast model typically uses gradient descent-based optimization algorithms, such as Adam or SGD, combined with learning rate decay strategies to gradually adjust the learning rate, ensuring the stability and convergence of the training process. Due to the complexity of the SlowFast model structure, gradient clipping, batch normalization, and other techniques may also be used during training. These techniques help prevent gradient explosion and accelerate convergence. Specifically, gradient clipping limits the size of gradients to prevent gradient explosion during training, stabilizing the process, while batch normalization alleviates the internal covariate shift problem, accelerating convergence, and improving training stability.

Overall, the training design of the SlowFast model fully considers the spatial-temporal characteristics of video data. By processing long-term and short-term information in parallel through the slow and fast pathways, the model improves the accuracy of video action recognition tasks. Its optimization strategies combine multi-task learning, data augmentation, and modern optimization algorithms, allowing the model to perform strongly in various scenarios. Through these carefully designed technical details, the SlowFast model not only efficiently captures multi-scale information in videos but also ensures the efficiency and stability of the training process, providing robust support for complex video understanding tasks.[1]

2.2. Cesium Platform

2.2.1. Role of Cesium in Video Data Visualization and Geospatial Analysis

Cesium is an open-source 3D geospatial visualization platform that plays a critical role in video data visualization and geospatial analysis. With the rapid growth of big data and real-time data, especially in smart cities and public safety, integrating massive geospatial data with video information has become a hot research topic. Cesium, with its powerful 3D viewing and spatial analysis capabilities, can accurately combine video data with geographic coordinates, presenting rich spatiotemporal information. This integration provides users with real-time geographic updates and enables deep dynamic monitoring and analysis of spatial data.

In video data visualization, Cesium supports a variety of video formats and can combine video with 3D maps through geospatial annotations and time sequences, providing a more intuitive display. This feature is especially useful for applications where analyzing and tracing specific behaviors or events in a particular region is essential. For example, in smart city development, combining surveillance video data with 3D models and map data of the city can enable real-time monitoring of streets, roads, or public areas, offering technical support for public safety management.[2]

In geospatial analysis, Cesium is more than just a data display tool; its integration with other spatial data processing tools and analysis models allows for precise measurement, simulation, and analysis of geographic data. Combined with video data, users can perform comprehensive spatiotemporal analysis in a 3D space, identifying regional behavior patterns. Especially in the combination of video GIS and behavior recognition, Cesium, integrated with deep learning and computer vision technologies, can improve the accuracy and

efficiency of behavior recognition, providing more precise and real-time decision support for urban management and security monitoring.[4]

2.2.2. Methods for Integrating Video Data into the Cesium Platform

Integrating video data into the Cesium platform for spatiotemporal analysis is an essential technique for implementing video GIS (Geographic Information System). With the widespread use of video surveillance equipment and the improvement of data collection capabilities, effectively integrating video data with geospatial information has become a key focus in smart cities and public safety research. Cesium, being a powerful 3D geospatial visualization platform, provides efficient spatial data visualization and analysis functions, making it well-suited for spatiotemporal video data analysis.

To integrate video data into Cesium, the first step is to georeference the video data. This involves associating video frames with locations in a geographic coordinate system. By incorporating camera location, orientation, focal length, and other parameters, the geospatial position for each video frame can be determined. Video data is typically integrated using formats like GeoJSON, KML, or other geospatial data formats in Cesium, with timestamps linking the video data to geographic locations on the 3D map, enabling spatial display of the video data.

Once the video data is successfully integrated with geospatial information, it can be visualized using Cesium's 3D features. The platform allows videos to be played in a 3D Earth view and overlayed on the map, with users able to view video content from different locations and times. At this point, the video is no longer static but can be analyzed alongside other dynamic geographic data. This dynamic spatiotemporal analysis provides strong data support for public safety, urban management, and emergency response.[5]

In spatiotemporal analysis, Cesium's timeline feature allows users to view video data and its corresponding geospatial information in chronological order. Users can set specific analysis regions and perform behavior recognition, anomaly detection, and other analyses based on video data. By combining deep learning or computer vision algorithms, the platform can identify specific behavioral patterns or abnormal events within the video and link them with geographic data for real-time alerts and event analysis. These spatiotemporal analysis functions provide users with real-time, accurate data on activities, traffic conditions, and more within specific geographic areas.

3. Research Methods

3.1. SlowFast Dataset Selection

The Kinetics-400 dataset was chosen for training the SlowFast model in this study. This dataset contains over 40,000 video samples across 400 action categories and is widely used in action recognition and video classification tasks. The major feature of the Kinetics-400 dataset is its large scale and diversity, with each category containing hundreds of videos ranging from everyday actions to complex movements like professional sports and dance. Compared to UCF101, Kinetics-400 offers more categories and samples, suitable for large-scale training and higher precision model optimization. However, Kinetics-400 presents challenges as some action categories are more complex or have longer durations, potentially lacking the distinct action differentiation found in UCF101.

Similar to Kinetics-400, the UCF101 dataset includes 101 action categories, covering a wide range of activities from daily life, with hundreds of videos per category, making it ideal for more streamlined model training. Although UCF101 contains fewer categories and samples than Kinetics-400, its action categories are more defined, and the video length is shorter, making it suitable for tasks involving short-duration action recognition.

Another common dataset is HMDB51, which contains 51 action categories, but with fewer samples per category. Compared to UCF101 and Kinetics-400, HMDB51 contains more fine-grained actions, such as "wave hand" versus "raise hand," providing higher discriminative power in certain scenarios, but with weaker generalization ability due to the smaller dataset size.

The Something-Something V2 dataset is representative for fine-grained video recognition tasks, suitable for recognizing subtle daily actions and brief interactive behaviors. The challenge with this dataset lies in its abstract actions and short video durations, which demand high model precision.

Overall, the Kinetics-400 dataset offers clear advantages for large-scale training with its wide variety of actions and samples, making it ideal for complex behavior recognition tasks. In contrast, UCF101 is better suited for clear action categories, while HMDB51 and Something-Something V2 focus on fine-grained action recognition. The dataset selection should be based on task requirements, dataset scale, and action complexity.[3]

3.2. SlowFast Model Training

The computational platform used in this research is configured as follows: the GPU is the NVIDIA GeForce RTX 4060, which boasts powerful graphics processing capabilities, suitable for large-scale deep learning tasks and image processing applications. The processor is the Intel Core i9-13900HX, offering high frequencies and a multi-core architecture that delivers exceptional computational performance, especially for high-concurrency computing and complex algorithmic operations. The deep learning framework used is PyTorch 1.10.1, supporting various advanced deep learning models and efficient GPU acceleration, ideal for large-scale data training and model optimization.

During SlowFast model training, hyperparameter tuning and training techniques are critical for enhancing model performance. Proper hyperparameter settings improve the model's accuracy, robustness, and generalization capability. This experiment employed cosine annealing and warm-up strategies to adjust the learning rate, which gradually decreases over training. This learning rate scheduling method helps avoid oscillations from an overly high learning rate in later stages, enhancing convergence speed and stability. Additionally, the batch size was set to 16, balancing training efficiency and model generalization. Smaller batch sizes help prevent overfitting and improve the model’s adaptability to noise in the data. The Adam optimizer was used, which adaptively adjusts the learning rate for each parameter through estimates of the first and second moments, accelerating training and effectively preventing gradient vanishing or explosion issues. The Adam optimizer demonstrated excellent stability and robustness in this training.

To avoid overfitting, L2 regularization (Weight Decay) was

also employed. L2 regularization penalizes overly large weights, encouraging the model to learn simpler and more generalizable feature representations. The regularization coefficient was set to 1e-5, which was verified experimentally to effectively control overfitting risk while maintaining the model's learning ability. Additionally, to address class imbalance in the dataset, Focal Loss was used as the loss function. Focal Loss increases the weight of hard-to-classify samples and decreases the weight of easy-to-classify samples, making the model focus more on difficult samples and improving classification performance.

Lastly, the model was initialized using the Kinetics-400 pre-trained model. Kinetics-400 is a large-scale video dataset, and the pre-trained weights provided strong support for the model, helping accelerate convergence and improve performance. In summary, this experiment applied appropriate hyperparameter tuning and training techniques to enhance the performance of the SlowFast model in video behavior recognition tasks, ensuring the model's stability and generalization capability.

4. Results and Discussion

4.1. Performance of the SlowFast Model

In the experiment, the SlowFast model was trained and tested on various video datasets, demonstrating high accuracy and robustness. Especially in complex and dynamically changing video scenes, SlowFast effectively distinguishes between different types of actions and maintains high recognition accuracy when processing videos with long time spans. By optimizing the model's hyperparameters, incorporating data augmentation, and using pre-trained models, the performance of action recognition was further enhanced.

Compared to traditional single-stream networks, the SlowFast model exhibits superior performance when handling videos with complex actions and multi-level dynamics. This is particularly evident in its ability to balance the capture of high-frequency details and low-frequency background information. As a result, the SlowFast model has proven to be an efficient and powerful solution for various action recognition tasks, particularly in applications requiring long-term dependencies and fine-grained action recognition, such as sports video analysis, security monitoring, and human-computer interaction.

The following figure illustrates the recognition performance of the SlowFast model.

Figure 4-1. Recognition Performance of the Improved SlowFast Model in Scenarios

4.2. Deep Learning-Based Cesium Video GIS Behavior Recognition

This paper designs and develops a behavior recognition system based on SlowFast for Cesium video GIS, as shown in Figure 4-2.

Figure 4-2. Interface of the SlowFast-based Cesium Video GIS Behavior Recognition System

(1) Model Loading Function

The SlowFast-based Cesium Video GIS Behavior Recognition System includes a model loading module. By clicking the "Load Model" button, users can add 3D models to the scene. Figure 4-3 shows the model loading effect.

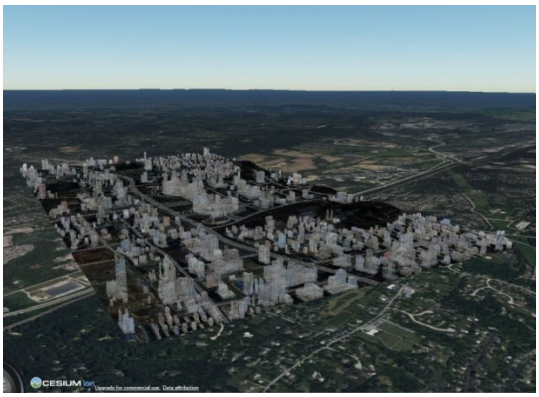

Figure 4-3. Model Loading Execution Effect

(2) Video Loading Function

The Cesium Video GIS Behavior Recognition System based on the SlowFast model includes a built-in video loading module. Users can simply click the "Load Video" button to directly load the video into the scene, displaying real-time behavior recognition results and providing an intuitive visualization experience.

Figure 4-4. Video Loading Runtime Effect

(3) Behavior Analysis

The behavior analysis module of the Cesium Video GIS behavior recognition system based on the SlowFast model simultaneously calls the model loading and video loading functions. It seamlessly integrates the behavior recognition results and visually displays them within the Cesium scene, offering a more immersive visual experience.

Figure 4-5. Operation Effect of Behavior Analysis

5. Conclusion

This paper presents an improved SlowFast model and successfully integrates it into the Cesium platform to achieve efficient spatiotemporal behavior recognition and visualization. First, addressing the limitations of traditional behavior recognition models in handling spatiotemporal behavior recognition tasks, the study utilizes the SlowFast model to improve behavior recognition accuracy.

Another key contribution of this research is the integration of the improved SlowFast model with the Cesium platform, forming a spatiotemporal behavior recognition method. In the Cesium platform, the model can combine the recognized behaviors extracted from video data with geospatial information in real-time, achieving accurate location-based spatiotemporal visualization of the behavior recognition results. This integration not only enhances the real-time performance and accuracy of behavior recognition but also provides more contextual information for decision-making within spatial data, allowing behavior recognition results to be more intuitively and effectively presented in a geographic information system (GIS).

Through this approach, the study demonstrates the superior performance of the SlowFast model in video behavior recognition and highlights the broad application potential of spatiotemporal data fusion in fields such as intelligent surveillance and urban security. The successful implementation of this method offers new insights for the development of multimodal data fusion and spatiotemporal behavior recognition technologies, with significant practical implications, especially in research directions integrating advanced deep learning models with GIS platforms.

References

[1] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[2] Zhang X, Li Q, Li X, et al. A new method for orthographic video map construction based on video and GIS[J]. Geocarto International, 2023, 38(1): 2289450.

[3] Kang X, Li J, Fan X. Spatial-temporal visualization and analysis of earth data under Cesium Digital Earth Engine[C]//Proceedings of the 2018 2nd International Conference on Big Data and Internet of Things. 2018: 29-32.

[4] Jiao Q, Sun G, Chen Z, et al. A 3D WebGIS-enhanced representation method fusing surveillance video information[J]. IEEE Access, 2023.

[5] Liu F, Han Z, Song H, et al. Crowd sensing and spatiotemporal analysis in urban open space using multi-viewpoint geotagged videos[J]. Transactions in GIS, 2023, 27(2): 494-515.