Artificial Intelligence Assists Drones in Motion Capture

Yu Cheng*

Xiamen No.6 High School, Xiamen, China

*Corresponding author: yuc40334@gmail.com

Abstract. With the development and construction of urbanization in the country, the scale of the city has expanded; meanwhile, the problem of urban traffic congestion is becoming increasingly severe. To ensure the efficient operation and stable development of society, it's necessary to use AI-assisted drones in motion capture of violations by motor vehicles and voice display. AI assists drones in motion capture can reduce traffic congestion to a certain extent, thereby minimizing the economic losses caused by traffic problems. The article conducts research and analysis the principles and function of the gesture recognition algorithm of neural networks and spatiotemporal convolution network algorithm At the same time, in combination with the small body and mass characteristics of drones, this paper analyzes the application feasibility of artificial intelligence technology in assisting drones with motion capture, and also analyzes the development trend of artificial intelligence technology in assisting drones with motion capture and application.

Keywords: Human Motion Capture; Convolutional Neural Network; Graph Convolutional Network; Markerless Motion Capture; Multi-Modal.

1. Introduction

With the development and construction of urbanization in the country, the scale of the city has expanded; meanwhile, the problem of urban traffic congestion is becoming increasingly severe. To ensure the efficient operation and stable development of society, it's necessary to use AI-assisted drones in motion capture of violations by motor vehicles and voice display. AI assists drones in motion capture can reduce traffic congestion to a certain extent, thereby minimizing the economic losses caused by traffic problems.

The application field of drones' motion capture systems is extensive, such as aerial photography, video shooting, mining and cargo transportation, security, and industrial inspection. It enables drones to operate autonomously by leveraging computer vision and AI, for instance, when drones deliver goods automatically, manual operation of drones for delivery requires a considerable amount of time and effort. Therefore, the development of AI, especially the spatiotemporal graph convolutional network algorithm, offers a new solution for this object. The recognition of human actions is carried out by constructing a model to calculate the convolutional mesh training [1-4].

Gesture recognition using convolutional neural networks is an effective method for object recognition, widely used in many fields. The process involves steps like data collection, model design, training, and optimization. The system is made up of four main parts: convolutional layers, which extract key features from images; pooling layers, which retain important features by reducing the image's size; fully connected layers, which transform the pooled features into vectors for further analysis; and activation functions, which apply nonlinear transformations to improve the model's ability to recognize complex patterns.

This essay, through the study of the core algorithms of motion capture nowadays, analyzes their functions advantages, at the same times, combined with the conditions of miniaturization and light weight of drones, and analyzes the feasibility of applying AI technology in motion captures by drones, deeply disserts the core algorithms of AI assisting drones in motion capture, and evaluates and compares gesture recognition algorithms based on convolutional neural networks. The advantages and effectiveness of network algorithms in assisting drones in their actions and capture to help relevant industry personnel choose the most suitable method for them [5][6].

2. Core Algorithms for Motion Capture Using Drones and AI

2.1. Gesture Recognition Algorithm Based on Convolutional Neural Networks (CNN)

Drones equipped with cameras capture motion video, from which keyframes are extracted. These keyframes undergo pre-processing, which includes cropping the palm region, normalizing the size, and applying techniques like brightness adjustment and slight rotations for data augmentation. Lightweight convolutional neural networks (CNN) are used to process the images. Shallow convolution layers with 3x3 kernels extract local detail features. Deeper layers alternate between convolution and pooling operations to integrate higher-level semantic features. At the output, a fully connected layer and softmax classifier map the extracted features to specific gesture categories. The recognition results are then translated into control commands and sent to the drone for real-time action. In dynamic gesture recognition, temporal dependencies are often modeled to improve the stability of continuous gesture recognition [7].

2.1.1 Faster R-CNN

Researchers have developed a multi-scale small object detection method to address the challenge of detecting small targets in drone-captured aerial images, especially in complex backgrounds. For instance, bird nests on high-voltage towers are used as the target objects. The method improves a convolutional network (such as ResNet101) for feature extraction, employs multi-scale sliding windows to generate initial candidate regions, and uses deconvolution operations to enhance the resolution of feature maps. The model consists of four main components: feature extraction, candidate region generation, region of interest (RoI) pooling, and classification/regression estimation. The process begins by extracting features from input images using convolution and pooling layers. These feature maps are passed to the Region Proposal Network (RPN) and further processed. The RPN generates candidate regions based on anchor points, which are evaluated through a softmax function. RoI pooling is applied to adjust the candidate regions to a fixed size, and the data is then passed to a final classification network for action determination [8].

2.1.2 Mask R-CNN

Using cameras and deep convolutional networks, a method called Mask R-CNN, along with algorithms for tracking and classification, is applied to detect the posture of cows in video footage. The system tracks key points on the cow's back and head to determine posture and detect lameness. The system uses a dataset containing images of cows for training and can accurately identify the coordinates of key points for posture analysis. Initially, infrared images are processed through a backbone convolutional network to extract features, which are then passed through a Region Proposal Network (RPN) to generate candidate boxes. These boxes are adjusted in size using the ROI Align layer, and finally, the processed features are classified and segmented to identify the object and its attributes [9].

2.2. Spatio-Temporal Graph Convolutional Network Algorithm (ST-GCN)

The system extracts key human body information using pose estimation tools, converting the video input into frames. Each frame is analyzed to detect specific points of interest, which are then organized into a spatio-temporal graph. In this graph, nodes represent human body keypoints, with features such as coordinates, while edges represent the spatial connections between keypoints in the same frame and the temporal connections between the same keypoints across frames. The spatial features are extracted using convolution operations, and temporal features are captured through a temporal mapping process. The model groups nodes based on their relationship (centripetal and centrifugal) to better reflect human movement. After multiple layers of spatio-temporal convolution, the extracted features are input into a classification layer to recognize the actions [10][11].

2.2.1 HSTGCN

A parking space prediction scheme based on HST-GCNs utilizes graph convolutional networks (GCN) and gated linear units (GLUs), alongside one-dimensional convolutional neural networks (CNNs), to extract spatial and temporal features. This method builds spatio-temporal convolution blocks and introduces an attention mechanism to capture complex mixed spatio-temporal dependencies, achieving higher accuracy in predicting parking space availability. The model works by fusing traffic flow and travel time data, applying domain conversion to align data types, using gated convolutions for temporal patterns, and leveraging graph convolutions for spatial dependencies. The final model predicts future traffic conditions based on these integrated spatio-temporal features HSTGCN integrates three core functions: dynamic heterogeneous graph construction f_H(•), spatiotemporal modeling T(•)/f_S(•), and ultra-short-term prediction f_F(•). f_H(•), consolidates multi-models into a unified heterogeneous graph. f_H(•),reshapes 1D sequences into 2D form to capture temporal dependencies ,while f_T(•) conducts hierarchical spatial link modeling using GCAM . Ultimately, f_F(•), projects fused features onto multi-station, multi-step prediction outputs [12].

2.2.2 Graph WaveNet

A model called Graph WaveNet is applied to traffic prediction using a dataset that records traffic speed from sensors on highways. This model learns spatial dependencies through graph convolution layers and captures temporal dependencies using dilated convolutions. It has demonstrated superior performance compared to traditional models such as ARIMA and LSTM, as well as convolution-based models like STGCN. The model consists of spatio-temporal layers, each containing multiple layers to capture dependencies at different time scales. The input data is processed through a linear layer, then by a gated temporal convolution module, followed by graph convolution layers. Dilated causal convolutions are used for the temporal aspect of the model, with residual and skip connections to improve performance and prevent issues like vanishing gradients [13].

3. Practical Application Cases of AI in Drone-Based Motion Capture

3.1. Power Line Inspection Action Monitoring

A system developed by the team of Wu Weizhong at the Guangdong Dongguan Power Supply Bureau of the Southern Power Grid enables full-process intelligent inspection of power transmission line joints. Using AI-powered drones to capture the actions of inspection personnel, the system recognizes actions such as climbing and tool usage. It then evaluates the standard of operations in real-time and provides feedback. This automation has significantly improved the efficiency of power line inspections, allowing for quicker and more accurate assessments of work quality and safety.

3.2. Film Action Assistance in Filming

Lionsgate, a Hollywood film studio, uses multiple AI drones to simultaneously capture the body movements of actors during outdoor scene shooting. These drones work with AI algorithms to stitch together the captured data in real-time, generating dynamic motion tracks that are used as references for post-production visual effects. This approach significantly shortens the time required to film outdoor action shots, providing both efficiency and creative flexibility for the filmmakers.

3.3. Agricultural Labor Behavior Analysis

Crux Agribotics, a Dutch agricultural technology company, uses AI drones to capture the labor actions of farm workers. The drones are able to identify the intensity and frequency of actions like sowing and harvesting, and they combine this data with crop growth metrics to generate reports on labor efficiency. After implementing the system in 2023, partnering farms saw a 25% improvement in labor efficiency per worker, contributing to better resource allocation and higher productivity.

4. Conclusion

In conclusion, the integration of drones with motion capture systems has proven highly effective, overcoming the limitations of traditional methods. The spatio-temporal graph convolutional network algorithm, with its precise capture and processing of spatio-temporal features, has significantly improved the continuity and accuracy of motion analysis and prediction.

The gesture recognition algorithms based on neural networks have efficiently extracted and accurately classified gesture features, even in complex environments. These advancements provide essential support for the development of motion capture technology and open up new possibilities for its future growth.

In sports training, this technology can capture athletes' movements and create detailed motion records, offering precise analysis and technical guidance. In industrial manufacturing, it can monitor workers' actions and machinery conditions, helping to assess equipment performance, predict failures, and enable maintenance. In action game development, multi-angle motion capture using drones, combined with data analysis, allows for the creation of realistic motion skills, helping to craft lifelike character actions and expressions in games.

By combining drones with motion capture, this technology has great potential to enhance various fields, including sports, manufacturing, entertainment, and more.

References

[1] Zhang Z, Huang K. Drone-based vision systems: A survey of algorithms and applications. IEEE Access, 2019.

[2] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press, 2016.

[3] Mohammed F, Idries A, Mohamed N, Al-Jaroodi J, Jawhar I. Smart City Drones: Opportunities and Challenges. In: 2014 International Conference on Unmanned Aircraft Systems (ICUAS). IEEE, 2014: 267-273.

[4] Ayamga M, Akaba S, Nyaaba A A. A comprehensive review of drone applicability in various fields. Technological Forecasting and Social Change, 2021, 167: 120677.

[5] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.

[6] Howard A G, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[7] Molchanov P, Gupta S, Kim K, Kautz J. Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2015.

[8] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the IEEE Conference on Neural Information Processing Systems (NeurIPS), 2015: 91-99.

[9] He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2961-2969.

[10] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 744-753.

[11] Shi L, Zhang Y, Cheng J, Lu H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[12] Xiao X, Li X, Liu W. HST-GCN: Hybrid spatio-temporal graph convolutional networks for complex traffic prediction. Journal of Advanced Transportation, 2021: 1-10.

[13] Wu Z, Shi X, Zhang C, Chen L. Graph WaveNet for deep spatial-temporal graph modeling. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019: 1907-1913.