





Motion and Action Analysis of Drama Performance Based on Convolutional Neural Network

Li Han¹  and Hongtao Niu² 

¹School of Film and Television, Handan University, Handan, Hebei 056005, China,
hlmoment@163.com

²School Performing Arts, Hebei Institute of Communications, Shijiazhuang, Hebei 050071, China,
nztking520@126.com

Corresponding author: Hongtao Niu, nztking520@126.com

Abstract. This article endeavours to develop an advanced drama performance analysis system, leveraging computer-aided design (CAD) and big data technology. By integrating CAD's geometric modelling capabilities with big data's statistical strengths, we've crafted an intelligent framework tailored for analyzing drama performances. Our methodology initially involves utilizing CAD to model human movements precisely, followed by collecting and organizing extensive drama performance video data via big data technology. Subsequently, deep learning algorithms are employed to extract action features, enabling automatic recognition and analysis of performance actions. Results indicate that our proposed system maintains high processing speed and image quality, even when confronted with complex actions, varying lighting conditions, and background shifts. Compared to existing methods, our system enhances motion capture accuracy and efficiency, offering robust technical backing for creating, instruction, and preserving drama art. Looking ahead, we plan to deepen our research and foster the progression of action analysis technology within drama performances.

Keywords: CAD; Big Data; Dramatic Performance; Motion Analysis

DOI: <https://doi.org/10.14733/cadaps.2025.S9.147-160>

1 INTRODUCTION

In the art of drama performance, action is an important means for actors to shape their roles. CAD technology provides a powerful tool for the analysis of drama performance actions with its accurate modelling ability. To uncover profound insights like audience behaviour and market trends, big data technology harnesses the collection, integration, and analysis of extensive datasets [1][2]. CAD's robust graphics processing and data analysis prowess allow designers to capture and simulate intricate action details [3] precisely. In drama, action design necessitates considering posture, expression, and movement fluency; CAD technology digitally models these complexities [4]. During performances, vast data resources emerge from audience feedback, box

office figures, and social media commentary, offering rich information to inform drama analysis [5]. Big data analysis grants a deep understanding of audience needs and market trends, guiding performance optimizations. Given the increasing popularity and accessibility of 3D skeleton data, skeleton-based human motion recognition technology is gradually becoming a research focus in theatrical performance motion analysis systems based on CAD and big data. However, a core challenge faced by action recognition is that the motion representations captured from different angles exhibit significant differences, greatly increasing the difficulty of recognition [6]. The core of this scheme lies in the two view adaptive neural networks we designed: VA-RNN (View-Adaptive Recurrent Neural Network) and VA-CNN (View-Adaptive Convolutional Neural Network). Both networks are equipped with a novel view adaptation module that can learn and determine the most suitable observation viewpoint [7]. VA-RNN is based on a recursive neural network (RNN) with long short-term memory (LSTM), while VA-CNN is based on a convolutional neural network (CNN). In order to effectively reduce the impact of view changes on action recognition, some scholars have proposed an innovative view adaptation scheme. And convert the skeleton data to these viewpoints for end-to-end action recognition with the main classification network. This transformation greatly eliminates the impact of viewpoint changes on action recognition, allowing the network to focus on learning specific features of the action itself, thereby achieving excellent performance. Through ablation research, we found that the proposed view adaptive model can convert skeletal data from different views into more consistent virtual viewpoints. In theatrical performances, precise capture and recognition of movements are crucial for analyzing actors' performance skills, emotional communication, and stage scheduling [8]. This scheme abandons the traditional method of relying on fixed human-defined prior criteria to reposition the skeleton, and instead adopts a learning-based data-driven approach to automatically determine the virtual observation viewpoint during the action process. In addition, we have also designed a dual-stream scheme, namely VA fusion. This scheme combines the scores of VA-RNN and VA-CNN networks to provide the final prediction results. This fusion strategy further enhances the performance of the model and improves the accuracy and robustness of action recognition.

Big data has shown great potential not only in radar applications involving target classification and imaging but also in theatrical performance action analysis systems based on CAD (computer-aided design) and big data [9]. Placing big data in the context of data-driven theatre performance action analysis methods, we can observe that compared to traditional methods that rely on manual feature extraction and analysis, DL technology has a higher degree of automation and stronger processing capabilities. In the drama performance action analysis system, big data can accurately capture and classify actor actions by processing and analyzing CAD-designed stage layouts, actor action data, and a large number of live performance records. Outside the field of indoor monitoring, DL technology has opened up new paths for precise analysis of theatrical performance movements while classifying daily human activities, detecting falls, and monitoring gait abnormalities [10]. Through DL, the system can provide real-time accurate statistical information on actors' motion joints, thereby helping directors, actors, and dance choreographers better understand performance details, optimize motion design, and improve performance quality. Behind this interest, there is a close connection with emerging application demands related to smart and secure homes, assisted living, medical diagnosis, and artistic performance innovation. This ability not only relies on advanced neural network structures but also on efficient data input representation methods and appropriate training strategies. And how to ensure the real-time accuracy of the system to meet the real-time feedback needs in actual performance scenarios [11]. However, despite the significant achievements of DL in theatrical performance action analysis systems, there are still some important challenges to fully unleash its potential. For example, how to effectively integrate CAD design and big data resources to improve the quality and usability of data.

Only relying on CAD technology and big data technology is not enough to realize the comprehensive analysis of drama performance. In practical application, advanced image processing and computer vision technology are also needed to achieve high-precision capture and recognition of actions [12]. As a deep learning algorithm, Convolutional Neural Network (CNN) is

especially suitable for processing image and video data. In the action analysis of drama performance, CNN can be used to capture the details of actors' actions, such as gestures, expressions and body postures [13]. Virtual Try-On Network (VITON) network realizes high-precision capture and virtual display of human movements by combining image segmentation, attitude estimation and image generation technology. The objective of this research is to develop an action analysis system for drama performances, leveraging CAD and big data. By incorporating cutting-edge technologies like CNN and VITON networks, we can achieve high-accuracy capture and recognition of actors' movements. As science and technology continually evolve, the realm of drama performance art stands to gain numerous new opportunities. Through the in-depth exploration of this study, I hope to inject new vitality into the development of drama performance art.

The drama performance action analysis system constructed in this article has the following innovations:

(1) The system combines the accurate modelling of CAD technology with the data analysis ability of big data technology and expands the application scope of CAD and big data technology in the art field.

(2) CNN technology is introduced in this study, which can capture the details of drama performance more accurately and improve the accuracy of action analysis.

(3) The system provides an interactive action analysis and feedback mechanism, which enables designers and actors to check the action analysis results in real-time and make adjustments.

Firstly, this article expounds on the goal and significance of the research; Then it introduces the application of CAD and big data in the action analysis of drama performance; Then, the construction and algorithm realization of the drama performance action analysis system are carried out; Finally, it is verified on a specially constructed data set and the system performance is assessed.

2 OVERVIEW OF RELATED THEORIES

The core of CAD is to utilize the powerful computing and graphic processing capabilities of computers to assist designers in efficient and accurate design work. CAD technology can achieve visual display of theatrical performances through methods such as 3D modelling and animation simulation. Pareek and Thakkar [14] used CAD software to create 3D models of actors and simulated different movements by adjusting the model parameters. CAD technology can also be combined with motion capture devices. The actor's motion data is collected by motion capture devices and then visually displayed by CAD software. Qi et al. [15] intuitively observed the trajectory of actors, thereby optimizing action design more accurately. In the field of theatrical performance, big data technology is mainly used for audience behaviour analysis and market trend prediction. By collecting and analyzing audience viewing data, social media comments, box office information, and other data resources, big data technology can reveal audience needs and market trends. By analyzing the audience's viewing behaviour and feedback data, we can understand their preferences for certain behaviours; By analyzing box office data and social media reviews, we can understand the market acceptance of a certain TV drama. By combining action data captured by CAD technology with big data technology, Sreenu and Durai [16] gained a deeper understanding of the relationship between actions and audience reactions. In motion analysis of theatrical performances, CNN is mainly used for motion capture and recognition. CNN can automatically learn the features of image or video data through convolutional layers, pooling layers, and fully connected layers. In action analysis of theatrical performances, actors' action videos can be used as input data, and CNN can be used for feature extraction and classification recognition. By training a large amount of action video data, CNN learned feature representations of different actions, achieving accurate recognition of actor actions. Compared with traditional motion capture

methods, CNN can recognize the basic movements of actors, and capture subtle facial expressions and pose adjustments.

In the field of human-computer interaction, visual gesture recognition technology has always been a hot research topic. However, its recognition performance is often limited by the performance of recognition algorithms. In the theatrical performance action analysis system based on CAD and big data, gesture recognition also plays a crucial role. On this basis, the system achieved effective segmentation of gestures, further reducing the interference of environmental factors on recognition performance. This measure not only enhances the model's ability to recognize complex gestures but also improves its generalization performance in different scenarios. These algorithms not only reduce the impact of shooting angles and complex environments on gesture recognition performance but also significantly improve recognition accuracy in constantly changing environments. In order to further improve the accuracy of gesture recognition, the system also fully utilizes big data resources. Wang et al. [17] extensively trained the convolutional neural network model by introducing rich gesture data such as the ASK gesture database. In order to improve the accuracy and robustness of gesture recognition, researchers have introduced skeleton algorithms and convolutional neural networks (CNN) to optimize the recognition process. In order to overcome the problem of recognizing the same gesture caused by the shooting angle, researchers optimized the skeleton algorithm using the concept of layer-by-layer peeling. This optimization strategy can accurately extract key node information from the hand skeleton graph, and then determine the direction of the gesture based on the spatial coordinate axis of the hand. The experimental results show that in the drama performance action analysis system based on CAD and big data, Zhang et al. [18] used optimized skeleton algorithms and convolutional neural network models for gesture recognition, with a recognition rate of 96.01%. Compared with SVM methods, dictionary learning+sparse representation, and traditional CNN methods, this achievement demonstrates significant advantages. This not only validates the effectiveness of optimization algorithms and big data resources in improving gesture recognition accuracy but also provides strong support for the further development and improvement of drama performance action analysis systems.

Video-based human motion recognition is currently one of the most active research directions in the field of computer vision, and technology based on Kinect cameras has injected new vitality into this field. To fill this gap, Zhang et al. [19] conducted an in-depth analysis and comparison of the 10 latest Kinect-based cross-topic and cross-view action recognition algorithms using six benchmark datasets. Since the emergence of Kinect cameras, many human motion recognition technologies based on Kinect have sprung up like mushrooms after rain. In addition, we have implemented and improved some of these technologies and included their variants in the comparison. The experimental results show that in cross-topic action recognition, most methods exhibit better performance than cross-view action recognition. This may be because cross-topic recognition relies more on the intrinsic characteristics of behaviour, which have stronger universality across different topics. This measure not only helps us to have a more comprehensive understanding of the advantages and disadvantages of these technologies but also provides a valuable reference for the development of drama performance action analysis systems based on CAD and big data. Meanwhile, Zhang et al. [20] found that skeleton-based features exhibit more robust performance in cross-view recognition than depth-based features. This may be because skeletal features can more accurately capture the movement trajectory and posture changes of the human body, thereby maintaining good consistency from different angles. In drama performance action analysis systems, especially in the context of CAD and big data, the performance of action recognition largely depends on the types of extracted features and the representation of actions. However, despite these technologies having their own advantages in terms of functionality, we have not yet seen a comprehensive comparison in the literature, particularly between handcrafted features and deep learning features, as well as between depth-based features and skeleton-based features.

3 ALGORITHM IMPLEMENTATION OF DRAMA PERFORMANCE ACTION ANALYSIS SYSTEM

3.1 Overall System Architecture

The drama performance action analysis system is a highly integrated technical framework. The core process starts from the data acquisition layer. It is responsible for capturing every subtle action of the actors in the drama performance with high precision. This includes, but is not limited to, the actor's posture change, rich expression display, and complex action trajectory.

After data collection, these data resources will be immediately transmitted to the data processing layer for further refined processing. In the data processing layer, the system will use a series of professional technical means to eliminate impurities and interference in the original data.

Then, the action recognition layer began to play its role. This layer mainly relies on advanced deep learning algorithms such as CNN to identify the preprocessed action data. Through long-term model training and optimization, the system can automatically extract the core features of the action.

On the basis of motion recognition, the system further goes to the level of data analysis. This layer uses big data technology to deeply mine action data. The system provides users with more diverse perspectives by revealing the internal relationship between actions and plot development, role shaping, audience feedback and other dimensions.

Finally, all the analysis results will be presented to users through the presentation layer. The presentation layer will help users to quickly understand the results of action analysis through charts, reports and animations.

3.2 Algorithm Realization

In drama performance action analysis, CNN serves to extract features and categorize action data. Initially, motion data is transformed into an image or video format compatible with CNN. Subsequently, extensive action data is utilized to train the CNN model, enabling it to learn the traits of action data and enhance recognition accuracy over time. Upon completion of training, the CNN model can identify fresh action data by inputting it and outputting the respective action category alongside its probability distribution.

As a deep learning model, CNN has shown excellent performance in the fields of image recognition and video processing. As shown in Figure 1, the core components of CNN architecture mainly include convolution operation, pooling process and full connection layer. These components work together, which enables CNN to process and analyze complex visual information efficiently.

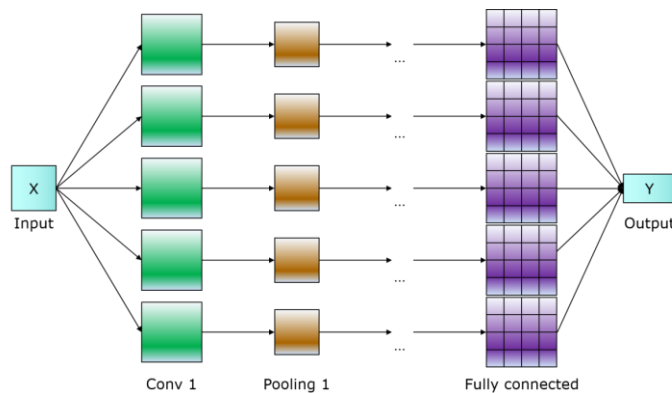


Figure 1: CNN.

In the discrete domain, the convolution operation is realized as follows:

$$f \cdot g [n] = \sum_{m=-\infty}^{\infty} f[m] g[n-m] \quad (1)$$

The convolution kernel is usually represented as an $n \times n$ matrix, which is the key element in convolution operation. As shown in Figure 2, a 3×3 convolution kernel example is adopted. Through the sliding operation of the convolution kernel on the image, the system can flexibly enhance or suppress specific feature patterns.

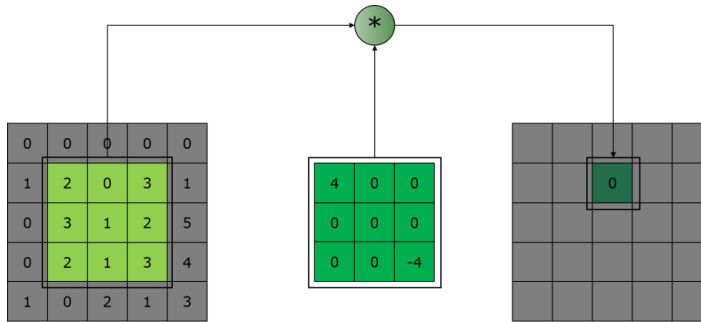


Figure 2: Convolution operation.

Actions in drama performances often contain rich visual information, such as gestures, expressions and body postures of actors. Through convolution operation, these action features can be accurately captured. The flexibility and trainability of the convolution kernel enable CNN to adapt to the demand of action feature extraction in different drama performance scenes.

Presume the feature output x^l of the fully connected layer l adheres to the formula:

$$x^l = f(w^l x^{l-1} + b^l) \quad (2)$$

Here w^l represents the weight parameter, b^l the offset term and the Softmax regression classifier requires iterative updates and learning focused on the subsequent functions:

$$h_w \left(\begin{matrix} \rightarrow \\ x \end{matrix} \right) = \frac{1}{\sum_{i=1}^k e^{\vec{w}_i \cdot \vec{x} + b_i}} \begin{matrix} \rightarrow \\ \vec{w}_1 \cdot \vec{x} + b_1 \\ \rightarrow \\ \vec{w}_2 \cdot \vec{x} + b_2 \\ \dots \\ \rightarrow \\ \vec{w}_k \cdot \vec{x} + b_k \end{matrix} \quad (3)$$

Here k indicates the number of classification categories, with b_i and \vec{w}_i representing the offset and weight vectors, respectively, for the category i . The sample \vec{x} corresponds to the probability of a class j , expressed by the formula:

$$P \left(y = j \mid \vec{x} \right) = \frac{e^{\vec{w}_j \cdot \vec{x} + b_j}}{\sum_{i=1}^k e^{\vec{w}_i \cdot \vec{x} + b_i}} \quad \sum_{j=1}^k P \left(y = j \mid \vec{x} \right) = 1 \quad (4)$$

After training and learning are finished, \vec{w}_i and b_i are ascertained, enabling the formulation of the target loss function is as follows:

$$J_{w,b} = -\frac{1}{m} \sum_{j=1}^m \sum_{l=1}^k 1_{\{y^j = l\}} \log \frac{e^{\vec{w}_l \cdot \vec{x} + b_l}}{\sum_{i=1}^k e^{\vec{w}_i \cdot \vec{x} + b_i}} \quad (5)$$

Here m signifies the sample count in the drama performance training set, k indicates the number of drama performance classification categories, and $1_{\{ \cdot \}}$ serves as the indicative function.

The mechanism of the pooling layer is to select representative regions from the feature map extracted by convolution to replace the whole region. As shown in Figure 3, the pooling operation realizes the down-sampling of image data by performing operations such as maximum pooling or mean pooling.

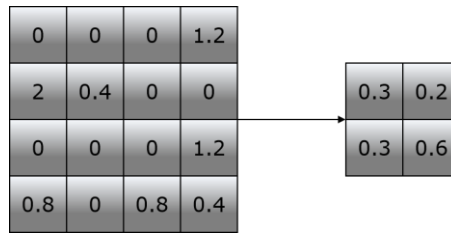


Figure 3: Mean pool operation.

The application of the pooling process can further refine the action characteristics of drama performance. Through pooling operation, the most representative features can be extracted from a large number of action data. The pooling process can also enhance the robustness of CNN to action changes so that the system can identify actions stably under different lighting, angles and backgrounds.

The fully connected layer, serving as the final element in a CNN, integrates and classifies features derived from prior convolution and pooling. Here, every neuron links to all neurons in the preceding layer, facilitating further abstraction and classification of features through weight and bias adjustments.

CNN can effectively process and analyze the dramatic action data generated by CAD technology and the visual information in the large-scale performance database. The application of CNN provides technical support for an in-depth understanding of the details of drama performance.

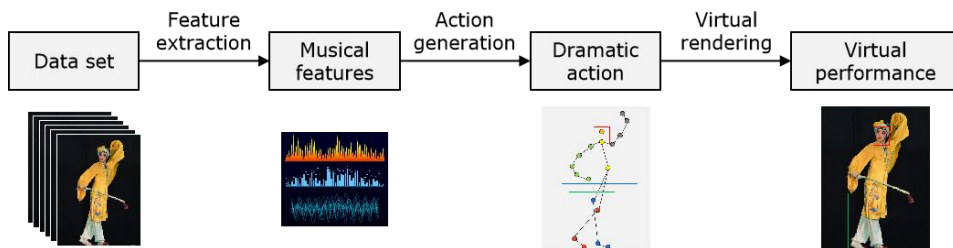


Figure 4: Basic framework of drama performance action generation.

Figure 4 shows the basic framework of drama performance action generation based on CNN. As the basis of CNN training, the data set contains a large number of carefully selected drama performance action samples. These samples cover movements with different styles and difficulties. By learning these samples, CNN can extract the key features of the action. In drama performance, music is often closely linked with actions, which together create a specific emotional atmosphere.

Therefore, in the process of action generation, musical features are introduced as important input information. In the motion capture stage, the system uses advanced motion capture equipment to obtain the skeleton information of actors. This information includes joint position, bone length and so on. In the action generation stage, the system uses the skeleton drama generated by CNN to perform actions, combined with the virtual character model, to generate realistic virtual character performances.

Figure 5 shows the way of dividing human body regions according to joints. By dividing the human body into different regions, such as the head, trunk, arms and legs, we can capture and analyze the action characteristics of each region more accurately. This division is helpful to deeply understand the structure and law of action in the study.

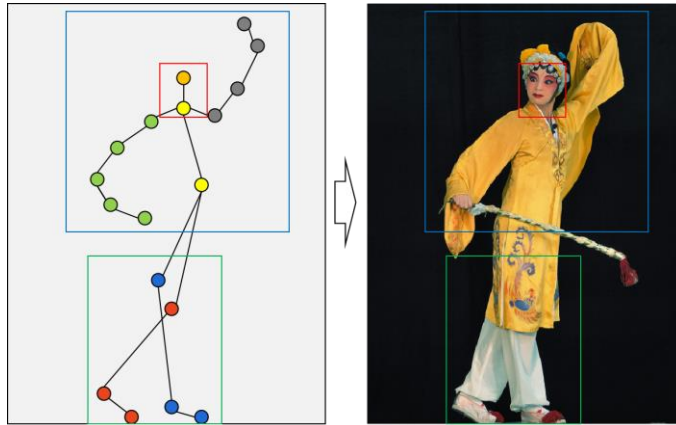


Figure 5: Dividing human body regions according to joints.

To ensure accurate and standardized comparison results when constructing the human skeleton template, given variations in height and body shape, the following standardized approach is employed:

$$P_c \ x,y = \left(\frac{sl_x + sr_x}{2}, \frac{sl_y + sr_y}{2} \right) \quad (6)$$

$P_c \ x,y$ denotes the standard central point of the acquired bone points. The x-coordinates for the left and right shoulder bone points are represented by sl_x and sr_x , respectively while sl_y, sr_y representing their y-coordinates. Subsequently, the difference between each skeleton point and the centre is computed to derive the center-standardized human skeleton coordinate points. Given the variation in shoulder width among individuals, the distance between the two shoulders is employed to normalize the scale of the human skeleton coordinates.

$$D_s = \sqrt{sl_x - sr_x^2 + sl_y - sr_y^2} \quad (7)$$

Here D_s signifies the Euclidean distance between the two shoulders. By dividing x,y the bone points, post-center standardization, D_s , we obtain skeleton coordinates standardized to a uniform scale for matching purposes.

In practical application, combined with CAD technology, the human body region is further refined and modelled. By constructing an accurate human model, we can simulate muscle movement and joint stress under different movements.

In the encoder-decoder stage, we begin with the given figure gesture representation p and target clothing image c . By learning the transition from c to p , a preliminary image I is synthesized. VITON utilizes a multi-task encoder-decoder framework to generate both a fitted image of a person and a corresponding clothing mask, which guides the network's attention to the clothing area and refines the ultimate output.

Formally, let G_c represent the objective function of the generator in the encoder-decoder structure, which takes the spliced c and p as inputs to generate 4-channel output:

$$I', M = G_c(c, p) \quad (8)$$

The first three channels represent the composite image I' , and the last channel M represents the segmentation mask of the clothing area. The goal of the network is to learn a generator to make I' it is close to the target image I, M and close to the real mask M_0 . The loss function of the encoder-decoder is expressed as the sum of perceptual loss and L_1 loss:

$$L_{G_c} = \sum_{i=0}^5 \lambda \|\varphi_i(I' - \varphi_i I)\|_1 + \|M - M_0\|_1 \quad (9)$$

Perceptual loss makes the generated image match the RGB value of the real image, thus making the image synthesis network learn the real pattern.

G_R transforms the clothing image through thin-plate spline transformation, thus generating the deformed clothing image c' . The goal G_R is to make the deformed clothing image c' seamlessly combine with the clothing area on the synthesized rough image I , and properly handle the occlusion when the arm or hair appears in front of the body. The method is to connect c' and rough output I' as the input of optimizing the network G_R . Then, the optimized network generates a 1-channel synthesis mask α , which indicates how much information the final result uses from two sources. VITON's final virtual fitting output is the combination of c' and I' :

$$\hat{I} = \alpha \odot c' + 1 - \alpha \odot I' \quad (10)$$

In the action analysis of drama performance, this article tries to combine the VITON network with motion capture technology to realize the virtual display of actors' actions. The skeleton information of actors is obtained by using motion capture equipment, and then this information is combined with the VITON network. By inputting skeleton information and target clothing information, the VITON network can generate the virtual image of actors wearing target clothing and realize the synchronous display of actions. During training, the image mode data has a high dimension, with the model emphasizing action learning but struggling with low-dimensional music features. For 2D skeleton sequence prediction, limited data causes the network to prioritize audio feature relationships.

4 EXPERIMENT AND RESULT ANALYSIS

In order to comprehensively assess the effectiveness of the model in video action analysis of drama performance, this article carries out experiments on a specially constructed video action data set of drama performance. This data set covers a wide range of action categories with uniform video parameters.

4.1 Data Set Overview

High-quality data sets are very important to the deep learning model. However, in the task of audio-driven drama performance action generation, there are few large-scale drama performance data sets with music, and these data sets have limitations in the diversity of available drama performance actions and music styles. Different from music, there is no universal definition of the

rhythm of human movements. According to the particularity of drama performance, we take the position where the drama performance suddenly slows down or the dancer's direction suddenly changes as the division position of drama performance units. Convolutional neural networks can't learn time sequence information for single frame input, but if multiple time sequence information are combined, it can learn the relationship between actions with fewer parameters. Different from some work that directly converts audio information into drama performance, some work will continuously input the action of the previous frame in different stages of the model, prompting the subsequent learning output to achieve the effect of correction. To assess the benefit of multimodal information on network learning and convergence, we varied the input: single skeleton mode, single image mode, and combined skeleton-image mode. The output remained as predicted skeleton information, with all other network parameters held constant across layers.

The parameters of the data set used in the experiment are shown in Table 1. This data set contains 105 different types of drama performances, ranging from subtle gestures to large-scale body movements, which fully reflect the diversity of actions in drama performances. The frame rate of the video is 24 frames per second (fps), which ensures the fluency of action and the accuracy of detail capture. The resolution of the video is 320×240 pixels. The duration of the video is between 2.35 seconds and 65.46 seconds.

<i>Parameter</i>	<i>Value</i>
Number of Action Categories	105
Video Frame Rate	24 fps
Video Resolution	320×240
Video Duration	2.35~65.46s
Parameter	Value
Number of Action Categories	105

Table 1: Data set parameters.



Figure 6: A video data set of drama performance action.

Figure 6 shows some video frames in the data set. The data set covers different lighting conditions, background complexity and changes in actors' costumes, which provides rich training samples for the model.

4.2 Experimental Results

Rendering speed is an important index for assessing the model's real-time performance. Figure 7 shows a continuous drama performance modelling action frame, which is used as the input of the rendering speed test.

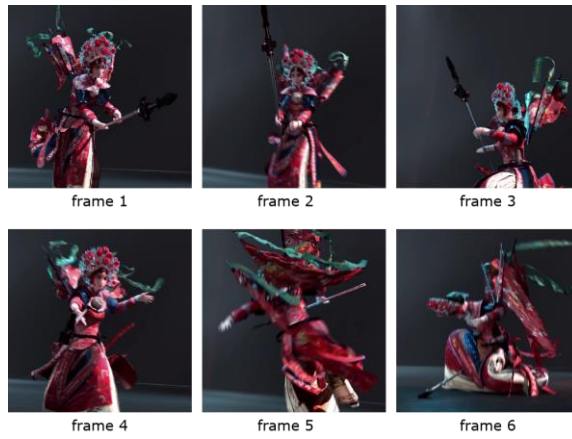


Figure 7: Modeling effect of drama performance.

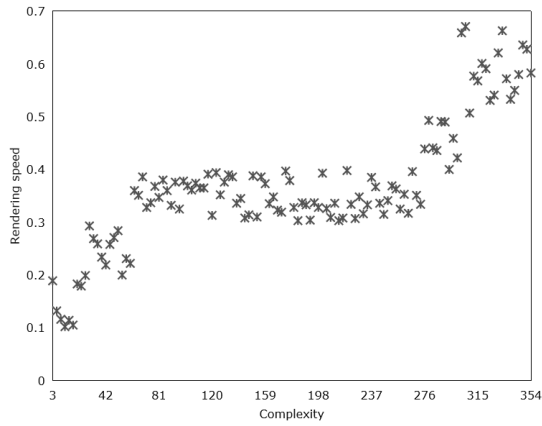


Figure 8: Rendering speed.

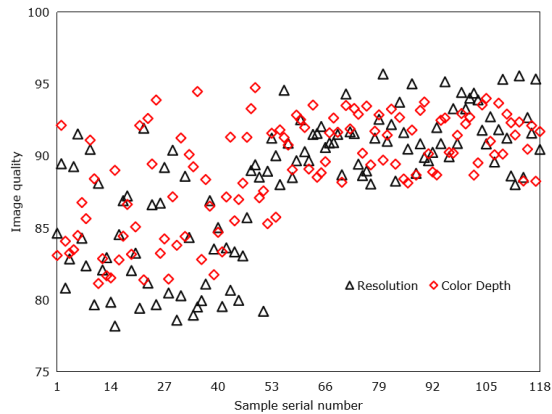


Figure 9: Image quality.

Figure 8 shows the rendering time of the algorithm under different test complexity. With the increase in test complexity, rendering time shows a certain growth trend. However, even at the highest complexity, the rendering time remains within an acceptable range. This shows that the model can effectively handle complex theatrical performances while maintaining efficient operation.

Figure 9 shows the performance of the image generated by the algorithm in terms of resolution and colour depth. From the resolution point of view, the details of the image generated by the algorithm remain intact and the edges are smooth. In terms of colour depth, the algorithm accurately restores the colour information of the original video without colour distortion or colour blocking.

4.3 Discussion

In terms of rendering speed, the efficiency of the model benefits from its optimized algorithm design and parallel processing mechanism. Through reasonable algorithm optimization, the model can significantly improve the processing speed while ensuring accuracy. The parallel processing mechanism enables the model to process multiple video frames at the same time, which further improves the rendering efficiency.

In terms of image quality, the performance of the model benefits from its advanced deep learning architecture and a large number of training data. Through the deep neural network, the model can learn the inherent laws of drama performance. At the same time, a large number of training data enable the model to be better generalized to unseen scenes.

However, some improvements were also found in the experiment. For example, in extreme lighting conditions, such as too dark or too bright environment, the performance of the model is reduced, and the image quality is affected to some extent. This may be because there are few samples of extreme light conditions in the training data, which leads to the lack of adaptability of the model to such scenes. In the future, the performance of the model in these scenes can be further improved by increasing training samples under extreme light conditions.

5 CONCLUSIONS

In the art of drama performance, action is an important means for actors to shape their roles. This article aims to improve the analysis effect of drama performance action through advanced technical means. By comprehensively applying CAD technology, big data analysis method and deep learning algorithm, a comprehensive system integrating motion capture, data analysis and visual display is constructed.

Firstly, this article describes the current situation of action analysis in drama performance. Then, the application of CAD technology in action modelling of drama performance is expounded, including key links such as human joint division, skeleton construction and action simulation. Simultaneously, the significance of big data in analyzing drama performance actions is examined. Through the collection, organization, and analysis of extensive video data, insights into action characteristics and patterns are uncovered.

In the aspect of system construction, a drama performance action analysis system based on CNN is designed and implemented. The system can automatically capture the action in drama performance, extract the action features through CNN, and then generate the drama performance action of virtual characters. The system also integrates the VITON network, which realizes virtual fitting and action display from rough to fine.

The experimental part assesses the performance of the system by comparing the rendering speed and image quality under different complexity. The results show that the system can maintain high processing speed and image quality when dealing with complex actions, different lighting conditions and background changes.

However, there are some limitations in this study. Under extreme lighting or intricate backgrounds, system performance requires enhancement. Future research can bolster applicability by augmenting training data for such scenarios and employing advanced image processing techniques.

6 ACKNOWLEDGEMENT

Handan Philosophy and Social Science Planning Project: Research on the Promoting Effect of Drama Performance Practice on the Mental Health of College Students in Handan Colleges and Universities, Project No. 2022063; School-level project seedling project of Handan University: Research on the Intervention Effect of Drama Performance Practice on College Students' Social Anxiety and Depression, Project No. XS2022407.

Li Han, <https://orcid.org/0009-0009-6441-9621>

Hongtao Niu, <https://orcid.org/0009-0005-3286-860X>

REFERENCES

- [1] Albert, J.-A.; Owolabi, V.; Gebel, A.; Brahms, C.-M.; Granacher, U.; Arrnrich, B.: Evaluation of the pose tracking performance of the azure Kinect and Kinect v2 for gait analysis in comparison with a gold standard: A pilot study, *Sensors*, 20(18), 2020, 5104. <https://doi.org/10.3390/s20185104>
- [2] Ben, X.; Ren, Y.; Zhang, J.; Wang, S.-J.; Kpalma, K.; Meng, W.; Liu, Y.-J.: Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 2021, 5826-5846. <https://doi.org/10.1109/TPAMI.2021.3067464>
- [3] Elhoseny, M.: Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems, *Circuits, Systems, and Signal Processing*, 39(2), 2020, 611-630. <https://doi.org/10.1007/s00034-019-01234-7>
- [4] Gurbuz, S.-Z.; Amin, M.-G.: Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring, *IEEE Signal Processing Magazine*, 36(4), 2019, 16-28. <https://doi.org/10.1109/MSP.2018.2890128>
- [5] Jiang, D.; Li, G.; Sun, Y.; Kong, J.; Tao, B.: Gesture recognition based on skeletonization algorithm and CNN with ASL database, *Multimedia Tools and Applications*, 78(21), 2019, 29953-29970. <https://doi.org/10.1007/s11042-018-6748-0>
- [6] Khare, S.-K.; Blanes, V.-V.; Nadimi, E.-S.; Acharya, U.-R.: Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations, *Information Fusion*, 102(1), 2024, 102019. <https://doi.org/10.1016/j.inffus.2023.102019>
- [7] Koch, S.-C.; Riege, R.-F.; Tisborn, K.; Biondo, J.; Martin, L.; Beelmann, A.-L: Effects of dance movement therapy and dance on health-related psychological outcomes, A meta-analysis update, *Frontiers in Psychology*, 10(1), 2019, 1806. <https://doi.org/10.3389/fpsyg.2019.01806>
- [8] Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; Zhu, S.; Hu, W.: Rethinking the competition between detection and reid in multiobject tracking, *IEEE Transactions on Image Processing*, 31(1), 2022, 3182-3196. <https://doi.org/10.1109/TIP.2022.3165376>
- [9] Mekruksavanich, S.; Jitpattanakul, A.: Lstm networks using smartphone data for sensor-based human activity recognition in smart homes, *Sensors*, 21(5), 2021, 1636. <https://doi.org/10.3390/s21051636>
- [10] Michielli, N.; Acharya, U.-R.; Molinari, F.: Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals, *Computers in Biology and Medicine*, 106(1), 2019, 71-81. <https://doi.org/10.1016/j.compbiomed.2019.01.013>

- [11] Mittal, A.; Kumar, P.; Roy, P.-P.; Balasubramanian, R.; Chaudhuri, B.-B.: A modified LSTM model for continuous sign language recognition using leap motion, *IEEE Sensors Journal*, 19(16), 2019, 7056-7063. <https://doi.org/10.1109/JSEN.2019.2909837>
- [12] Mujahid, A.; Awan, M.-J.; Yasin, A.; Mohammed, M.-A.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K.-H.: Real-time hand gesture recognition based on deep learning YOLOv3 model, *Applied Sciences*, 11(9), 2021, 4164. <https://doi.org/10.3390/app11094164>
- [13] Nweke, H.-F.; Teh, Y.-W.; Mujtaba, G.; Garadi, M.-A.: Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions, *Information Fusion*, 46(1), 2019, 147-170. <https://doi.org/10.1016/j.inffus.2018.06.002>
- [14] Pareek, P.; Thakkar, A.: A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, *Artificial Intelligence Review*, 54(3), 2021, 2259-2322. <https://doi.org/10.1007/s10462-020-09904-8>
- [15] Qi, J.; Jiang, G.; Li, G.; Sun, Y.; Tao, B.: Intelligent human-computer interaction based on surface EMG gesture recognition, *IEEE Access*, 7(1), 2019, 61378-61387. <https://doi.org/10.1109/ACCESS.2019.2914728>
- [16] Sreenu, G.-S.-D.-M.-A.; Durai, S.: Intelligent video surveillance: a review through deep learning techniques for crowd analysis, *Journal of Big Data*, 6(1), 2019, 1-27. <https://doi.org/10.1186/s40537-019-0212-5>
- [17] Wang, L.; Huynh, D.-Q.; Koniusz, P.: A comparative review of recent Kinect-based action recognition algorithms, *IEEE Transactions on Image Processing*, 29(1), 2019, 15-28. <https://doi.org/10.1109/TIP.2019.2925285>
- [18] Zhang, H.-B.; Zhang, Y.-X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.-X.; Chen, D.-S.: A comprehensive survey of vision-based human action recognition methods, *Sensors*, 19(5), 2019, 1005. <https://doi.org/10.3390/s19051005>
- [19] Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8), 2019, 1963-1978. <https://doi.org/10.1109/TPAMI.2019.2896631>
- [20] Zhang, Z.; He, T.; Zhu, M.; Sun, Z.; Shi, Q.; Zhu, J.; Lee, C.: Deep learning-enabled triboelectric smart socks for IoT-based gait analysis and VR applications, *NPJ Flexible Electronics*, 4(1), 2020, 29. <https://doi.org/10.1038/s41528-020-00092-7>