www.cad-journal.net
Computer-Aided Design AND APPLICATIONS

Intelligent Construction of Animation Scenes and Dynamic Optimization of Character Images by Computer Vision

Nan Zhang[1,2], Han Meng[3] and Mingyu Ju[4]

[1,4]School of Film, Jilin Animation Institute, Changchun 130013, China,
[1]ZhangN2000@cuc.edu.cn, [4]workjmy@163.com
[2]School of Theater, Film and Television, Communication University of China, Beijing 100020, China,
ZhangN2000@cuc.edu.cn
[3]School of Animation Industry, Jilin Animation Institute, Changchun 130013, China,
smallfish524422@gmail.com

Corresponding author: Nan Zhang, ZhangN2000@cuc.edu.cn

Abstract. The aim of this article is to investigate an intelligent approach for constructing animation scenes and dynamically optimizing character images driven by computer vision, with a particular emphasis on its potential applications in Chinese animated film production. To accomplish this objective, the study integrates multi-modal emotion recognition techniques, enhancing the emotional impact of animations by simultaneously processing facial expressions and voice emotions of characters. Utilizing state-of-the-art convolutional neural networks (CNN) and long short-term memory (LSTM) models, the research intelligently analyzes and refines animation scenes and character depictions. The findings reveal that this strategy notably elevates the emotional ambiance of the animated scenes and enhances the dynamic expressiveness of the characters, thereby contributing to technological advancements in Chinese animated filmmaking. This study not only shows the potential of computer vision technology in animation production but also provides a new direction for the modernization process of animated films in China. Through the integration of technological innovation and artistic expression, it is expected that Chinese animated films will present a more wonderful visual feast in the future.

Keywords: Animated Film; Computer Vision; Multimodal; Emotional Expression; Role Image
DOI: https://doi.org/10.14733/cadaps.2024.S25.233-248

1 INTRODUCTION

In the past few years, Chinese animated films have achieved significant success both domestically and internationally, garnering widespread acclaim for their artistic merit, technological advancements, and engaging narratives, alongside impressive box office performance. From the delicate emotional description in Big Fish & Begonia to the modern interpretation of traditional myths

in I AM THE DESTINY, and to the reinterpretation of classic legends in White Snake: Origin, China's animated films are gradually getting rid of the past stereotypes and showing a new look of diversification and internationalization. In the global animation market, China's animated films have gradually changed from a follower to a leader, providing a unique oriental perspective for the global audience. Driven by computer vision technology, many innovative research results have emerged in the field of animation production. These achievements not only improve the efficiency and quality of animation production but also bring a broader creative space for animation creators. Bao [1] introduced intelligent algorithms into the construction process of 3D graphics engine animation scenes, greatly improving the efficiency and intelligence level of scene construction. Intelligent algorithms can be used to automatically generate and optimize scene layouts. Early intelligent scene layouts were usually manually completed by designers, but now intelligent algorithms can automatically calculate the optimal layout scheme by analyzing the relationships between scene elements. This can not only save a lot of labor costs but also ensure that the layout of the scene is more reasonable and beautiful. Intelligent algorithms can achieve intelligent placement and combination of objects in a scene. In 3D graphics engines, the positional relationship and combination of objects are crucial for the overall effect of the scene. Intelligent algorithms can automatically calculate the optimal placement and combination method based on the shape, size, material, and other properties of objects, making the scene more realistic and vivid.

In computer animation, precise modeling and animation design can present the complex dynamic behavior of characters with ball and socket joints and three-dimensional scissor-like structures. By utilizing advanced physics engines and dynamic simulation techniques, it is possible to simulate the real motion patterns of characters in the virtual world, making their movements more natural and smooth. Meanwhile, by adjusting the visual characteristics of characters, such as materials, lighting, and textures, their visual expression and realism can be further enhanced [2]. Eom et al. [3] analyzed the visual motion system model predictive control based on physics for character animation. In this type of animation, the motion of characters is not only controlled by the animator but also constrained by the laws of physics, presenting a more realistic and natural effect. As an advanced control method, visual motion system model predictive control provides a more accurate and efficient control method for physics-based character animation. The core idea of visual motion system model predictive control is to establish a character's motion model and predict and control the model based on real-time visual information. In physics-based character animation, this means that we need to first build a physical model that can accurately describe the character's motion behavior. This model needs to take into account factors such as the mass, inertia, and force conditions of the characters, as well as their interactions.

The modeling methods for urban animation scenes are also constantly evolving. Machine learning provides powerful tools for modeling animated city scenes, making the modeling process more efficient and accurate, and able to generate more realistic and vivid scenes. Feng et al. [4] reviewed computer graphics methods used for modeling animated city scenes from the perspective of machine learning. In the model construction phase, we utilize these features to construct a 3D model. Finally, the model is transformed into a visualized animated scene through rendering techniques. In terms of feature extraction, early computer graphics methods often relied on manually designed feature descriptors. This is not only time-consuming and labor-intensive, but also difficult to capture the complexity and diversity of data. The VR animation interaction system builds a virtual environment that allows users to experience and interact with the virtual world firsthand. In the Internet of Things, studying character modeling and behavior control in virtual reality animation interaction systems, as well as the application of real-time image processing technology is of great significance for improving user experience and system performance. Real-time image processing technology plays a crucial role in virtual reality animation interaction systems. Real-time image processing technology can achieve efficient processing and optimization of images in virtual environments, improving image clarity and realism. Gan et al. [5] updated the state and behavior of virtual characters in real time, achieving seamless integration with the real world.

In the field of computer graphics, cloud-based character scene modeling, rendering, and animation production have been research hotspots in recent years. With the development of cloud

computing technology and the improvement of computer performance, Goswami [6] brings users a more realistic visual experience through efficient rendering and animation technology. Goswami reviewed the relevant technologies and applications of cloud-based character scene modeling, rendering, and animation in computer graphics. By collecting real-world character and scene data and utilizing technologies such as 3D scanning and point cloud processing, highly realistic virtual character and scene models are constructed in the cloud. These models can accurately reflect the shape, expressions, and actions of characters, as well as the layout, lighting, and other information of the scene, providing a foundation for subsequent rendering and animation. The advancement of deep learning and computer vision technology provides strong technical support for generating video animations from a single static image. Hu et al. [7] explored the methods, applications, and prospects of generating video animations from a single static image on social media based on intelligent computing. The video animation generation method based on intelligent computing mainly relies on deep learning and image processing techniques. By training deep neural network models, useful information such as contours, colors, textures, etc. can be extracted from static images. Then, use this information to infer and predict the motion trajectory and dynamic changes of objects in the image. Through a series of calculations and synthesis, a continuous and dynamic video animation is ultimately generated. Jing and Song [8] discussed the application of 3D reality technology and CAD in animation design, as well as the importance of computer-aided design. 3D reality technology is a technology that can create and present 3D virtual environments, allowing users to immerse themselves in a realistic virtual world. In animation design, 3D reality technology can accurately simulate the shape, material, and lighting effects of objects, making the animation scene more realistic and three-dimensional. Designers can use 3D modeling software to create various complex objects and scenes and present high-quality images and videos through rendering techniques. In animation design, CAD technology can assist designers in precise size measurement, shape analysis, and design optimization. Designers can use CAD software to draw 2D drawings and convert them into 3D models. With the help of CAD technology, designers can more accurately grasp the proportions and details of objects, improving the accuracy and efficiency of design.

Beyond advancing China's animation filmmaking techniques, this research contributes to the industry's innovative growth. It explores how computer vision and multimodal emotion recognition can streamline production, raising the bar for animated films. Furthermore, this article delves into market demands and audience feedback, both domestically and internationally, to chart a path for enhancing the global appeal of Chinese animated films. This holistic approach aims to expand overseas markets and offer a more diverse animated film catalog to a worldwide audience.

Before further exploration, it is necessary to clarify the main innovations of this study, which constitute the unique significance of this study:

(1) In this study, the multimodal emotion recognition model combining CNN and LSTM is applied to the intelligent construction of animation scenes and the dynamic optimization of character images.

(2) By constructing a multi-modal emotion recognition model, this study realizes the synchronous processing of voice and facial expression data and emotion recognition.

(3) By combining with computer vision technology, this research has realized the efficient and intelligent construction of animation scenes.

(4) This study also pays attention to the innovative development of China's animated films under the background of globalization. By analyzing the current situation and demand for China's animated films, this article puts forward targeted solutions.

Firstly, this article analyzes the current situation and demand for Chinese animated films in scene construction and role dynamic optimization. Then, aiming at the limitations of the existing technology, a multi-modal emotion recognition model combining CNN and LSTM is constructed, and its effectiveness and superiority are verified by experiments. The experiment uses real animated movie data sets to carry out experiments and compares the performance of different algorithms in scene intelligent construction and role dynamic optimization.

2 RELATED WORK

In the creation of animated films, multi-style adaptation, and visual attention-based character detection are two crucial techniques. They not only enrich the visual presentation of animation but also enhance the audience's understanding and emotional investment in the characters. Kim et al. [9] explored the application and significance of these two technologies in animated films. The adaptation of multiple styles into animated films brings endless creative possibilities. Traditional animated films often adopt a unified visual style, while multi-style adaptations break through this limitation by integrating different artistic styles and elements together. This adaptation can be reflected in the design of characters, the arrangement of scenes, and the overall visual atmosphere. The protagonist in an animated movie may have a cute cartoon-style appearance, while the background is delicately depicted in a realistic style. This style of collision not only brings a brand new visual experience to the audience but also makes the characters and scenes more vivid and interesting. The significance of multi-style adaptation lies in its ability to highlight the characteristics and emotions of the character. By changing the appearance, actions, and expressions of characters, their personality traits, emotional states, and inner world can be emphasized. In the dynamic optimization design of computer-aided animated character images, spline curves play a crucial role as an important mathematical tool. Spline curves, with their unique properties such as smoothness, controllability, and flexibility, provide strong support for the dynamic optimization design of animated character images. Li [10] discussed the application of spline curves in this field and the advantages they bring. Spline curves play a crucial role in the design of animated character contours and forms. Through spline curves, designers can precisely control the shape and changes of character contours, achieving smooth and natural transitions. This fine control allows animated characters to maintain fluency and coherence during their movements, greatly enhancing the viewing and realism of the animation. Spline curves can also help designers quickly modify and adjust animated character images. In the process of animation production, it is often necessary to modify and optimize character images multiple times. By parameterizing spline curves, designers can easily adjust the shape and position of the curves, thereby achieving rapid modification of character images. This flexibility greatly improves the efficiency and quality of animation production.

In terms of dynamic optimization of character images, Li and Li [11] focus on improving the realism and expressiveness of character images. By finely processing and analyzing character images, more facial features and details are extracted, making virtual characters more vivid and realistic. At the same time, it uses dynamic optimization technology to adjust and optimize the actions and expressions of characters in real-time, making them more natural and smooth. 2D digital animation has become an important medium for inheriting and showcasing folk stories. Among them, the dynamic optimization of character images is crucial for improving the viewing and narrative effects of animation. Mayowa [12] explored how to develop dynamic optimization of 2D digital animated character images for folktale storytelling. The characters in folk stories often have distinct personalities, and their behavior, language expression, and psychological state are all key to the development of the story plot. Therefore, in the dynamic optimization process of character images, delicate animation expressions are used to showcase the inner world of characters. Smooth movements can enhance the coherence and visual appeal of animation, allowing the audience to better immerse themselves in the story plot. Advanced animation techniques and software can be used to finely process character movements, achieving smooth and natural transitions of movement.

The dynamic optimization of synthesized character images aims to finely adjust and optimize them through algorithms and technical means, making them more realistic, natural, and in line with actual scenes. Paulin and Ivasic [13] reviewed and analyzed the dynamic optimization application of synthetic human body images in computer vision, exploring their development history, technological status, and future trends. Early synthetic images of characters often had obvious flaws and unnatural features, making it difficult to meet the requirements of realism. By training a large amount of data, this model can learn the features and patterns of character images, thereby making more accurate adjustments and optimizations to synthesized images. Pose estimation technology can analyze the posture and movements of characters, making the movements in synthesized images more natural

and smooth. Expression synthesis technology can generate different expressions according to needs, enhancing the diversity and expressiveness of synthesized images. Artificial intelligence is gradually penetrating into every aspect of our lives, and the field of animation and film production has also been deeply influenced. The emergence of intelligent construction technology for animation scenes not only greatly improves the production efficiency of animated movies, but also brings unprecedented cutting-edge visual effects to the audience. Reddy et al. [14] analyzed the power of artificial intelligence in the animation revolution, which has played a certain role in cutting-edge visual effects in movies. Compared with early intelligent animation production methods, current intelligent construction technology can greatly shorten the production cycle, reduce labor costs, and improve the accuracy and realism of scenes. The emergence of intelligent construction technology for animation scenes has not only improved the production level of animated movies but also brought audiences a more colorful visual feast.

Group animation design has become a highly anticipated research field. Group animation design involves a large number of character actions, interactions, and collaborations, which are crucial for showcasing authentic and natural group behavior. Traditional group animation design methods often rely on complex rules and preset animation sequences, making it difficult to cope with complex and ever-changing scenes and real-time interaction needs. Therefore, Tian et al. [15] explored group animation design techniques based on artificial intelligence algorithms to improve the realism and naturalness of animations. Among them, the group animation design technology based on an artificial fish swarm algorithm has received much attention. The artificial fish school algorithm is an optimization algorithm that simulates the behavior of natural fish schools. In group animation design, artificial fish swarm algorithms can be applied to character path planning, behavioral decision-making, and interactive collaboration, enabling characters to exhibit more realistic and natural group behavior. T ü men and Sezgin [16] explored the application and advantages of computer vision-driven dynamic programming in offline sketch scene segmentation and recognition. Computer vision technology provides powerful support for offline sketch scene segmentation and recognition. By utilizing techniques such as deep learning, image processing, and pattern recognition, it is possible to train models that can automatically recognize and segment sketch scenes. These models can learn the features and patterns of sketch scenes from a large amount of training data, and accurately segment and recognize new sketch scenes in practical applications.

With the rapid development of deep learning technology, Generative Adversarial Networks (GANs) have become a powerful tool for generating highly realistic images and videos. In the field of voice-driven facial animation, the application of GANs enables us to generate facial animations that match speech signals, making the expressions of virtual characters more vivid and natural. Facial animation images generated solely by GANs often have some issues, such as unnatural expressions and stiff movements. In the dynamic optimization process, Vougioukas et al. [17] used facial keypoint detection algorithms to extract and locate key points in the generated facial animation images. By comparing the key point positions of real faces, it can fine-tune the generated images to make their expressions more natural. Secondly, motion estimation techniques such as optical flow can be used to smooth facial animations between consecutive frames. Traditional modeling methods often rely on manual operations and complex physical calculations, which are inefficient and difficult to handle complex scenes and actions. Machine learning-based modeling methods have brought new possibilities for modeling three-dimensional dynamic images of characters in film and television animation due to their efficiency, accuracy, and automation. Attitude estimation is another important step in the modeling process. Through machine learning algorithms, Wang [18] trained a model that can accurately estimate human posture. These models can automatically recognize the joint positions and motion trajectories of the human body based on input image or video data. Based on this pose information, a three-dimensional character model with realism and dynamism can be further generated. Through machine learning algorithms, it is possible to reconstruct a complete surface of a person from sparse point cloud data. These algorithms can learn and understand the geometric structure and texture information of character surfaces, thereby generating high-quality 3D character models. These models can be further applied to the production of movies and animations, giving characters vivid appearances and actions.

With the rapid development of computer vision and deep learning technology, depth cameras have shown great potential in real-time animation scene reconstruction of dynamic human scenes. A depth camera can capture depth information in three-dimensional space, enabling Xu et al. [19] to reconstruct more realistic and three-dimensional animated scenes and character dynamics in real-time. A depth camera obtains depth information for each point in the scene by emitting infrared light and receiving reflected light. This depth information can be used to generate 3D point clouds, thereby constructing a 3D model of the scene. Compared to traditional 2D images, depth cameras provide richer and more accurate information, enabling us to better understand and reconstruct the 3D world. In the reconstruction of dynamic scenes of characters in real-time animation scenes, depth cameras play a crucial role. Secondly, depth cameras can also be used for real-time reconstruction of dynamic backgrounds in animated scenes. By capturing the depth information of the scene, it can generate a 3D model of the background and update it in real time based on the movement and position changes of the characters. The reconstruction of this dynamic background makes the animation scene more realistic and natural, providing the audience with a richer visual experience. Ocean animation, as an important branch of 3D animation, has unique visual charm and artistic value. In recent years, 3D animation automatic generation technology based on Unreal Engine 4 (UE4) engine has provided new solutions for the creation of ocean animation. Zhang [20] explored the application research of 3D animation automatic generation technology based on the UE4 engine in ocean animation. Simulating waves is a key technical challenge in ocean animation. Traditional wave simulation methods often rely on complex physical models and computational processes, making it difficult to achieve real-time generation and efficient rendering. The 3D animation automatic generation technology based on the UE4 engine can automatically generate realistic wave effects by introducing physics-based simulation algorithms and real-time rendering technology.

3 THE CURRENT SITUATION OF CHINESE ANIMATION MOVIES

Chinese animated films have undergone significant development and transformation in recent years, not only achieving impressive box office results but also demonstrating impressive strength in visual effects, storytelling, and technological applications. This section will analyze the current situation of Chinese animated films.

3.1 The Achievements of China's Animation Films

In the past few years, Chinese animated films have made remarkable achievements in domestic and foreign markets. From the box office data, many Chinese animated films rank among the top in the global box office, showing strong market appeal. In terms of visual effects, Chinese animated films constantly break through the limitations of traditional production technology and use advanced computer-generated image (CGI) technology and animation rendering technology to bring viewers a more realistic and shocking visual experience. In story narration, China animated films gradually got rid of the single narrative mode and began to try diversified themes and narrative techniques, providing more choices for the audience.

3.2 Technology and Strategy

In the process of making animated films in China, scene construction and role dynamic optimization have always been the focus of the production team. Through computer vision technology, the production team can capture the movements and expressions of the characters more accurately and realize a more natural dynamic effect of the characters. Using a deep learning algorithm, the production team can quickly generate realistic scene images and improve production efficiency and quality.

At present, there are still some challenges in the scene construction and role dynamic optimization of Chinese animated films. For example, it is still a difficult problem to realize more natural and true emotional transmission in the expression of role emotions. In addition, in the

construction of complex scenes, how to balance the scene details and rendering efficiency is also a problem that needs to be solved.

3.3 Case Presentation

Case 1: I AM THE DESTINY

I AM THE DESTINY is an animated film with the traditional myth of China as the background. Through precise motion capture and expression rendering technology, this film makes the movements and expressions of Nezha and other characters more natural and vivid. Especially in emotional expression, the film realizes the true transmission of the character's emotions through delicate expression changes and tone adjustment. Figure 1 shows some animation scenes in the movie I AM THE DESTINY.

Figure 1: "I AM THE DESTINY" animation scene.

Figure 2: "Legend of Deification" animation scene.

Case 2: Legend of Deification

Legend of Deification is another animated film with the theme of ancient Chinese mythology. From mountains and rivers to palaces and castles, every detail has been carefully designed and rendered. In the dynamic optimization of characters, the film makes the movements of characters such as Jiang Ziya smoother and more natural through innovative action design and expression capture technology. In terms of emotional expression, the film successfully shapes the inner world and emotional changes of the characters through rich facial expressions and changes in pronunciation and intonation. Figure 2 shows some animation scenes in the movie Legend of Deification.

Case 3: New Gods: Yang Jian

New Gods: Yang Jian is an animated film featuring Yang Jian, an ancient mythological figure in China. The film has shown excellent strength in scene construction and role dynamic optimization. Through careful design and rendering, the production team combined the elements of ancient mythology with modern aesthetics, presenting a unique visual effect.

In the aspect of dynamic optimization of characters, this film makes the movements and expressions of Yang Jian and other characters more natural and vivid through accurate motion capture and expression rendering technology. In the battle scene, Yang Jian's moves and movements were smooth and powerful, which brought a strong visual impact to the audience. In terms of emotional expression, the film also realizes the true transmission of the character's emotions through delicate expression changes and tone adjustments.

The film reinterprets the image and story of Yang Jian, a traditional mythical figure, through a unique perspective and narrative technique. Yang Jian in the film is no longer a deity on high, but a character full of flesh and blood and emotion. His interaction and growth with characters such as Aquilaria sinensis gave the audience a deeper understanding of the inner world and emotional changes of this role. Figure 3 shows some animation scenes in the movie "New Gods: Yang Jian".

Figure 3: "New Gods: Yang Jian" animation scene.

From the above cases, we can see that Chinese animated films have made remarkable progress in scene construction and role dynamic optimization. These advances are not only reflected in the technical level and innovation but also in the in-depth excavation of the story connotation and role

emotion. These successful cases not only set a new benchmark for Chinese animated films but also provide a useful reference for future animation production.

4 THE FUSION OF COMPUTER VISION AND MULTIMODAL EMOTION RECOGNITION TECHNOLOGY

This section will discuss how to construct a multimodal emotion recognition model that integrates CNN and LSTM to realize the synchronous processing of voice and facial expression data. In order to realize the synchronous processing of speech and facial expression data, a multimodal emotion recognition model combining CNN and LSTM is constructed in this study. Firstly, the model uses CNN to process facial image data and extract facial features. At the same time, LSTM is used to process speech data and capture the time sequence information in speech. Then, the extracted facial features and speech time sequence information are fused to form a multimodal feature vector. Finally, the multi-modal feature vectors are classified by a fully connected layer to realize the synchronous recognition of facial expressions and voice emotions.

CNN is a deep learning algorithm, especially suitable for processing image data. In facial expression recognition, CNN can automatically extract facial features, such as the shape and position information of eyes, mouth, and eyebrows, by learning a large number of facial image data. These features are very important for recognizing facial expressions. Through multi-layer convolution and pooling operation, CNN can gradually abstract the high-level features of the face, thus realizing accurate recognition of facial expressions. The CNN model structure is shown in Figure 4.

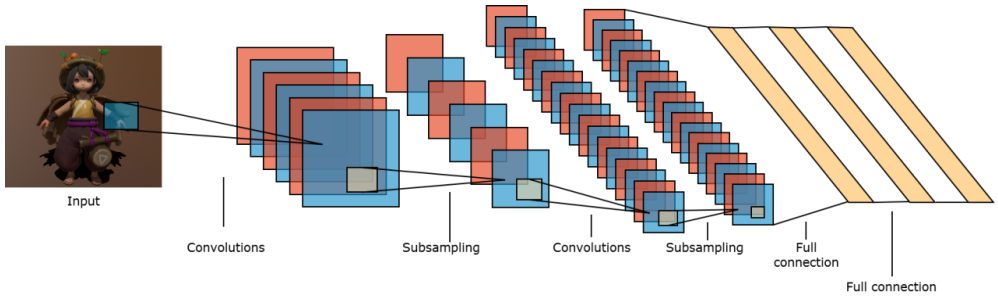

Figure 4: CNN model structure.

When using CNN to process facial image data, it is necessary to preprocess the image first, including converting it into a format suitable for network input, resizing, and normalizing. Then, the CNN model structure is designed according to the specific task, and the model parameters are continuously optimized by training the labeled facial image data to minimize the difference between the prediction and the actual tag. After the training is completed, the performance of the model is assessed on the independent test set to ensure its accuracy and reliability. Finally, the trained CNN model is applied to practical facial image processing tasks, such as feature extraction, identity recognition, or expression classification, so as to realize automatic and intelligent processing. The whole process needs to be adjusted and optimized according to specific tasks and data conditions, and the performance of the model should be continuously improved with the help of new CNN architecture and technology.

The pool layer is used to reduce the dimension of the feature map, and the common maximum pool operation is as follows:

$$y_{ij} = \max_{i-1\ S\le m\le iS,\ j-1\ S\le n\le jS} x_{mn} \tag{1}$$

Where S is the pool step size? This operation is helpful to reduce the calculation amount of the model, improve the calculation efficiency, and enhance the robustness of the features.
The fully connected layer is used to map the feature map to the output space, and its formula is:

$$y = Wx + b \tag{2}$$

Where W is the weight matrix and b is the bias term? In interior design evaluation, these outputs may correspond to different design quality grades or styles.
Cross entropy loss function for multi-classification problems;

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log\ p_{ic} \tag{3}$$

Where N is the number of samples, C is the number of categories, y_{ic} is the real label (0 or 1), and p_{ic} is the probability of model prediction. By minimizing this loss function, the model is encouraged to predict the labels of interior design images more accurately.

Random Gradient Descent (SGD) is used to update the model weights, and its formula is:

$$W_{t+1} = W_t - \eta \nabla L\ W_t \tag{4}$$

Where η is the learning rate and $\nabla L\ W_t$ is the gradient of the loss function concerning the weight? In each iteration, the weight is adjusted according to the gradient of the loss function about the weight so as to optimize the performance of the model step by step.

Dropout is used to prevent the model from over-fitting, and its formula is:

$$r_j^{\ l} \sim \text{Bernoulli}\ p \tag{5}$$

$$\tilde{y}^{\ l} = r^{\ l} * y^{\ l} \tag{6}$$

Where $r_j^{\ l}$ is a vector of 0 or 1 randomly generated with probability p, and $\tilde{y}^{\ l}$ is the output of the l layer. By turning off some neurons at random, it reduces the parameters of the model and helps to prevent over-fitting on the training data.

In China's animated films, the facial expressions of characters are an important means to convey emotions. By applying CNN technology, the facial expressions of characters can be automatically identified and classified, such as happiness, sadness, and anger. This will help the production team to grasp the emotional changes of the role more accurately and bring a more realistic viewing experience to the audience.

LSTM is a special RNN, which has memory ability and can deal with long-term dependencies in sequence data. In speech emotion recognition, LSTM can effectively capture the time sequence information in speech signals, such as the changes in pitch, sound intensity, and speech speed. This time series information is very important for recognizing emotions in speech.

The LSTM was originally designed to guarantee information integrity during transmission and stabilize error gradients during reverse propagation. It manages the influence of varying time-point information in memory by selectively filtering out irrelevance. The output gate, regulated by the Sigmoid function, determines which information to release from the cell's current state, as illustrated in Figure 5.
The input gate comprises two distinct components. Firstly, it identifies fresh information that must be incorporated into the cell's state. Secondly, it establishes the proportion of this novel data to be merged into the memory cell's state. Its inputs encompass the prior hidden layer state, denoted as h_{t-1}, and the present input, designated as x_t. The computation for this process follows the formula:

$$\tilde{C}_t = \text{Tanh}\ W_c h_{t-1} + U_c x_t + b_c \tag{7}$$

Where W, U_c, b_c represents the weight and offset values.

The output gate determines the quantity of information that will be transmitted from the neuron's internal state to its external state at the present t_i moment, utilizing the following expression:

$$O_t = \delta\ U_o h_{t-1} + W_o x_t + b_o \tag{8}$$

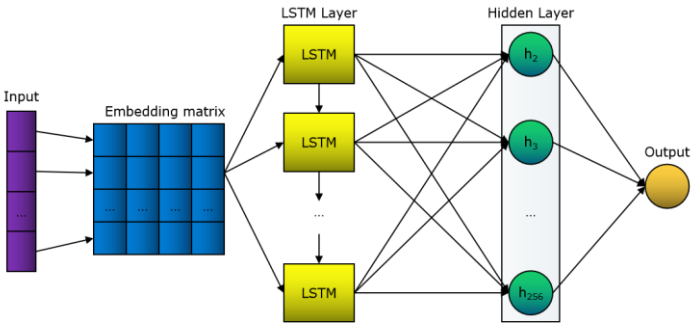

Figure 5: Working mechanism of LSTM.

For the nonlinear function of the output gate, the sigmoid function is chosen. As O_t the value nears 1, a higher volume of information is transmitted from t_i the internal state c_t to t_i external state h_t in the present moment. The emotional disposition associated with characteristic words is as follows:

$$tendency = \frac{1}{n}\sum_{i=1}^{n} sim\ \ word, seed_{1i}\ \ - \frac{1}{m}\sum_{i=1}^{n} sim\ \ word, seed_{2j} \tag{9}$$

Upon discovering that when $a_i \neq 0$ the corresponding vector X_i serves as the support vector, the decision function can be reformulated as follows:

$$f\ \ x\ \ = \mathrm{sgn}\left(\sum_{i=1}^{M} a_i y_i K\ \ x, x_i\ \ + b\right) \tag{10}$$

In this context, M denotes the number of support vectors present.

Instead of calculating similarity between concepts, the focus will shift to determining similarity between sememes:

$$Sim\ \ p_1, p_2\ \ = \frac{a}{d+a} \tag{11}$$

This formula p_1 stands for sememe 1, p_2 represents sememe 2, d denotes the path length p_1, p_2 within the initial tree structure, and a serves as an adjustable parameter.

By utilizing LSTM technology, we can achieve automated recognition and categorization of characters' voice emotions. This aids the production team in precisely capturing emotional shifts in voice, enhancing the audience's viewing experience.

During model training, extensive labeled data, including facial images, voice recordings, and corresponding emotional tags, are employed to refine the model's parameters. Through repeated iterative training, the model progressively learns effective methods for extracting emotional features from facial images and voice data, enabling precise emotion recognition.

5 EXPERIMENTAL DESIGN AND RESULT ANALYSIS

In this section, we aim to contrast the performance of the proposed multi-modal emotion recognition model, which integrates CNN and LSTM, against single-modal CNN and LSTM algorithms in emotion

recognition tasks. To comprehensively evaluate the effectiveness of these algorithms, we consider experimental metrics such as accuracy, recall, F1 score, and processing time. The study employs a unified dataset for training and testing both the multi-modal model and the single-modal algorithms. This dataset encompasses facial expression images, corresponding voice recordings, and emotional tags. To ensure the credibility of our findings, we have partitioned the dataset into training, validation, and test sets.

We use the CNN algorithm to process facial expression image data and extract facial features. The LSTM algorithm is used to process speech data and capture the time sequence information in speech. For the multimodal emotion recognition model, the extracted facial features and speech time sequence information are fused and classified through the full connection layer. In the process of model training, cross-entropy loss function and gradient descent optimization algorithm are used to optimize model parameters.

Figure 6 shows the comparison results of different algorithms in accuracy. The multi-modal emotion recognition model proposed in this article is significantly better than the single-modal CNN and LSTM algorithms in accuracy. This shows that the multimodal information of facial images and voice data can provide richer emotional clues.

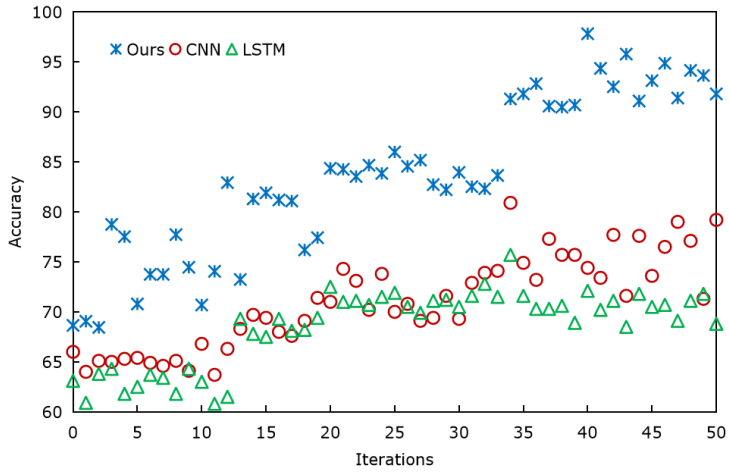

Figure 6: Accuracy performance.

Figure 7 presents a comparative analysis of recall rates among various algorithms. Recall, in this context, quantifies an algorithm's proficiency in correctly identifying positive samples, specifically emotional categories. The multimodal emotion recognition model also shows good performance in recall, which shows that it can capture and identify samples of different emotion categories more comprehensively.

Figure 8 displays a comparative overview of the F1 scores achieved by distinct algorithms. The F1 score, serving as the harmonic mean of precision and recall, provides a holistic evaluation of an algorithm's performance. The multi-modal emotion recognition model is also superior to the single-modal algorithm in the F1 score, which shows that it has good performance in accuracy and comprehensiveness.

Figure 9 shows the comparison results of different algorithms in processing time. Processing time is particularly important for real-time application scenarios. Although the processing time of the multi-modal emotion recognition model is slightly longer than that of the single-modal algorithm, it is still within the acceptable range. This shows that the multimodal emotion recognition model has a certain real-time performance while maintaining high accuracy.

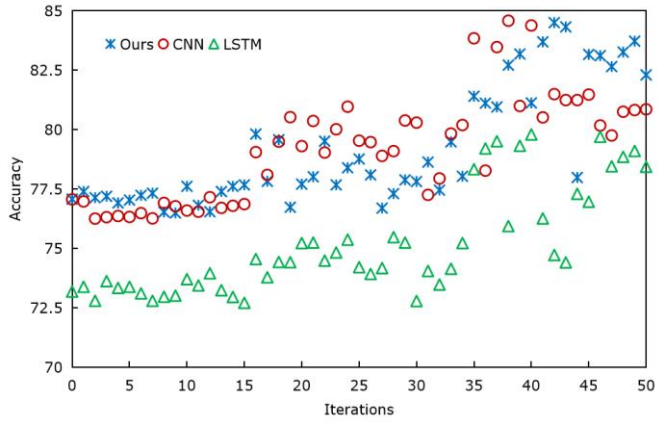

Figure 7: Recall performance.

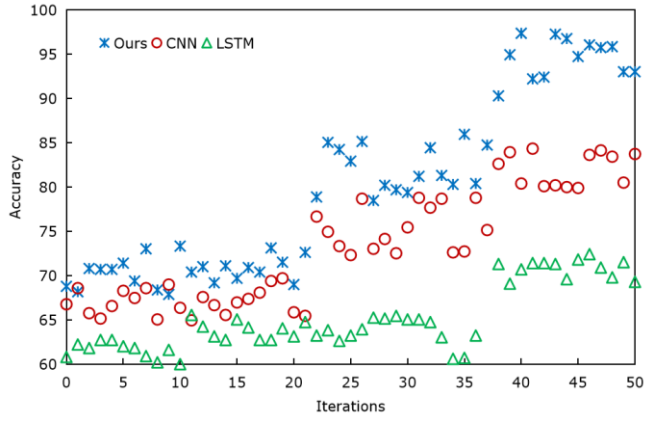

Figure 8: F1 score performance.

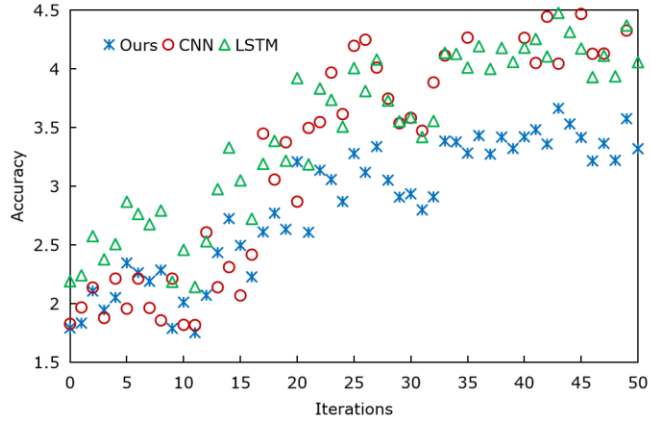

Figure 9: Processing time performance.

Comparative experiments reveal that the multi-modal emotion recognition model introduced in this article outperforms single-modal CNN and LSTM algorithms in terms of accuracy, recall, and F1 score. This underscores the importance of multimodal information, combining facial images and voice data, for achieving more comprehensive and precise emotion recognition outcomes. While the processing time for the multi-modal model is marginally longer than that of single-modal algorithms, it still exhibits adequate real-time capabilities, making it well-suited for practical applications. Future endeavors could focus on further optimizing the model's structure and parameters to enhance its processing efficiency.

6 DISCUSSION

The multi-modal emotion recognition model that combines CNN and LSTM discussed in this article shows remarkable advantages in the synchronous processing of voice and facial expression data. In the process of making animated films, the application of this technology will provide new possibilities for intelligent scene construction and dynamic optimization of character images.

The emotional expression of characters in animated films is an important means to convey emotions and shape the characters' images. In the early animation production process, animators needed to manually adjust the facial expressions and gestures of characters to express their emotions based on the script and the director's intentions. This way is limited by the animator's subjective understanding and technical level, and it is difficult to ensure the accurate transmission of emotions. The application of multimodal emotion recognition technology can accurately identify the current emotional state of the character by automatically analyzing the voice and facial expression data. The application of multi-modal emotion recognition technology can identify the emotional atmosphere needed by the current scene by automatically analyzing the voice and facial expression data, and intelligently adjusting the light, color, and other elements of the scene.

The integration and application of multi-modal emotion recognition technology and other technologies will bring more possibilities for animated film production. For example, the combination with virtual reality (VR) technology can realize the real-time interaction and emotional resonance between the audience and the animated characters; Combined with augmented reality (AR) technology, animated characters and scenes can be integrated into the real world. The combination with the Internet of Things (IoT) technology can realize the linkage between animated films smart homes and other devices.

7 CONCLUSIONS

This research endeavors to investigate strategies for intelligent animation scene construction and dynamic character image optimization, leveraging computer vision techniques with a focus on their applicability in Chinese animated filmmaking. By integrating multi-modal emotion recognition technology, the study aims to achieve synchronized facial expressions and voice emotions for animated characters, thereby imparting a heightened sense of realism and emotional depth to animated narratives. Comparative experiments demonstrate the superiority of the proposed multi-modal emotion recognition model over single-modal CNN and LSTM algorithms in terms of accuracy, recall, and F1 score.

As Chinese animation undergoes modernization, technological advancements, and their application emerge as pivotal forces. This study not only elevates the level of intelligence in animation production but also marks significant progress in scene development and character image enhancement. With computer vision as a driving force, animation scenes can adapt intelligently based on plot twists and character emotions, delivering richer, more captivating visual experiences.

In conclusion, this study offers technical underpinnings for enhancing animated films through intelligent scene construction and dynamic character optimization. Looking ahead, with technology's relentless progression, Chinese animated films are poised to unfold an even more spectacular and diverse visual tapestry.

8 ACKNOWLEDGEMENTS

This work was supported by The Youth Fund Projects for Humanities and Social Sciences Research of the Ministry of Education in 2021, Innovation Research on the Construction of National Cultural Image in the Overseas Communication of Chinese Animation (2004-2020), No.21YJC760106; The Planning Fund Projects for Humanities and Social Sciences Research of the Ministry of Education in 2023, Research on the History of Chinese Red Comics in the Past 100 Years (1921-2021), No.23YJA760071; The key project of vocational education research in Jilin Province in 2023, Research on the Construction Path of Animation Vocational Education Industry College from the Perspective of Cluster Development of Animation Industry in Jilin Province, No. 2023XHZ019.

Nan Zhang, https://orcid.org/0009-0006-1746-3396
Han Meng, https://orcid.org/0009-0007-1979-7944
Mingyu Ju, https://orcid.org/0009-0009-9967-8885

REFERENCES

[1] Bao, W.: The application of intelligent algorithms in the animation design of 3D graphics engines, International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 13(2), 2021, 26-37. https://doi.org/10.4018/IJGCMS.2021040103
[2] Chen, X.; Jiang, H.; Xuan, T.; Huang, L.; Liu, L.: Designing deployable 3D scissor structures with ball-and-socket joints, Computer Animation and Virtual Worlds, 30(1), 2019, e1848. https://doi.org/10.1002/cav.1848
[3] Eom, H.; Han, D.; Shin, J.-S.; Noh, J.: Model predictive control with a visuomotor system for physics-based character animation, ACM Transactions on Graphics (TOG), 39(1), 2019, 1-11. https://doi.org/10.1145/3360905
[4] Feng, T.; Fan, F.; Bednarz, T.: A review of computer graphics approaches to urban modeling from a machine learning perspective, Frontiers of Information Technology & Electronic Engineering, 22(7), 2021, 915-925. https://doi.org/10.1631/FITEE.2000141
[5] Gan, B.; Zhang, C.; Chen, Y.; Chen, Y.-C.: Research on role modeling and behavior control of virtual reality animation interactive system in Internet of Things, Journal of Real-Time Image Processing, 18(4), 2021, 1069-1083. https://doi.org/10.1007/s11554-020-01046-y
[6] Goswami, P.: A survey of modeling, rendering and animation of clouds in computer graphics, The Visual Computer, 37(7), 2021, 1931-1948. https://doi.org/10.1007/s00371-020-01953-y
[7] Hu, T.; Liang, C.; Min, G.; Li, K.; Xiao, C.: Generating video animation from single still image in social media based on intelligent computing, Journal of Visual Communication and Image Representation, 71(1), 2020, 102812. https://doi.org/10.1016/j.jvcir.2020.102812
[8] Jing, Y.; Song, Y.: Application of 3D reality technology combined with CAD in animation modeling design, Computer-Aided Design and Applications, 18(S3), 2020, 164-175. https://doi.org/10.14733/cadaps.2021.S3.164-1-175
[9] Kim, H.; Lee, E.-C.; Seo, Y.; Im, D.-H.; Lee, I.-K.: Character detection in animated movies using multi-style adaptation and visual attention, IEEE Transactions on Multimedia, 23(1), 2020, 1990-2004. https://doi.org/10.1109/TMM.2020.3006372
[10] Li, L.: Application of cubic b-spline curve in computer-aided animation design, Computer-Aided Design and Applications, 18(S1), 2020, 43-52. https://doi.org/10.14733/cadaps.2021.S1.43-52
[11] Li, L.; Li, T.: Animation of virtual medical system under the background of virtual reality technology, Computational Intelligence, 38(1), 2022, 88 https://doi.org/10.1111/coin.12446
[12] Mayowa, A.-S.: Development of a 2D digital animation for Yorùbá Folktale narrative, International Journal of Art, Culture, Design, and Technology (IJACDT), 9(1), 2020, 47-61. https://doi.org/10.4018/IJACDT.2020010104
[13] Paulin, G.; Ivasic, K.-M.: Review and analysis of synthetic dataset generation methods and techniques for application in computer vision, Artificial Intelligence Review, 56(9), 2023, 9221-9265. https://doi.org/10.1007/s10462-022-10358-3

[14] Reddy, V.-S.; Kathiravan, M.; Reddy, V.-L.: Revolutionizing animation: unleashing the power of artificial intelligence for cutting-edge visual effects in films, Soft Computing, 28(1), 2024, 749-763. https://doi.org/10.1007/s00500-023-09448-3
[15] Tian, Y.; Li, Y.; Pan, L.; Morris, H.: Research on group animation design technology based on artificial fish swarm algorithm, Journal of Intelligent & Fuzzy Systems, 38(2), 2020, 1137-1145. https://doi.org/10.3233/JIFS-179475
[16] Tümen, R.-S.; Sezgin, M.: Segmentation and recognition of offline sketch scenes using dynamic programming, IEEE Computer Graphics and Applications, 42(1), 2021, 56-72. https://doi.org/10.1109/MCG.2021.3069863
[17] Vougioukas, K.; Petridis, S.; Pantic, M.: Realistic speech-driven facial animation with gans, International Journal of Computer Vision, 128(5), 2020, 1398-1413. https://doi.org/10.1007/s11263-019-01251-8
[18] Wang, Y.: 3D dynamic image modeling based on machine learning in film and television animation, Journal of Multimedia Information System, 10(1), 2023, 69-78. https://doi.org/10.33851/JMIS.2023.10.1.69
[19] Xu, L.; Cheng, W.; Guo, K.; Han, L.; Liu, Y.; Fang, L.: Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera, IEEE Transactions on Visualization and Computer Graphics, 27(1), 2019, 68-82. https://doi.org/10.1109/TVCG.2019.2930691
[20] Zhang, L.: Application research of automatic generation technology for 3D animation based on UE4 engine in marine animation, Journal of Coastal Research, 93(SI), 2019, 652-658. https://doi.org/10.2112/SI93-088.1