Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study on the Application of Motion Capture Technology and Animation Generation Methods in Film and Television Performances

Na Zhao[1,] ✉

[1] *College of Ministry of Sports, Xi'an Aeronautical University, Xi'an, Shaanxi, 710089, China*

ABSTRACT
Artificial Intelligence Generated Content (AIGC), as a computer technology mainly characterized by intelligent content generation, has caused significant changes in film and television performances and creations, and has greatly broadened the creation and development space of film and television performances. In this paper, we use motion capture technology to obtain the character movement data in film and television performances, and combine it with the skeletal motion data generation algorithm to realize the mapping of skeletal motion data. Using ResNet-122 as the backbone network, a 3D action pose estimation model is constructed by combining multi-view and multi-feature fusion networks. Based on the 3D action pose estimation sequence, the character animation generation model is constructed by combining GAN and action detail attention mechanism, and the action detail feature loss function is designed to improve the generalization ability of the animation generation model. In order to verify the effectiveness of the above method, data analysis is carried out through simulation verification. The average value of PCP3D index of the 3D action pose estimation model is 98.37, which is 0.28 percentage points higher than the sub-optimal model, and the average joint position error is only 16.07 mm. The animation generation model combining GAN and the action detail attention mechanism has the values of animation generation diversity and richness index of 5.104 and 3.997, respectively, and the animation generation diversity and richness indexes of the animation generation model combining GAN and the action detail attention mechanism are 5.104 and 3.997, respectively. 3Ds MAX software can map the generated animation sequences into the virtual space, providing assistance for optimizing the motion design of film and television performances.

Keywords: motion capture technology, action pose estimation, GAN, attention mechanism, animation generation

1. Introduction

Motion capture technology is a widely used technology in the field of movies, games and animation in

✉ Corresponding author.
E-mail address: zhaona89722390@126.com (N. Zhao).
Received 15 April 2024; Revised 05 July 2024; Accepted 25 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-071

recent years [1-2]. By capturing actors' movements and expressions and then applying them to virtual characters or special effects, it can produce realistic effects and bring new visual experiences to the audience. The principle of animation generation by motion capture technology is mainly divided into three steps: data acquisition, data processing and data application [3-6].

Data acquisition is the first step of motion capture technology, mainly through the sensor and camera to record the actor's movements and expressions. The sensors can monitor the position and posture of the actor's body parts, while the camera can capture the actor's facial expression and eye changes. During the data acquisition process, actors wear specific sensor devices in order to accurately capture their movements and expressions [7-10]. Data processing is the process of processing and transforming the collected data. The accuracy of the data is ensured by cleaning and calibrating the data to eliminate noise and errors. Then, mathematical models and algorithms are used to analyze and reconstruct the data to transform the actor's movements and expressions into those of virtual characters or special effects [11-14]. The process of data processing requires the support of computer software and hardware in order to efficiently process large amounts of data. Data application is the process of applying the processed data to movie and television production. By matching and integrating actors' movements and expressions with virtual characters or special effects, the virtual characters or special effects can be made to show similar movements and expressions as those of the actors [15-18]. This application can realize a variety of amazing effects in movies, games and animations, making the audience feel a more realistic and vivid visual experience [19-20].

Literature [21] constructs a film and television animation production system by combining motion capture technology and proposes a simple correction method to eliminate the instability of animation frames. It is emphasized that the film and television animation system based on motion capture technology can effectively improve the animation production effect. Literature [22] explores the application of digital media technology in FATA post-production, discusses the animation production process based on MCT, and demonstrates the revolutionary potential of MCT in 3D animation technology. Literature [23] introduces the development of motion capture technology and addresses the problems exposed in the development of China's current 3D motion capture technology based on the realization of 3D motion capture technology to promote the development of China's film and television animation. Literature [24] emphasizes the important role of motion capture, pointing out that it has revolutionized 3D animation production techniques, significantly simplifying animation production operations and improving the quality of animation production. [25] revealed that motion capture technology is widely used, and based on the Unity3D platform, optical positioning technology is applied to accurately locate the position of the camera and the virtual character, realizing real-time virtual filming with multi-person motion capture. Literature [26] describes the application of motion capture technology in film and television works to bring excellent visual effects. It discusses the development and improvement of motion capture technology in recent years in comic book movies, and puts forward the views on motion capture. [27] explored the method of combining group animation motion capture with virtual reality technology. And a crowd motion capture system based on virtual reality technology is constructed, which proves that the method has an important role in group animation motion capture and animation design. Literature [28] affirms the effect of motion capture technology in film and television animation, and reveals the difference between virtual capture technology and traditional 3D animation, and describes the motion capture technology, as well as the application of the body in film and television animation.

The rapid development of intelligent generation technology has brought great changes to the film and television industry, changing the traditional way of film and television creation, and enabling extensive and in-depth cooperation between people and technology. In order to further improve the optimization effect about action design in film and television performances, this paper extracts the character action data of film and television performances by motion capture technology, and obtains their character action sequences by the three-dimensional action gesture estimation model fused with

multi-feature and multi-view angle. Then the generative adversarial network is combined with the attention mechanism to establish a model for character animation generation, and the optimization of loss function is proposed to enhance the animation generation effect. Based on the generated character animation sequences, the introduction of Unity3D and 3Ds MAX software can realize the rendering of animated actions, which provides inspiration for improving the action design in film and television performances.

2. Motion Capture Technology in Film and Television Performance

Due to the rapid development of modern computer technology, computer animation technology is also constantly changing, especially the use of three-dimensional technology, making film and television performances more realistic, giving people a sense of immersion. In the current production process of film and television performances, people through computer technology to produce high-quality images, making the animation picture more vivid. Performance motion capture technology is through the film and television performances of moving objects, people and other things in the three-dimensional movement of the trajectory of real-time capture, and then through the computer technology to capture the picture analysis, so as to give people a new visual experience.

2.1. Motion Capture Technology

2.1.1. Motion Capture Technology Flow. Motion capture technology is to set up trackers in the key parts of the moving objects, record the process of object movement, and then get three-dimensional spatial data after computer processing, the technology is widely used in digital protection, games, animation, film and television performances, ergonomics, simulation training, virtual reality, and other research fields [29].

Motion capture technology can be divided into expression capture and body capture according to different application methods, and can be divided into mechanical, acoustic, electromagnetic and optical motion capture according to different working principles. This study adopts an optical motion capture system, which uses a high sensitivity near-infrared CMOS sensor to capture and record the near-infrared reflecting marker ball through the settings of camera focal length, aperture, grayscale color gradient, and frame angle in different capture environments to accurately capture the motion trajectory of the key point where the marker ball is located. The application process of motion capture technology is shown in Figure 1, which mainly includes spatial calibration, motion capture, data processing, data analysis and editing conversion.

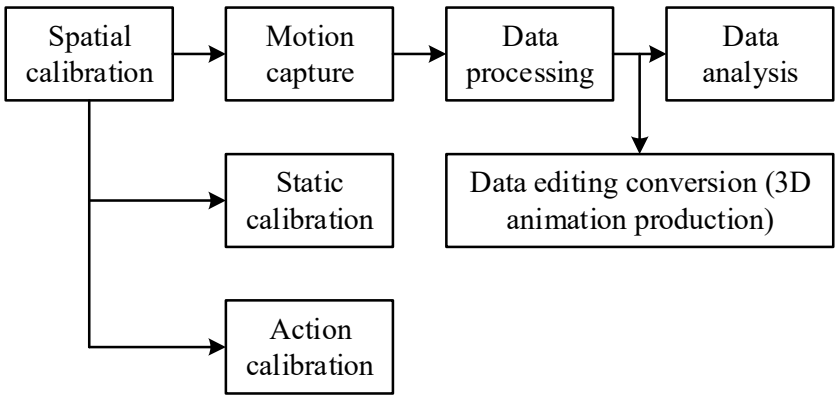

Fig. 1. Motion capture system operation flow

2.1.2. Skeletal motion data generation. In the operation process of motion capture technology, one of the important steps is to process the motion data, which aims to better obtain the skeletal motion

data. Based on this, this paper proposes a skeletal motion data generation algorithm to help better obtain motion sequences and provide data support for motion pose estimation [30].

Let $X_{1,T}^{A}=\left[p_1,\cdots,p_t,\cdots,p_\tau\right]^{'}$ be a skeletal motion sequence data of character A, where $p_t=\left[x_{t,1},y_{t,1},z_{t,1},\cdots,x_{t,J},y_{t,J},z_{t,J},v_t\right]$ is frame t of the sequence, $J=22$ is the number of joint points, and $v_t\in R^4$ is the root joint motion in frame t.

Firstly, the motion segment $X_{1,\tau}^{A}$ is encoded into the encoder F_{enc} to get the hidden variables, i.e.:

$$\varphi_A=F_{enc}(P_{1:T}^{A},W_{enc}) \tag{1}$$

Where W_{enc} is the parameter of encoder F_{enc} and φ_A is the hidden variable of $X_{1.T}^{A}$. After obtaining the hidden variables, the hidden variables φ_A are regularized once using the fully connected layer $F_{qlinear}$, i.e.:

$$q_A=\frac{F_{qlinear}(\varphi_A,W_{qlinear})}{\|F_{qlinear}(\varphi_A,W_{qlinear})\|_2} \tag{2}$$

where the regularization serves to make q_A a unit vector.

The skeleton information is then encoded using the fully connected layer F_{skel}. Let $\overline{p}^{j}$ denote the joint coordinates of the character's skeleton in T pose with root joint coordinates 0. $\overline{p}^{parent(j)}$ is the father node of the j th node, and let:

$$\overline{s}^{j}=\overline{p}^{j}-\overline{p}^{\text{parent}(j)} \tag{3}$$

where $\overline{S}_A=\left[\overline{s}_A^2\quad\overline{s}_A^3\quad\cdots\quad\overline{s}_A^J\right]\in R^{3\times J}$ is the skeleton information for role A, order:

$$\overline{S}_A^{'}=\frac{F_{skel}(\overline{S}_A,W_{skel})}{\|F_{skel}(\overline{S}_A,W_{skel})\|_2} \tag{4}$$

The above equation is also a regularization operation.

Next, the skeleton encoded information and the regularized hidden variables are connected and then go through the fully connected layer F_{Δ} to get:

$$q_{A_\Delta}=\frac{F_{\Delta}(Concat(\varphi_A,\overline{S}_A^{'}),W_{\Delta})}{\|F_{\Delta}(Concat(\varphi_A,\overline{S}_A^{'}),W_{\Delta})\|_2} \tag{5}$$

Then the order:

$$q_A=Hamiton(q_A,q_{A_\Delta}) \tag{6}$$

$$p_{1,T}^{A^{'}}=FK(q_{A_\Delta},\overline{S}_A) \tag{7}$$

where the Hamiton operation is essentially the multiplication of 2 quaternions, which in 3D space can be understood as fine-tuning q_A with rotational information q_{A_Δ}. The role of forward kinematics is to assign rotational information to a given skeleton $\overline{S}_A$.

Finally, the network training loss function in this paper is defined as:

$$L_{roon}(X_{1:T}^{A},X_{1:T}^{A^{'}})=\frac{\lambda_{recon}}{3TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\|p_t^{j.x_{1,T}^{A}}-p^{j.x_t^{A}}\|_2 \tag{8}$$

where the training coefficients $\lambda_{recon}=0.5$, $p_t^{i,x_{1,T}^{\Lambda}}$ in this paper are the coordinates of the j th node

in the tth frame of motion clip $X_{i,\tau}^{A}$.

2.2. Estimation of movement gestures

In this paper, a 3D pose estimation method based on multi-view fusion is used to realize the action pose estimation of characters in film and television performances, and the specific architecture of the model is shown in Fig. 2. Firstly, the basic motion sequence is obtained by the skeletal motion data generation algorithm given in the previous paper [31]. Then single-view 2D joint point estimation is performed for each viewpoint, followed by 3D spatial projection as well as 3D joint point reconstruction. The overall network first extracts 2D joint points for each viewpoint using HRNet network, which uses ResNet-122 as the backbone network. After extracting the 2D joint point feature representations for each viewpoint, the multi-view fusion network is first utilized to fuse the joint points in different viewpoints to reduce the joint point estimation error in each single viewpoint. Then the multi-stage fusion network is used to fuse the joint points in different stages to achieve optimization. Then the joint points of each viewpoint are projected to the 3D space to obtain the 3D joint point estimation by 3D reconstruction.

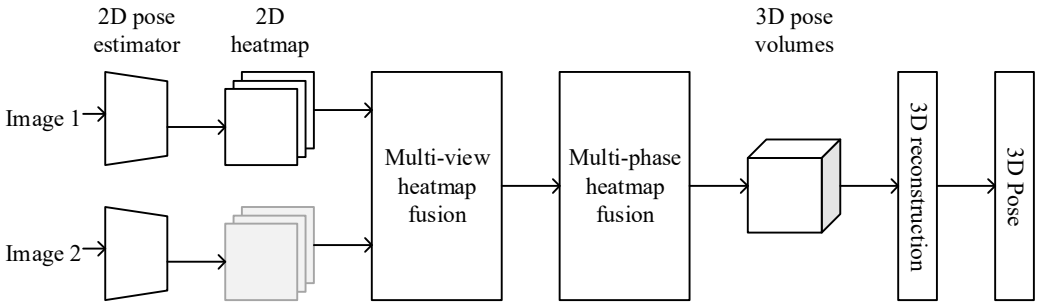

Fig. 2. Three-dimensional human image estimation network

2.2.1. Multi-perspective convergence networks. The thermogram of view v is denoted as H^{v}, and the response at position x on the thermogram is denoted as $H^{v}(x)$. The antipodal line corresponding to position x on view u is denoted as $I^{u}(x)$, and consists of a series of discrete points on H^{u}. The exact position of the antipodal line can be calculated from the camera parameters. The feature fusion formula is then:

$$\hat{H}^{v}(x)=\lambda H^{v}(x)+\frac{1-\lambda}{N}\sum_{u=1}^{N}\max_{x^{'}\in I^{u}(x)}H^{u}(x^{'}) \tag{9}$$

where $\hat{H}$ denotes the fused heatmap and N is the number of views that influence the current view. Parameter λ is used to balance the weights of the current viewpoint and other viewpoints in the fusion.

The above fusion strategy leads to degraded fusion results when the detection results of some viewpoints are incorrect, because the viewpoints with originally better detection quality may be disturbed by the bad detection results, which leads to degraded 2D pose estimation results.

To address this problem, a weighted learning network is proposed to assign weights to the human pose detection results for each viewpoint to characterize its detection quality.

The network takes a heat map as input and regresses to obtain the weights for each detection result ω^{u}. Therefore, the multi-view fusion method is rewritten as:

$$\hat{H}^{v}(x)=\omega^{v}H^{v}(x)+\sum_{u=1}^{N}\omega^{u}\max_{x^{'}\in I^{u}(x)}H^{u}(x^{'}) \tag{10}$$

Based on the thermogram output from the 2D human pose estimation network, two features are

further extracted for weight prediction. The first feature is the appearance embedding feature which encodes properties such as energy distribution on the thermogram. The second feature is the geometric embedding feature, which encodes the consistency of the detection results between multiple views. These two features are of complementary nature. Multi-view weighted fusion human learning networks can be trained end-to-end jointly with 2D human pose estimation networks without the need for separate supervised signals for the weighted results.

2.2.2. Multi-stage converged networks. In order to obtain better 2D joint points, after the multi-view fusion network, a multi-stage fusion method is applied to each single-view estimation network to optimize the accuracy of the 2D pose estimation using the multi-stage cascade structure thereby improving the accuracy of the 3D pose estimation. The specific fusion method is:

$$h_t = w_t * y_t + (1 - w_t) * \hat{y}_t \tag{11}$$

where y_t is the value at the specific pixel location in the 2D articulation node in stage t and $\hat{y}_t$ is the value at the specific pixel location in the projected 2D articulation node in stage t. $w_t \in [0,1]$ is a weight parameter to be learned for better fusion of two different 2D joints to get the value h_t at the specific pixel location in the fused node in stage t. The 2D joints contain multiple channels, each representing a joint, and are fused in the nodes of each channel corresponding to each pixel location using the above method to get a set of fused 2D joints in the end. This is then fed into the next stage and used as input to optimize the original 2D joint points obtained after the 2D pose detector, instead of using the original 2D joint points alone.

3. Animation generation techniques based on motion sequences

With the development of computer technology, computer animation technology is also changing, especially the use of three-dimensional technology, so that the film and television performance animation is more realistic, giving people a sense of immersion. Traditional film and television performance animation production process, people through computer technology to produce high-quality images, making the animation picture more vivid, but there will be action incoherent phenomenon, the audience's viewing experience is not optimistic. Therefore, in order to improve the visual effect of the audience when watching, through the action sequence obtained by motion capture technology, combined with intelligent generation technology can make the animation generation more visualized, the audience's visual experience is more optimistic.

3.1. GAN and Attention Mechanisms

3.1.1. Generating Adversarial Networks. Generative Adversarial Network (GAN) consists of a generator G (generative network) and a discriminator D (discriminative network). The role of the generator is to try to learn the feature distribution $P_{data}(x)$ of the real data x under the guidance of the discriminator, with the aim of fitting the probability distribution $P_z(z)$ of the random noise z to the probability distribution $P_{data}(x)$ of the real data as much as possible, so as to generate similar data with the features of the real data. The discriminator is a discriminator responsible for distinguishing whether the input data is real data x or fake data generated by the generator $G(z)$. The two networks are trained alternately, and constantly fight against each other, so that the ability of the generator and the discriminator is synchronized to improve in the process of continuous iterative optimization until the data generated by the generator network can be fake, and the discriminator can not discriminate between the authenticity of the input data, i.e., the ability of the generator and the discriminator is improved in the process of continuous iterative optimization, which means that the ability of the generator and the discriminator is improved in the process of continuous iterative optimization. Nash equilibrium is reached in the process of continuous iterative optimization [32].

The generator and the discriminator are trained alternately during training, and the role of the discriminator is to be able to have the ability to accurately judge, and its loss function is:

$$\max_D V(D,G) = \mathrm{E}_{x\sim P_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim P_z(z)}\left[\log(1-D(G(z)))\right] \tag{12}$$

Where E represents the mathematical expectation, $G(z)$ is the output of the generator, $P_{data}(x)$ is the probability distribution of the real data, $x \sim p_{data}(x)$ represents the randomly selected data x from the real data set, $P_z(z)$ is the random noise z probability distribution, $D(x)$ and $D(G(z))$ are the discriminant scores of the discriminator on the real data x and the generated data $G(z)$ by the generator, respectively.

The role of the generator is to fit data that can be faked as much as possible by capturing the characteristic distribution $P_{data}(x)$ of the real data x. The discriminator discriminates the real data output as 1, then the generator should maximize the discrimination score of its generated data, so the loss function of the generator is:

$$minV(D,G) = E_{z\sim P_z(z)}[log(1-D(G(z)))] \tag{13}$$

The discriminator needs to maximize its objective function and the generator needs to minimize its objective function. The final objective function of the adversarial generative network is:

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[log\ D(x)] + E_{z\sim P_z(z)}[log(1-D(G(z)))] \tag{14}$$

Adversarial generative networks train the generator and the discriminator alternately, fixing the parameters of the generator updating the discriminator using error back-propagation and gradient descent to maximize the loss function V(D,G) of the discriminator, and locking the parameters of the discriminator updating the parameters of the optimization generator to minimize the objective function V(D,G) of the generator.

3.1.2. Attention mechanisms. The essence of the attention mechanism is the rational allocation of resources according to the importance of information. The attention mechanism will use different degrees of attention to different features in the data, and assign corresponding weights according to the importance of the feature location, so as to improve the network's ability to extract key features [33].

The computational process of the classical attention framework is divided into three stages, Query represents the query vector, Value is generally the target information, and Key is the index of these target information.

The first stage of the attention mechanism is to compute the correlation between Query and different Keys, also known as the attention score, there are different methods for different attention mechanisms, for example, the multiplicative and additive models can be represented as:

$$s(Query, Key) = Query \cdot Key_t \tag{15}$$

$$s(Query, Key) = v_t \tanh(W \cdot Query + U \cdot Key_t) \tag{16}$$

where both W, U and v are learnable parameter matrices.

The second stage normalizes the weight coefficients by mapping the range of values between 0 and 1, i.e:

$$a_t = \mathrm{Softmax}(s_t) = \frac{e^{s_t}}{\sum_{j=1}^{L} s_j} \tag{17}$$

where L represents the word length.

The third stage calculates the attention value, which is generally a multiplication and then accumulation of the Value and the weight corresponding to each Value. I.e:

$$Att = \sum_{i=1}^{lx} a_t \cdot Value_t \tag{18}$$

Attention mechanisms can be categorized into hard and soft attention. Hard attention mechanisms are more aggressive, selecting only the most important or most probable information, while the remaining information is ignored, with weights of 0 and 1. In contrast, soft attention mechanisms assign weights in a more moderate manner, allowing weights to take on a range of values between 0 and 1.

3.2. Character Animation Generation Model

3.2.1. Generating Network Architecture Designs. Combined with the needs and objectives of this study, this paper refers to the representative work in the field of motion style migration, treats the motion content and motion style separately, and uses ST-GCN to model the motion sequence data to build a diversified animation generation model that can be transferred to multiple target style domains. Drawing on the spatial adaptive normalization method in existing research, it is improved and used in 3D human animation generation to realize the application of spatial adaptive normalization method from 2D image space to 3D motion space. The overall framework of the network is shown in Fig. 3, which contains four modules: action generator, discriminator, style encoder and mapping network.

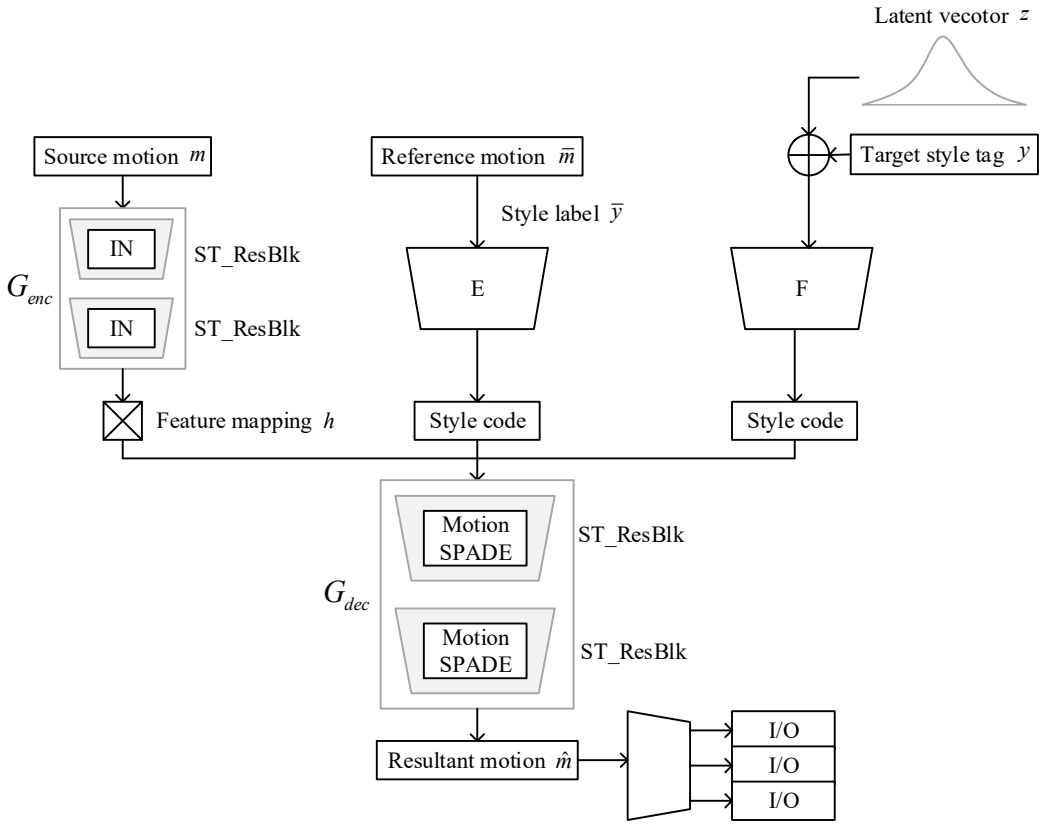

Fig. 3. 3d animation style generation model

Where the action generator is an encoder-decoder structure, improved spatial adaptive normalization methods are used in the decoder of the network. The network model can be used to train the generator to learn to transform the input actions into different stylistic domains more

accurately by adding labeling information of the target stylistic domain.

3.2.2. Attention to motor details. Subtle movements can often convey rich emotions and provide infectiousness for characters in film and television performances, while existing stylized movement generation methods often lack attention to movement details, and shallow features of movements are often ignored. To address this problem, this paper proposes an attention and normalization module called Motion Detail Attention (MD-ATN), which employs a multi-layer strategy to simultaneously acquire global and local features of the motion. The network integrates the MD-ATN module on two ST-GCN residual blocks of the decoder. In order to fully utilize the shallow elements of the network, the MD-ATN module connects the features of the current layer with the downsampled features of the previous layer. To wit:

$$f_{1x}^{*} = D_x(f_1^{*} \oplus f_2^{*} \oplus \cdots \oplus f_x^{*}) \tag{19}$$

where * is c or s for content and style features respectively, x is the current number of layers, D_x represents a bilinear interpolation layer which downsamples the input to the same shape as f_x^{*}; and $\oplus$ denotes a crosstalk operation along the channel dimension.The embedding characteristics of the MD-ATN module at layer x are denoted as:

$$f_x^{s} = MD - ATN(f_x^{c}, f_x^{s}, f_{1x}^{c}, f_{1x}^{s}) \tag{20}$$

where f^{c}, f^{s} and f^{es} are content features, style features and embedding features, respectively.

The generated stylized motion is denoted as:

$$m^{cs} = G_{de}(f_1^{cs}, f_2^{cs}) \tag{21}$$

The MD - ATN module processing can be divided into three steps, firstly, computing the attention graph using shallow and deep content and style features, secondly, computing the weighted mean and variance of style features using the attention graph, and thirdly, adaptively normalizing the content features and adjusting the distribution of the features in order to make the content consistent with the style features.

1) Computation of the Attention Graph. The MD-ATN module computes the attention graph using content and style features. $\overline{f}_{1x}^{*}$ represents the result of f_{1x}^{*} normalized by instances. j, k, l represents 1 × 1 learnable convolutional layers for normalizing mean-variance, respectively. The module performs matrix multiplication between $l(\overline{f_{1x}^{c}})$ and $k(\overline{f_{1x}^{s}})$ to construct the attention graph A using Softmax layers, which are computed as:

$$A = Softmax(l(\overline{f_{1x}^{c}})^{\cdot} \otimes k(\overline{f_{1x}^{s}})) \tag{22}$$

Where $\otimes$ stands for matrix multiplication.

2) Calculation of weighted mean and variance. Attention score matrix A is used for style feature statistics, where the target style feature points are considered as the distribution of all weighted style feature points, and the attention weighted mean M and weighted standard deviation S are denoted as:

$$M = j(f^{s}) \otimes A^{\cdot} \tag{23}$$

$$S = \sqrt{(j(f^{s}) \cdot j(f^{s})) \otimes A^{\cdot} - M \cdot M} \tag{24}$$

which denotes the dot product of elements in $j(f^{s})$.

3) Adaptive Normalization. For each position and each channel of the normalized content feature, the corresponding scales in S and the displacements in M are used to generate the style-migrated

features f_x^{cs}. i.e:

$$f_x^{cs} = S \cdot \overline{f_x^c} + M \tag{25}$$

where $\overline{f}_x^c$ denotes the result after instance normalization for f_x^c.

3.2.3. Loss of motion detail features. The loss function of a GAN plays an important role in supervised generation, discriminator training, facilitating confrontation between the two, and ensuring training stability and convergence. For this reason, the total loss function selected for GAN is formed by the network adversarial loss as well as the edge loss. Namely:

$$L_t = x_i \lambda_a L_a + x_i \lambda_p L_p \tag{26}$$

Where λ_a, λ_p is used to describe the weights corresponding to loss functions L_a and L_p, respectively, and the presence of $\lambda_a + \lambda_p = 1$. L_a is used to describe the network adversarial loss function, which measures the authenticity of the generated complex action images and accelerates network convergence. L_p is used to describe the contour loss, which serves as an improvement of the original $L_1.L_2$ loss function to capture images with a greater sense of detail, while making the images captured by the network consistent with the real complex action images. The Euclidean distance between the feature matrices of the generated complex action image and the real complex action image processed by convolution is used to describe L_p the expression for L_a the:

$$L_a = E_{X \sim P_x}[D(X)] - E_{\tilde{X} \sim P_x}[D(\tilde{X})] + \mu E_{\hat{X} \sim P_x}[(\| \nabla_{\hat{X}} D(\hat{X}) \|_2 - 1)^2 gr] \tag{27}$$

Where $D(\cdot), E(\cdot), \mu$ is used to describe the generator model, mathematical expectation, and penalty coefficient, respectively, and $X, \tilde{X}$ is denoted by $\hat{X}$ for the random sampling result. $P_x, P_{\tilde{X}}$ is used to describe the distribution of complex action target image and generated complex action image, respectively, and $\nabla_{\hat{x}} D(\hat{X})$ denotes the bias result of $D(\hat{X})$ against $\hat{X}$.

The expression of L_p is given by:

$$L_p = \frac{1}{w \times h} E_{\tilde{X}, Y}[\phi(Y) - \phi(\tilde{X})]^2 \tag{28}$$

Where $\phi(\cdot), w, h$ describes the feature map, feature map corresponding to width and height obtained by the network after convolution process.

4. Motion Capture Technology Application and Animation Generation

In the context of the rapid development of computer science and technology, computer animation technology has also made breakthrough progress, especially after the introduction of the three-dimensional concept, the animation of its own three-dimensional sense and the sense of reality has become more obvious, more and more close to the real world. In relation to the existing animation production, people are gradually able to use computers to generate high-quality images, and virtual reality technology is distributed in all aspects of people's daily life. Performance animation relies on motion capture technology to implement the process of capturing the movement of people, animals or objects in three-dimensional motion, and can be digitally analyzed, which belongs to a high-tech. Nowadays, the virtual human animation in film and television animation belongs to the more rapid development of computer graphics research field, the core of which is the motion capture and animation generation technology.

4.1. Basis of Animation Generation Technology

4.1.1. Unity3D Development Engine. Unity3D, as a cross-platform development engine, can be published on PC, web, iOS, Android, PS3 and other platforms at the same time, and is one of the mainstream applications in the fields of 2D and 3D game development, 3D animation and interaction, and 3D architectural displays nowadays. It has a powerful and simple comprehensive editor, which is able to WYSIWYG, save a lot of development time and improve efficiency. This paper is mainly based on this software to realize the 3D modeling application of motion capture.

Unity3D is based on Mono, which can support running on Windows, Linux, Unix, OSX and other platforms. As an integrated development environment, it contains the kernel of the game engine as well as a graphical development interface, which can be visually operated when it is set up, and it supports the use of the three scripting languages of C#, JavaScript, and boo, which can be used to realize different functions by hooking up scripts to the model objects, of which the C# language is the only one. Different functions, of which C# language is widely used because of its rich resources.

In terms of resource import, Unity3D supports the resource import of many mainstream 3D modeling software, such as 3dMax, Maya, Blender, SolidWorks, and other professional 3D software, and skeleton, animation, and texture material import can work with most of the related applications, which makes the creation of a simple 3D model more precise than Unity3D itself. Unity3D will be automatically updated with the changes of resource files.

Unity3D has a built-in PhysX physics engine, which can efficiently and realistically simulate the physical effects of rigid body collision, gravity, and fabric, making the screen real and vivid, and the user experience more intuitive to improve the participation, thus it becomes the most widely used physics engine, which is adopted by many big works.

4.1.2. Three-dimensional modeling techniques. 3D modeling technique is one of the core techniques in the field of digital art and computer graphics and is mainly used to create virtual 3D scenes, objects and characters. This technique can be divided into various methods, such as polygonal modeling, surface modeling, and voxel modeling. In this paper, polygonal modeling and surface modeling are mainly used, and these two methods are widely used in modeling software such as Maya and 3Ds Max.
1) Polygonal modeling is a method of constructing a 3D model using a polygonal mesh, by connecting and combining polygons (usually triangles or quadrilaterals) to form the basic structure of the model. Polygon modeling is suitable for complex and organic surfaces, and can be used to create highly detailed models by adding, deleting, and editing polygons.
2) Surface modeling describes the shape of a model by using mathematical surfaces. This method creates smoother and more detailed surfaces by controlling the curvature and shape of the surface, and is therefore widely used in character modeling and scene modeling where a detailed and natural look is desired.

4.2. Validation of motion pose estimation

4.2.1. Comparative experiments. In order to verify the feasibility of the action-posture estimation network given in this paper in practical applications, the Human dataset is selected to carry out the validation analysis in this paper, and the Percentage of Correct Parts (PCP3D), Precision Rate (PRE), Recall Rate (RE), Average Precision (AP), and Mean Positional Error of Joints (MPJPE) are used as the evaluation indexes. The current mainstream algorithms of each action pose estimation are used as a comparison, and Table 1 shows the comparison results of different algorithms.

As can be seen from the table, the mean value of the PCP3D metrics obtained from the 3D action pose recognition model based on multi-view fusion in this paper is 98.37, which is 0.28 percentage points higher than that of the TEMPO model with the sub-optimal performance, and the precision and recall of the model are 92.61% and 92.84%, which are 2.15 and 1.05 percentage points higher than

the performance of the best existing method, respectively. This indicates that the action data of characters in movie and television performances can be restored from various perspectives by starting from multi-view fusion, and can be accurately restored by combining multi-view and multi-feature fusion. In addition, the model in this paper achieves an average accuracy of 99.85% in 3D action pose estimation, which is 0.06 percentage points higher than the result of the current best method. Although the MPJPE value obtained by this paper's method is 16.07 mm, which is still 3.95 percentage points higher than the optimal method, based on the comprehensive consideration of the accuracy rate, average accuracy and other indexes, it can be concluded that the 3D action pose recognition model based on multi-view fusion has a better performance. The model in this paper can accurately select the correct key points of the skeleton and can produce results closer to the real situation. Applying it to 3D action pose recognition after motion capture in film and television performances helps to better understand the character's action performance in film and television performances, and provides support for optimizing the creation of film and television performances.

Table 1. Comparison results of different algorithms

Method	PCP3D	PRE	RE	AP	MPJPE
VoxelPose	97.15	55.18%	66.38%	99.71%	17.75mm
PlaneSweepPose	97.84	88.24%	90.17%	99.79%	16.83mm
MvP	97.46	72.35%	80.26%	97.52%	15.46mm
Faster VoxelPose	97.68	83.97%	88.43%	99.34%	18.37mm
VoxelTrack	97.23	79.53%	82.51%	99.56%	17.08mm
TEMPO	98.09	90.46%	91.79%	99.69%	16.14mm
Ours	98.37	92.61%	92.84%	99.85%	16.07mm

4.2.2. Ablation experiments. The 3D action pose estimation model proposed in this paper is a combination of Multi-view Fusion Network (MTFN) and Multi-stage Fusion Network (MSFN) to enhance the effect of capturing and estimating character's actions in film and television performances. In order to verify the effectiveness of the above two modules on the 3D action pose estimation model, this paper conducts ablation experiments on 2D and 3D poses on the Human dataset.The results of the ablation experiments for 2D and 3D poses are shown in Table 2 and Table 3, respectively.

As can be seen from Table 2, after the inclusion of the multi-view fusion network module, the average accuracy of the 3D action pose estimation increased by 1.54% compared to the base network. With the inclusion of the multi-stage fusion network module, the average accuracy increased by 2.52% compared to the base network. And with the addition of both multi-view fusion network and multi-node fusion network modules, the average accuracy of 3D action pose estimation reaches 93.96%, which is 11.61 percentage points higher than that of the base network. In Table 3, the baseline model showed a decrease of 4.34 mm in the mean joint position error with the addition of MTFN compared to the original network. With the inclusion of MSFN, the average joint position error decreased by 4.27 mm compared to the original network. And after adding both MTFN and MSFN modules, the average joint position error is minimized to 16.07 mm, which is 34.35% lower than the average joint position error of the base network. In summary, the data in the table show that the 3D action pose estimation model based on multi-view fusion and multi-stage fusion in this paper is applied to motion capture in film and television performances with high accuracy, and the multi-view fusion and multi-stage fusion network module can effectively improve the accuracy of 3D human pose estimation.

Table 2. Experiment results of the experiment of 2D gestures (%)

Backbone	Baseline	Baseline+ MTFN	Baseline+ MSFN	Ours
Root	95.48	95.94	96.17	96.57

Belly	77.26	79.35	79.37	94.98
Nose	86.38	87.74	88.42	96.32
Head	86.29	86.65	86.85	96.45
Hip	79.37	81.38	83.19	96.01
Knee	81.52	82.94	84.54	92.63
Wrist	70.16	73.21	75.36	84.79
Means	82.35	83.89	84.87	93.96

Table 3. Experiment results of the experiment of 3D gestures (mm)

Method	MTFN	MSFN	MPJPE
Baseline	×	×	24.48
Baseline	√	×	20.14
Baseline	×	√	20.21
Baseline	√	√	16.07

4.3. Animation Generation Validation Analysis

4.3.1. Model generation capability. This paper proposes a GAN-based animation generation method, which can quickly generate a realistic and detailed human animation given only the motion sequence of the target movie and television performance. In order to verify the generating ability of the model, this paper carries out experiments based on hidden space sampling and analyzes the experimental results. In this paper, FID, accuracy (Acc), richness (Div) and diversity (Multi) are selected as evaluation indexes. Among them, FID, as a common generative model evaluation index, is used to measure the difference between the sampled human motion distribution and the real human motion distribution, and the smaller its value is, the more realistic the generated human motion is. Accuracy is used to evaluate the accuracy of the categories to which the generated human motions belong, and the larger its value, the more accurate the generated human motions are. Richness is used to evaluate the diversity of the generated movements in all categories, the closer its value is to the real data, the richer the generated human movement categories are. Diversity is used to evaluate the diversity of the generated movements of the same category, and the closer the value is to the real data, the richer the generated movement postures of a specific category are.

In order to fully demonstrate the animation generation capability of the model in this paper, the experiments are carried out simultaneously on three datasets, AMASS, Human and UESTC, and compared with Trans-VAE and CycleGAN methods. It is worth noting that due to the difference in dataset partitioning methods may lead to bias in the experimental results, the experiments were based on the same partitioning method to retrain the Trans-VAE and CycleGAN models on the Human and UESTC datasets, respectively. The animation generation performance evaluation results of different models are shown in Table 4.

From the data in the table, it can be seen that the models in this paper and the CycleGAN model have obvious improvement in the four indexes of FID, Acc, Div, and Multi compared with the Trans-VAE model, which further indicates that the animation model generating ability based on the GAN and the motion detail attention mechanism is better than that based on the single-frame probability distribution modeling method. Further analysis of the data in the table shows that the model in this paper further improves the animation generation ability compared to the CycleGAN model. In summary, the animation style generation model constructed based on the 3D action gesture sequence data obtained by motion capture technology, combined with the GAN model and the action detail attention mechanism has a better generation ability. Therefore, given an action gesture sequence of film and television performances, the model in this paper can generate diversified human motion sequences, providing feasible guidance and decision-making for action optimization of film and television performances.

Table 4. Animation generation performance assessment of different models

Datasets	Index	Truth	Trans-VAE	CycleGAN	Ours
AMASS	FID	0.041	0.175	0.069	0.043
	Acc	0.949	0.892	0.906	0.946
	Div	7.062	6.989	7.015	7.059
	Multi	2.913	2.604	3.078	2.898
Human	FID	0.019	3.084	1.663	0.425
	Acc	0.999	0.923	0.954	0.971
	Div	5.227	4.246	4.469	5.104
	Multi	4.445	2.808	3.472	3.997
UESTC	FID	0.035	2.034	0.431	0.052
	Acc	0.985	0.892	0.927	0.963
	Div	6.774	6.438	6.635	6.766
	Multi	3.889	5.716	4.467	3.905

4.3.2. User research experiments. In order to further evaluate the practicality of this paper's method, 25 users were invited to participate in the research experiment, and each user was required to complete 10 groups of experiments. For each segment of the animation results, the key frames were intercepted, and the users were allowed to mark the positions of the joints they thought should be located by dragging and dropping them with the mouse on the operation interface. The purpose of this experiment is to verify whether the character animation results generated by the method in this paper have good part recognizability, i.e., whether the character's posture and various parts can be effectively recognized.

By recording the user labeling results and comparing them with the true values, the Euclidean distance in image space is calculated for each corresponding joint, and the percentage of joints that are correctly labeled is measured using the Correct Percentage of Keypoints (PCK) metric. PCK calculates the percentage of the number of joints whose normalized distance error between the prediction result of the joints and the corresponding true value is less than a set threshold, and is a commonly used evaluation index for joint detection in character pose recognition. In this experiment, the pixel length of the person's head in the image was used as the data normalization reference, and PCK@0.35 means that the error is not more than 35% of the length of the person's head. Table 5 shows the results of user experiments.

It can be seen that when the PCK threshold is 0.35, the accuracy of user labeling can reach 94.55%, and its accuracy instead decreases by 11.2 percentage points after the threshold exceeds 0.35. Specifically for each body part, its accuracy rate is above 90%, in which for the head, elbow, wrist and knee, which have a clear range and are easy to recognize, their accuracy rates are above 95%. The hip, however, does not have a strong correspondence with the outline of the character like the limbs, and most of the experimental participants did not have a full understanding of the human skeletal structure, so they did not have a clear perception of the location of this part, resulting in a low user marking accuracy rate (around 90%). However, in a comprehensive view, the results of this user research experiment show that the character's posture in the animation generation results of this paper's model has a high degree of recognizability, and different limb parts can be correctly identified, which in turn indicates that its animation generation results are in line with expectations. The diversity and richness of the animation generation results can provide a reliable reference basis for the movement guidance of movie and television performances, and provide assistance for improving the movement expression in movie and television performances.

Table 5. User experiment results (%)

Site	PCK@0.15	PCK@0.25	PCK@0.35	PCK@0.45
Head	56.76	93.27	97.65	95.24
Shoulder	45.27	91.06	94.48	92.18
Orb	53.38	84.34	95.32	88.75
Wrist	42.69	79.83	97.94	82.39
Hip	13.57	46.58	90.23	56.69
Knee	52.65	90.27	95.88	88.78
Ankle	48.79	72.51	90.37	79.42
Means	44.73	79.69	94.55	83.35

4.3.3. Animation generation results. Unity3D is used as the basis for the validation of animation generation after motion capture in film and television performances in conjunction with 3Ds MAX software.AMASS is a large-scale automated motion capture dataset designed to provide rich human motion data to support research in the areas of pose estimation, human body modeling, and so on. The dataset contains motion capture data from multiple sources covering a wide range of poses, movements and human morphologies. The dataset contains a large amount of human motion capture data, in which the motion data is aligned, normalized and parameterized, making it highly compatible with the parametric representation of the animation generation model based on GAN and motion detail attention mechanism in this paper. Therefore, by combining the motion data in the AMASS dataset with the models in this paper, a direct mapping from the motion data to the pose of the human model can be realized, thus facilitating the animation generation of the human model. By utilizing the motion data in the AMASS dataset to drive the parametric model, highly realistic human body animation can be achieved, adding more vividness and naturalness to the performance of virtual characters.

First, each frame in the motion sequence of the dataset is converted one by one into the corresponding model parameters, which control the pose and shape changes of the character model. Then the model is rendered by calibrating the camera parameters to obtain the corresponding 2D image. Finally, the 2D image results of each frame are combined to form a continuous motion picture to show the movement process of the character. Figure 4 shows the sampling results of a number of frames in the character animation synthesis sequence (represented only by the key points of the human motion skeleton), in which the first column is the input image, and the last five columns are from the partial screenshots of the animation frames generated from the two action sequences, respectively.

The experimental results show that this paper can realize the animation generation of human motion gestures based on the 3D motion gesture data obtained by motion capture technology, combined with GAN and motion detail attention mechanism. Compared with the traditional image-based animation generation method, the common artifacts in the animation generation process are effectively eliminated, making the animation more realistic and detailed, and the animation generation based on motion capture technology has a high degree of realism. The generated human animation results are then input into Unity3D and 3Ds MAX software to render the character animation action sequences, which can be applied to virtual reality scenes. This means that the generated animation can be applied in the virtual reality environment to provide a virtual demonstration scene for the action sequence planning of film and television performances, so as to refine the expression of the action in the process of film and television performances, and better convey the connotation of the spirit of the character's action in film and television performances.

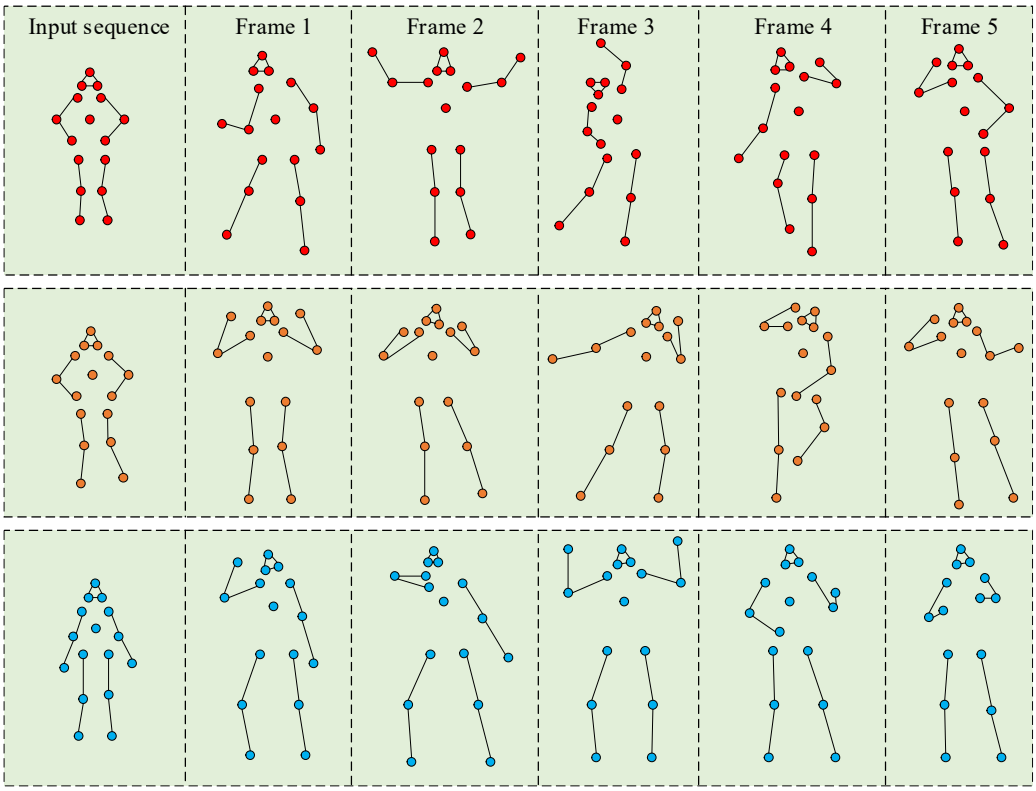

Fig. 4. Generate the sampling results of several frames in the animation

5. Conclusion

The action design of film and television performances can reflect the spiritual expression of the characters, so the optimization of action design has become an important means to enhance the infectious force of film and television performances. Based on this, this paper takes the action data obtained by motion capture technology as the basis, and combines the three-dimensional action gesture estimation model with multi-view fusion to obtain the action sequence. Then a character animation generation model based on GAN and attention mechanism is constructed by combining the action gesture sequence. Aiming at the effectiveness of the above model, a quantitative study is carried out through the dataset.

1) After the 3D action data acquired by motion capture technology is recognized by the multi-view fusion action pose estimation model, the mean value of its PCP3D index is 98.37, which is 0.28 percentage points higher than that of the TEMPO model, which has the second best performance. After adding multi-view fusion network and multi-feature fusion network, the average joint position error of the 3D action pose estimation is only 16.07 mm, which is the smallest, and a more accurate action sequence can be obtained by using motion capture technology to obtain the action data of characters in film and television performances and combining them with the 3D action pose estimation.

2) The diversity and richness index values of the animation generation model designed in this paper on the Human dataset are 5.104 and 3.997, respectively, and the accuracy rate of user labeling can reach 94.55% when PCK is equal to 0.35. Combined with Unity3D and 3Ds MAX software, the generated animation sequences can be placed into the virtual scene, and the animation generation results are closer to reality through rendering. It can provide decision-making assistance for action optimization design of film and television performances, and can also express the connotation of character actions in film and television performances more effectively.

Acknowledgements

1) This work was supported by Analysis of the impact of "Internet +" on the sports industry in the context of the new crown pneumonia epidemic - using Xi'an as an example.
2) This work was supported by Research on the Development Path of Aerobics Dance in Shaanxi Province from the Perspective of Healthy China.

References

[1] Chen, H. (2024). Advancements in Human Motion Capture Technology for Sports: A Comprehensive Review. Sensors & Materials, 36.

[2] Zhu, X., & Li, K. F. (2016, July). Real-time motion capture: An overview. In 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS) (pp. 522-525). IEEE.

[3] Hasler, N. (2021). Motion capture. In Computer Vision: A Reference Guide (pp. 818-822). Cham: Springer International Publishing.

[4] Zhu, Y. (2019, April). Application of Motion Capture Technology in 3D Animation Creation. In 3rd International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2019) (pp. 452-456). Atlantis Press.

[5] Zafer, A. Y. A. Z. (2023). Research on Current Sectoral Uses of Motion Capture (MoCap) Systems. Current Studies in Technology, Innovation and Entrepreneurship, 22.

[6] Shi, G., Wang, Y., & Li, S. (2014). Human motion capture system and its sensor analysis. Sensors & Transducers, 172(6), 206.

[7] Komisar, V., Novak, A. C., & Haycock, B. (2017). A novel method for synchronizing motion capture with other data sources for millisecond-level precision. Gait & Posture, 51, 125-131

[8] Ivanov, A. V., & Zhilenkov, A. A. (2018, January). Acquisition and processing data system based on technologies of inertial capture of the movement. In 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (pp. 882-885). IEEE.

[9] Patrona, F., Chatzitofis, A., Zarpalas, D., & Daras, P. (2018). Motion analysis: Action detection, recognition and evaluation based on motion capture data. Pattern Recognition, 76, 612-622.

[10] Sutopo, J., Ghani, M., & Burhanuddin, M. (2019). Synchronization of dance motion data acquisition using motion capture. International Journal of Innovative Technology and Exploring Engineering, 9(2), 1-4.

[11] Holden, D. (2018). Robust solving of optical motion capture data by denoising. ACM Transactions on Graphics (TOG), 37(4), 1-12.

[12] Benter, M., & Kuhlang, P. (2020). Analysing body motions using motion capture data. In Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2019 International Conference on Human Factors and Systems Interaction, July 24-28, 2019, Washington DC, USA 10 (pp. 128-140). Springer International Publishing

[13] Hegarini, E., Mutiara, A. B., Suhendra, A., Iqbal, M., & Wardijono, B. A. (2016, October). Similarity analysis of motion based on motion capture technology. In 2016 International Conference on Informatics and Computing (ICIC) (pp. 389-393). IEEE.

[14] Balazia, M., & Plataniotis, K. N. (2017). Human gait recognition from motion capture data in signature poses. IET Biometrics, 6(2), 129-137.

[15] Van der Kruk, E., & Reijne, M. M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. European journal of sport science, 18(6), 806-819.

[16] Petrosyan, T., Dunoyan, A., & Mkrtchyan, H. (2020). Application of motion capture systems in ergonomic analysis. Armenian journal of special education, 4(2), 107-117.

[17] Zeng, J., He, X., Hu, Y., Zhang, Y., Yang, H., & Zhou, S. (2021, May). Research status of data application based on optical motion capture technology. In 2021 2nd International Conference on Artificial Intelligence and Information Systems (pp. 1-8).

[18] Rybnikár, F., Kačerová, I., Hořejší, P., & Šimon, M. (2022). Ergonomics evaluation using motion capture technology—literature review. Applied Sciences, 13(1), 162.

[19] Sharma, S., Verma, S., Kumar, M., & Sharma, L. (2019, February). Use of motion capture in 3D animation: motion capture systems, challenges, and recent trends. In 2019 international conference on machine learning, big data, cloud and parallel computing (comitcon) (pp. 289-294). IEEE.

[20] Fu, Y., Li, Q., & Ma, D. (2023). User experience of a serious game for physical rehabilitation using wearable motion capture technology. IEEE Access.

[21] Guo, Y., & Zhong, C. (2022). Motion capture technology and its applications in film and television animation. Advances in Multimedia, 2022(1), 6392168.

[22] Liu, Z. (2024). Optimisation of digital media technology for film and television animation post-production considering motion capture technology. International Journal of Information and Communication Technology, 25(8), 1-13.

[23] Wendong, & Wang, C. (2021, December). Design and Realization of 3D Movie Animation Production Management System Based on Motion Capture Technology. In 2021 International Conference on Aviation Safety and Information Technology (pp. 631-635).

[24] Zhang, M. Y. (2013). Application of performance motion capture technology in film and television performance animation. Applied Mechanics and Materials, 347, 2781-2784.

[25] Wang, Y., Wang, Y., & Lang, X. (2021). Applied research on real-time film and television animation virtual shooting for multiplayer action capture technology based on optical positioning and inertial attitude sensing technology. Journal of electronic imaging, 30(3), 031207-031207.

[26] Zhu, R., Kim, H. G., & Moon, J. (2019). The Application and Improvement of Motion Capture Technology in Marvel Films. TECHART: Journal of Arts and Imaging Science, 6(4), 46-49.

[27] Zhang, S., & Xun, W. (2022). Group Animation Motion Capture Method Based on Virtual Reality Technology. Advances in Multimedia, 2022(1), 6107335.

[28] Jingde, N. (2020). Application analysis of motion capture technology in film and television animation production. The Theory and Practice of Innovation and Enntrepreneurship, 3(5), 141.

[29] Xiang er Li. (2024). Research on the Application and Development of Motion Capture Technology in Film and Television Animation Production. Science Social Development and Engineering Technology Research(2).

[30] Ali Nasr,Kevin Zhu & John McPhee. (2024). Using musculoskeletal models to generate physically-consistent data for 3D human pose, kinematic, dynamic, and muscle estimation. Multibody System Dynamics(prepublish),1-34.

[31] Liu Yamei,Guo Li & Guo Qiang. (2024). Retraction Note: Dynamic light collection system based on human posture estimation application in martial arts action teaching simulation. Optical and Quantum Electronics(10),1755-1755.

[32] Wenjun Wang,Chao Su & Guohui Han. (2024). Enhancement of Motion Blurred Crack Images Based on Conditional Generative Adversarial Network. Arabian Journal for Science and Engineering(prepublish),1-17.

[33] Xiaohui Li,Xinhai Han,Jingsong Yang,Jiuke Wang,Guoqi Han,Jun Ding... & Jun Yan. (2024). A Generative Adversarial Network with an Attention Spatiotemporal Mechanism for Tropical Cyclone Forecasts. Advances in Atmospheric Sciences(1),67-78.