International Journal of Information Technology and Web Engineering
Volume 19 • Issue 1

Multimedia Human-Computer Interaction Method in Video Animation Based on Artificial Intelligence Technology

Linran Sun, Xinyang Agriculture and Forestry University, Xinyang, China & Cheongju University, Cheongju, South Korea*
Nojun Kwak, Cheongju University, South Korea

ABSTRACT

With the development of computer technology innovation, be able to deal with the media comprehensive information and real-time information interaction with the computer multimedia technology arises at the historic moment, it promotes the application fields of computer widen to industrial all aspects of life. As the product of digital technology, animation technology plays an irreplaceable role in the production of multimedia courseware. However, the existing human-computer interaction methods have shortcomings such as incomplete extraction of video features and poor human-computer interaction effect. In this context, this paper designs a multimedia human-computer interaction method for animation works based on CNN model. First of all, the original video data is collected and preprocessed. Then it is input into the HCI framework based on CNN model for feature extraction. Finally, the effectiveness and practicability of the proposed method are proved by simulation experiments, which provides a reference and basis for the research of modern human-computer interaction.

KEYWORDS

Artificial Intelligence Technology, CNN, Human-Computer Interaction, Multimedia, Video Animation

With the popularization of computer technology and the advancement of teaching equipment and computer-aided teaching, multimedia technology has gradually replaced the conventional interactive methods of teaching (Dai & Yang, 2022). The percentage of the population of China that participates in multimedia has increased in the modern era (Ge & Darcy, 2022). A medium is sometimes referred to as media.

A medium is the means through which information is stored. *Medium* also refers to a form of expression or carrier of information. The corresponding term *multimedia* refers to the integrated application of sound, graphics, images, audio, video, animation, and other media by means of media technology. It can be explained from two aspects: on the one hand, it refers to the performance and communication forms of various information media; on the other hand, it refers to the methods and means by which people use computer technology to process multimedia information. Multimedia human-computer interaction (MHCI), defined from a certain level, can be considered as an interactive

DOI: 10.4018/IJITWE.344419 *Corresponding Author

method that uses modern teaching multimedia technology to participate in the whole process of learning according to the preset goal of the instructor. Computer-aided interaction is a new method of applying the functions of a computer to human-computer interaction (Alnuaim et al., 2022). It usually refers to the process by which learners interact with a computer as an auxiliary means to achieve corresponding goals. MHCI is gradually being applied in various fields (Chen, 2021). The aim is in applying MHCI are to realize. interactive learning processes, diversified learning resources, visualization of teaching methods, intelligent teaching processes, diversity and vividness of teaching content, and changes in teachers' role and students' status. To achieve these goals, we must make good use of multimedia interactive technology (Al-Hunaiyyan et al., 2021). So what are the applications of multimedia teaching in the specific field video animation? This question requires an examination of MHCI methods in video animation (Feng et al., 2021).

Since the mid- and late 1980s, human society has been increasingly shaped by information and intelligence. In this process, the development and application of multimedia technology has become an important topic of research, has become increasingly popular in people's lives, and has become an irreplaceable element of work. Multimedia technology arises from the intersection of computer technology and people's needs. Its emergence enables people to obtain information in their own accustomed way, shortens the time of information transmission, improves the quality of information acquisition, and makes rich and lively content appear on home computer. Not only can users get information in textual form, but they can also use audio and video to convey their feelings in the presentation of information. In the late 1990s, with the further development of computer technology and the popularization of multimedia technology, the new interactive media composed of computer, projection screen, video display table, and central control system became increasingly popular. Multimedia courseware integrates various technologies, such as sound, image, text, video, and animation. It makes the information field more colorful and improves the quality and efficiency of video analysis. Due to the rapid development of network information technology and the upgrading of related hardware, China has now entered the era of big data. Pictures, audio, text, and other multimedia information are widely spread on the internet every day, and the use of this technology shows a continuous explosive growth (Gurcan et al., 2021). These massive amounts of data and intricate communication forms undoubtedly bring great challenges to the tasks of personalized recommendation, statistical analysis, big data retrieval and other applications (Jiang et al., 2021). The question of how to classify and understand multimedia information more effectively and conveniently with less manual intervention is a widely discussed topic in the field of cross-modal research. With the popularity of short video software, video data has become an important way for people to acquire knowledge and understanding and has also become the main processing object and application direction in the current era of big data.

With the development of markets and technology, the interactive electronic whiteboard has appeared. The interaction here mainly refers to the information communication between computers. The electronic whiteboard is connected to the computer, and the content that needs to be displayed on the computer is displayed on the electronic whiteboard by means of transmission equipment. With the support of special application programs, an interactive teaching or conference environment can be formed to realize the information exchange and interaction between the speaker and the audience. The electronic whiteboard is operated by using a special pen. It can run all applications, edit files, annotate files, and perform all operations on the computer through the mouse and keyboard. It can be said that the emergence of the electronic whiteboard has constructed a large-screen, interactive, collaborative conference or teaching environment and a fully realized human-computer interaction, that is, information exchange with the computer (Fong et al., 2021). However, there are also some problems hindering the wide application of the electronic whiteboard; for instance, a large whiteboard screen presents difficulties associated with a large increase in input (Sulehu et al., 2022). At present, the cheapest electronic whiteboard on the market costs more than 5000 Yuan. In addition, this white

board is generally used as an input additional screen; portability is poor, and using the projection screen directly as the input screen of the white board requires additional devices, whose price is not cheap.

Video description not only has the value of promoting the integration of vision and text in theory, but also has very broad prospects for practical application. With the popularity of video software, everyone is receiving a large amount of video data every day. Analyzing these videos through video description can efficiently realize big data recommendation, video classification, and video screening. Secondly, in the field of intelligent monitoring, video description of the monitored area can realize some specific requirements. For example: when theft and other illegal behavior takes place in the monitoring area, it can trigger the alarm. In addition, for some people with visual impairment, human-computer interaction technology can be used to describe the field of vision in front of them and then generate language and pass it to people with visual impairment.

In recent years, a variety of new information display and interaction devices have appeared in the public eye. With the development of equipment and technology (O'Dwyer et al., 2021), new application demands for exhibition, education, and other industries are also emerging constantly (Kamariotou et al., 2021). People are no longer satisfied with playing the role of passive receivers of information, but hope to become participants in creation. The question of how to use equipment and technology to meet people's interactive needs is an important topic of research in the current industry. In this context, this paper focuses on the visual interaction of popular science exhibits. Visual tracking interaction technology is greatly influenced by the tracking algorithm; the accuracy, speed, and reliability of the tracking algorithm will affect the interaction effect. A delay in tracking interaction and the loss of target will make the interaction fail. Therefore, it is very necessary to study visual tracking technology to achieve natural real-time interaction. In the exhibition environment, target tracking is vulnerable to the interference of background images, so often algorithms are carried out under the background of a single moving target tracking, even tracking in complex background. Users have also asked for moving targets as well as more rigid objects, so research on visual tracking under complex background interaction research has become very important. It is also very promising to use visual tracking technology as an interactive form in popular science exhibits, and many scholars have carried out research in this area. In recent years, more and more researchers have devoted themselves to the study of visual tracking interaction with complex backgrounds and achieved remarkable results (Lou et al., 2022). In modern science and technology museums, a variety of new displays and interactions appear in public view, from the ordinary display equipment to a variety of handheld display equipment (Chauhan et al., 2021). There is display equipment, and even all kinds of physical display.

The research of this paper deals with the situation described above. In this paper, a deep learning-based, multi-feature, multi-mode video analysis method aims to automatically generate a human-readable natural language description for a given video clip. After watching a video, humans can easily describe the content of the video and describe it in detail. However, it is very difficult for a computer to accurately generate a description of the video. On the one hand, it has to deal with the presentation of the video. There are many people, objects, and scenes in the video, so it is a challenge for the computer to accurately and comprehensively express the video content(Dessi et al., 2019). On the other hand, once you have a video representation, it is a challenge for the computer to efficiently use that representation; there is the question of whether simple stitching can improve the quality of the description and whether the model can really understand the video content (Liang, 2019). Faced with a series of such problems, it can be seen that video description is a very complex task, involving two completely different modal mapping problems, which has a profound impact on promoting the research of MHCI in time-review animation (Wik & Pettersson, 2019).

This paper first introduces the application of multimedia technology in education, big data, and interactive technology and then reviews previous research in the field of convolutional neural network (CNN) modeling in animated videos. The experimental section first introduces the CNN

model, including the model's mechanism, training process, and prediction process. The Video Data Mining System Requirements Analysis section provides a detailed analysis and summary of the system requirements and functionality. The conclusion summarizes the whole paper, reviews the research results, and looks forward to possible future research directions and room for improvement.

RELATED WORK

With the development of a variety of interactive design platforms and software, micro-videos with interactive functions have emerged in an endless stream. The development of the internet has promoted the development of micro-video with interactive functions in the field of internet advertising, television, and education. At present, micro videos integrated with interaction design have been spread on the network. For example, interaction design is applied in interactive movies, where the audience can choose the beginning and end of the movie by themselves. In such a case video has to adapt to multiple interactions in a video broadcast platform; in addition, an online teaching platform also provides learners with a combination of interaction design technology in courses with a series of short videos. Mainly through human interaction, learners interact with content and other learners using the interface interaction design, in such concrete forms as barrage, embedded, hyperlinks, etc. (Tian & Tsai 2021). Scholars in this field, as shown in Table 1, have made progress, which can be divided into two aspects: interaction design development in micro-video and application research of interaction design in micro-video (Gao, 2018). Interaction design-related technologies developed earlier in foreign countries, with high maturity, and are applied mostly in human-computer interaction, APP design, digital media, and other commercial fields. The integration of interaction design in different fields aims to enhance users' positive experience and create easy-to-use interactive products to a certain extent. Interaction design was first used in advertising to enhance public participation in products and attract consumers, so as to achieve the ideal marketing purpose.

Table 1. Research Topics and Methods of Related Research

Author/Year	Research Topic	Method/Model
Wu et. al., 2015	Video classification	Hybrid deep learning framework with two CNNs for spatial and short-term motion features
Yao et. al., 2015	Video description generation	Approach considering both local and global temporal structure
Xiong et. al., 2017	Crowd counting	ConvLSTM model to exploit temporal information
He et. al., 2019	Video spatial-temporal modeling	Novel spatial-temporal network (StNet) architecture
Isobe et. al., 2020	Video super-resolution	Comparison of 2D CNN, 3D CNN, and RNN for temporal modeling
Fu et. al., 2021	Video language modeling	VIOLET network with Masked Visual-token Modeling pre-training task
Nir et. al., 2022	Semantic representation for cartoons/ animation videos	Method for refining semantic representation for specific animated content
Wang et. al., 2022	Video spatial-temporal modeling	Video Mobile-Former with 3D-CNNs and Transformer modules
Li et. al., 2022	Video matting	VMFormer, a Transformer-based end-to-end method
Zhao et. al., 2022	Human modeling and rendering	Comprehensive neural approach based on dense multi-view videos

METHOD

The Proposed CNN-Based Multimedia Human-Computer Interaction Method in Video Animation

Deep CNN Model

CNN is used to process gridded data with local association; this operation is shown in Figure 1. It is usually used to extract features of visual data. CNN has three important features. The first is local connection, that is, nodes of the convolution layer are connected only to some nodes of the upper layer, so that only local features of data are acquired. Second, weight sharing reduces the complexity of modeling and the parameters to be learned. Finally, there is pooled computation, which enables dimensionality reduction of data structures and a degree of translation invariance that makes the architecture impervious to small changes in position. In video description models based on deep learning, CNN is usually needed to capture rich semantic information in videos. Commonly used CNNs include AlexNet, VGGNet, ResNet, etc.

The core idea of CNN is to realize the mapping of inputs to outputs without explicitly determining the expression of the relationship between inputs and outputs. By means of local connectivity and weight sharing, CNN reduces the complexity of the network model and reduces the number of weights for better network optimization. A typical CNN consists of an input layer, a hidden layer, and an output layer, where the hidden layer includes a convolutional layer, a pooling layer, and a fully connected layer.

The convolutional layer is one of the most important parts of the CNN and is used to extract features from the input data to obtain a higher-level feature representation. Assuming that the lth layer is a convolutional layer, the output value of the lth layer can be expressed as shown in Equation 1.

$$y_l^j = \sigma\left(\sum\left(i=1\right)^{\left(k_l\right)}\left(x_{l-1}^i * W_l^{ij}\right) + b_l^j\right) \quad (1)$$

where y_l^j denotes the output value of the jth convolutional computation of the lth layer, σ denotes the activation function, x_{l-1}^i denotes the ith input value of the l-1th layer, k_l is the number of input eigenvectors, * denotes the element-by-element multiplication, W_l^{ij} denotes the convolutional kernel weights, and b_l^j denotes the jth bias vector of the lth layer.

In CNN-related research, model inputs and outputs are usually in the form of images. When the parameters of the deep learning model are fixed, the same input values produce the identical output values. If the original image is divided into two subset images that do not overlap each other, embedding information in one subset image will not affect the other subset image. The use of the image division method ensures that the input images at the sender and receiver sides are identical, as shown in Figure 2.

The calculation process of convolution is shown in Equation 2.

$$CONV_{(ij)} = \sum_{i}^{m-1}\sum_{j}^{n-1} u_{ij} \times w + b \ \ (i = 1,2\cdots m-1; j = 1,2\cdots n-1) \quad (2)$$

Figure 1. Schematic of a CNN

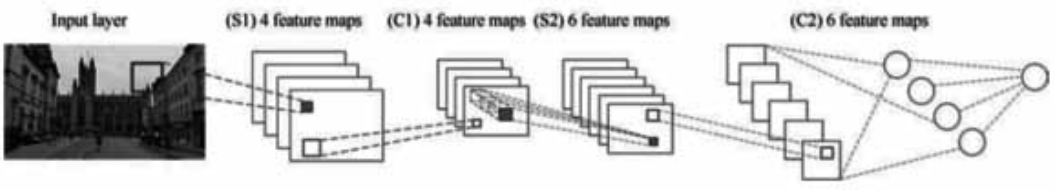
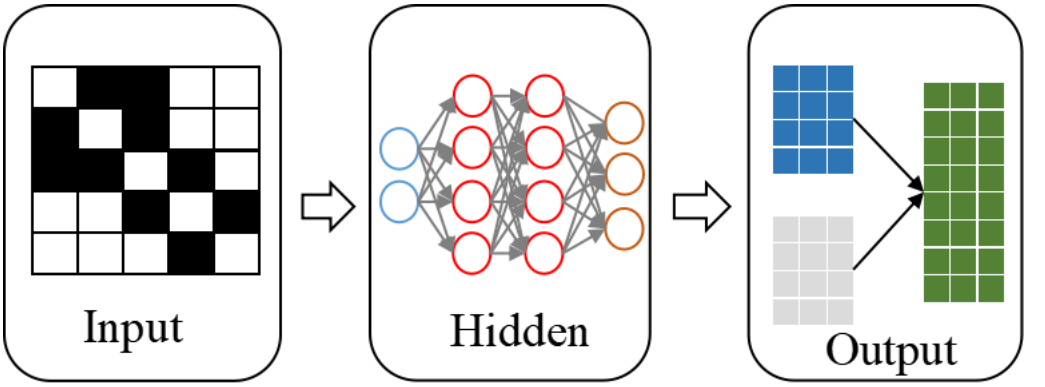


Figure 2. Typical Schematic Diagram of the CNN Model



where, u_{ij} is the input image, m and n are the size of the input image, w is the size of the convolution kernel, and b is the bias constant of the convolution kernel. $CONV(ij)$ is the characteristic graph output after convolution operation.

CNN adds an activation function layer to the network and analyzes the model better by adopting the feature mapping method of nonlinear function. Then, the activation functions are introduced one by one as shown in Equations 3–5.

The sigmoid function is shown in Equation 3.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The tanh function is shown in Equation 4.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4}$$

The ReLu function is shown in Equation 5.

$$f(x) = \max(0, x) \tag{5}$$

The full name of the ReLU function is *rectified linear unit*. The function is one of the commonly used activation functions, which is characterized by low computational complexity and no exponential operation. However, it is worth explaining that the ReLU function has certain defects in the calculation process. When the data passes through the negative range of the ReLU function, the output value is equal to 0. The Leaky-ReLu function can solve the above problem, as shown in Equation 6.

$$f(x) = \begin{cases} x, x \geq 0 \\ \alpha x, x < 0 \end{cases} \tag{6}$$

The corresponding equations of Sig and Tanh are shown in Equations 7 and 8.

$$\begin{cases} sig(x) = \dfrac{1}{1+\exp(-x)} \\ \tanh(x) = \dfrac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} \end{cases} \tag{7}$$

where x is the input.

$$h_{w,b}\left(x_i\right) = \begin{bmatrix} p\left(y_i = 1 \middle| x_i; w, b\right) \\ p\left(y_i = 2 \middle| x_i; w, b\right) \\ p\left(y_i = 3 \middle| x_i; w, b\right) \\ \cdots \\ p\left(y_i = n \middle| x_i; w, b\right) \end{bmatrix} = \frac{1}{\sum_{j=1}^{n} e^{w_j x_i + b_j}} \begin{bmatrix} e^{w_1 x_i + b_1} \\ e^{w_2 x_i + b_2} \\ e^{w_3 x_i + b_3} \\ \cdots \\ e^{w_n x_i + b_n} \end{bmatrix} \tag{8}$$

The *w* is the weights, and *b* is the bias. The cross entropy (CE) formula is shown in Equation 9.

$$\text{loss} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ji} \log\left(\hat{y}_{ji}\right) \tag{9}$$

The original form of the gradient descent method is shown in Equation 10.

$$\theta := \theta - \alpha \frac{\partial}{\partial \theta} J(\theta) \tag{10}$$

where θ is the parameter. The Adam optimizer is given as shown in Equations 11 and 12.

$$m_t = \beta_1 m_{t-1} + \left(1-\beta_1\right) g_t \tag{11}$$

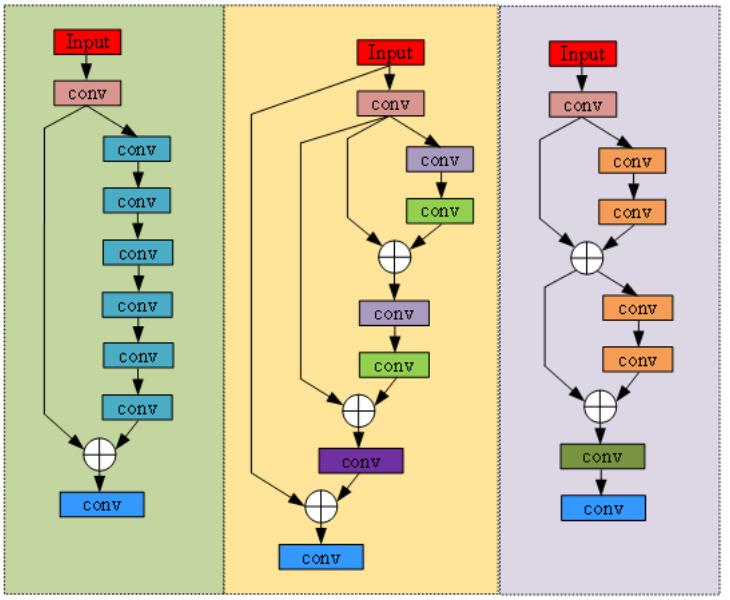
$$v_t = \beta_2 v_{t-1} + \left(1-\beta_2\right) g_{\mathrm{t}}^2 \tag{12}$$

where g_t is the object function and m_t and v_t are the parameters. β_1 and β_2 are the regularization parameter. The wrong actions identification accuracy can be measured *RMSE* index as shown in Equation 13.

$$\text{RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(y_i - \hat{y}_i\right)^2} \tag{13}$$

As shown in Figure 3, the normal cell is on the left, the simplified cell is in the middle, and the final network structure is on the right. Cells can also be stacked in more complex ways, such as neural networks with multiple layers of structure. Since the advent of cell-based structures, cell-based search

Figure 3. Cell-Based Structure Description



spaces have been successfully used in many recent works. In theory, cells can be combined in any way, such as the previously mentioned multibranched structure, simply replacing layers with cells. Ideally, the network structure and the cell structure should be optimized together, not just the micro network structure. Otherwise, the CNN model cannot be built. People also need to manually design the macro structure, which increases the design and operation burden of the model. In addition, some researchers also use network morphology, which is obtained by continuous updating on the basis of the existing network structure, rather than by designing a new neural network from scratch. Therefore, the aim of this study is to embed a CNN model into MHCI to realize the analysis of video animation.

The Framework of the Proposed Method

Based on the above discussions, the proposed CNN model and its application to an MHCI method for video animation is shown in Figure 4.

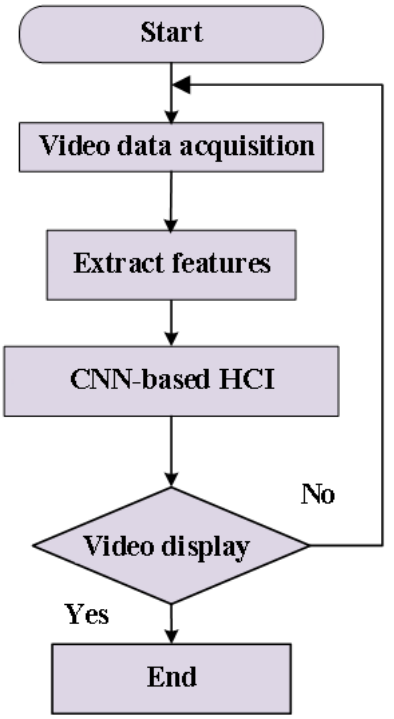
EXPERIMENTAL RESULTS AND ANALYSIS

Data Collection and Experimental Environment

So far, there are two publicly accessible dense video description datasets, namely ActivityNet and YouCook2. The ActivityNet dataset, which has been widely used in intensive video description tasks, consists of 100,000 event description sentences from 20,000 YouTube videos covering a wide range of complex human activities. In this study, we chose ActivityNet as the main experimental dataset to verify the effectiveness of the proposed method. The YouCook2 dataset contains more than 2,000 long, unedited videos from 89 cooking recipes. The average duration of each video was 5.26 minutes and consisted of about 7 events.

According to the official division, we allocated 1,000/491/504 videos for training, validation, and testing, respectively. Each video had an average duration of 5 minutes and contained 7.65 statements describing the event, with each description sentence being approximately 17.65 words. Like everyone

Figure 4. Model Structure Diagram of the Proposed Method



else, we used the validation set to report all our results because the test set did not provide the correct statement annotations.

The following steps were taken in preprocessing:

First, a subset from the ActivityNet dataset was selected for experimentation that was suitable for the objectives of the study. Then, the provided event description sentences were used to divide each video into multiple segments, and text processing techniques such as word splitting, stop word removal, and word stemming were applied to the text description of each segment.

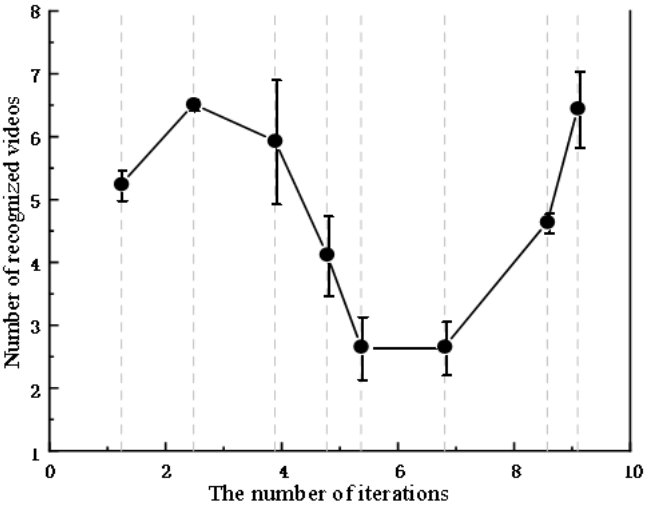
Next, the textual descriptions were converted into vector representations using word embedding techniques (e.g., Word2Vec or GloVe) and used as inputs to the CNN model. At the same time, the video data was sampled or the motion information was extracted using the optical flow method and was converted into a suitable format and size. Data enhancement techniques such as random cropping, flipping, rotating, etc. were also used in order to ensure the diversity and quantity of training data.

Through these preprocessing steps, the dense video descriptive data in the ActivityNet dataset was successfully prepared, and the consistency, availability and appropriateness of the data was ensured.

Experimental Results Analysis

First, Figure 5 shows the relationship between the number of iterations of the model and the number of recognized videos. As can be seen from the figure, with the increase of the number of iterations, the number of recognized videos shows an increasing trend. However, when the number of iterations continues to increase, the number of identified videos tends to decrease, which may be mainly due to the phenomenon of overfitting. When the number of iterations further increases, the number of recognized videos continues to increase, which indicates the effectiveness and practicability of the proposed method.

Figure 5. The Relationship Between the Number of Iterations of the Model and the Number of Recognized Videos



There is a certain relationship between the number of iterations and the number of recognized videos. When the number of iterations is 3 and less, the number of recognized videos increases as the number of iterations increases; when the number of iterations is between 3 and 7, the number of recognized videos decreases gradually as the number of iterations increases; when the number of iterations is between 8 and 10, the number of recognized videos increases again as the number of iterations increases.

Figure 6 shows the amplitude distribution of each frequency pixel in the video, with the three-dimensional coordinates representing the time, frequency, and amplitude respectively. As can be seen from the figure, videos with different colors, amplitudes, and frequencies can be uniformly distributed in the whole coordinate space. This shows that the method in this study is well suited to extract the key features of video animation works and to provide good data features for the subsequent human-computer interaction platform. It shows that this method has a better human-computer interaction function.

Furthermore, Figure 7 shows the comparison of video image sharpness before and after feature extraction. The image at the far left is the original video picture; the middle image is the feature clustering result after CNN extraction; and the image at the far right is the clustering result after feature extraction of the conventional neural network model. As can be seen from the figure, the clustering of features extracted by CNN can restore the sharpness of the original image to the greatest extent, while the sharpness of features extracted by the conventional model is significantly less. Therefore, it is evident that the proposed method can effectively extract the key features in video animation and achieve better clustering effect.

To provide even more evidence for the usefulness of CNN, the recognition rate was compared with some commonly used traditional recognition methods, and the results are shown in the Table 2.

In Table 2 it can be seen that the recognition rate of the CNN algorithm is 92%. It is better than multilayer perceptual network and Hidden Markov Model.

In addition, Figure 8 shows the satisfaction rate of the proposed human-computer interaction method. As can be seen from the figure, the method proposed in this paper has achieved relatively satisfactory results among most people: 30.3% of people were very satisfied, and the proportion of satisfied people was 21%. The proportions of people who were more satisfied and who found the results acceptable were 10.7% and 17.7%, respectively. The non-feeling group and the dissatisfied

Figure 6. Amplitude Distribution of Each Frequency Pixel in the Video

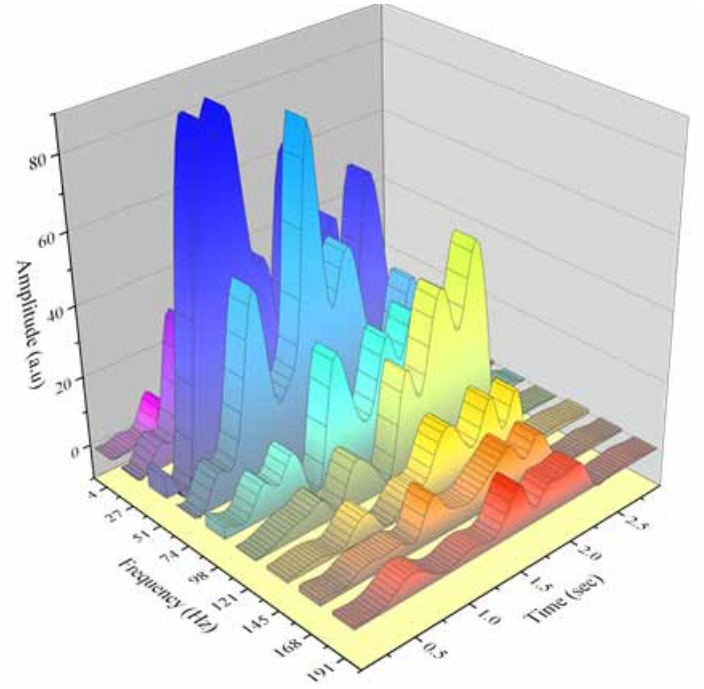
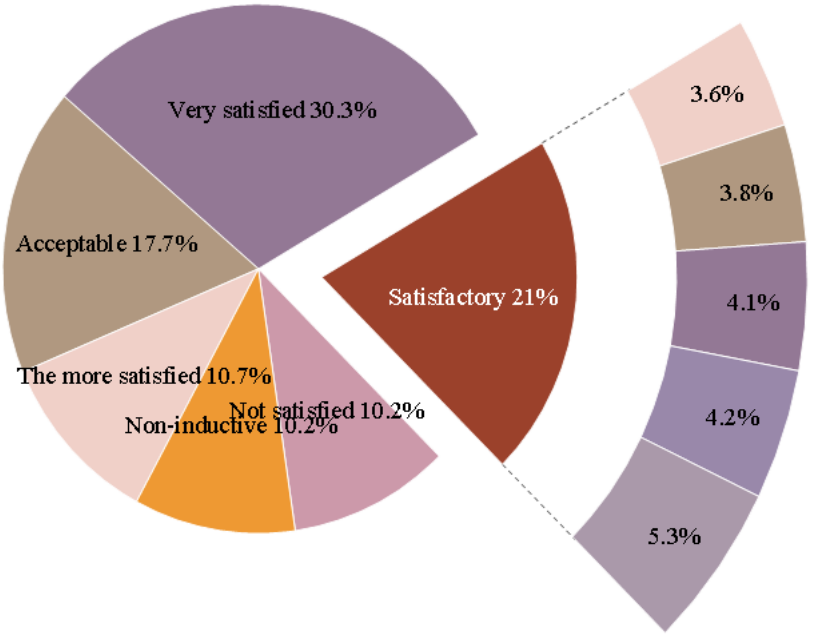


Figure 7. Comparison of Video Image Sharpness Before and After Feature Extraction

Table 2. Recognition Rate of Three Algorithms

Algorithm	Recognition Rate (%)
Multilayer Perceptron	60.4
Hidden Markov Model	87
CNN	92

Figure 8. Satisfaction Rates of the Proposed Human-Computer Interaction Method



group were each 10.2%. Thus, it is evident that the method in this paper has a good application satisfaction rate, demonstrating its reliability and practicability.

CONCLUSION

Video description is a very challenging cross-modal task with a wide range of application scenarios, such as video quizzing, video retrieval, and video obstacle assistance. Facing rapidly increasing quantities of surveillance videos, it is difficult to extract valuable laws or knowledge from the recorded phenomena. This project takes video animation as the object in studying the video data mining problem and finally designs a video data mining method for human-computer interaction.

This study makes three main contributions: The characteristics of video animation and its data mining are analyzed, and a video analysis method for human-computer interaction based on a deep learning model is proposed. Taking different video animations as research objects, the methods and processes of video animation human-computer interaction mining are designed respectively according to their working characteristics. On this basis, the requirements of the video data mining system are analyzed, and some functions are analyzed and summarized.

On the basis of our research and analysis of the video data mining problem, this project proposes a deep learning model-based video analysis method for human-computer interaction. By analyzing video animations and their data mining characteristics, we design the HCI mining methods and processes for different video animations.

The results show that CNN exhibits the highest recognition rate, 92%, in the video animation recognition task. This indicates that CNN has excellent performance in processing image and video data. The Hidden Markov Model also performs well, with a recognition rate of 87%. The Multilayer Perceptron (MP) algorithm has a relatively low recognition rate of 60.4%.

The results of this research provide an effective human-computer interaction method for video data mining. However, the method proposed in this paper also has some limitations. First, the CNN

model requires a large amount of training data and computational resources, which may not be feasible in some scenarios. Second, the performance of event recognition heavily depends on the quality of video animation and the diversity of events, which may limit the generalizability of our proposed method. Third, the human-computer interface may need to be further optimized and customized to suit different user groups and application scenarios.

DATA AVAILABILITY

The figures used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

The authors would like to express thanks for the financial support they received from the Xinyang Agricultural and Forestry University teaching and research project "Xinyang Agricultural and Forestry College First-class undergraduate course *Urban Planning and Design 2"* (Grant No: 902539).

PROCESS DATES

Received: October 31, 2023, Revision: March 14, 2024, Accepted: March 14, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Linran Sun, sunny838868@163.com

ACKNOWLEDGMENT

The authors sincerely thank all those who contributed to this study.

REFERENCES

Al-Hunaiyyan, A., Alhajri, R., Al-Sharhan, S., & Bimba, A. (2021). Human-computer interaction perspective on mobile learning: Gender and social implications. [IJIM]. *International Journal of Interactive Mobile Technologies*, *15*(11), 4–20. doi:10.3991/ijim.v15i11.21367

Alnuaim, A. A., Zakariah, M., Alhadlaq, A., Shashidhar, C., Hatamleh, W. A., Tarazi, H., Shukla, P. K., & Ratna, R. (2022). Human-computer interaction with detection of speaker emotions using convolution neural networks. *Computational Intelligence and Neuroscience*, *7463091*, 1–16. Advance online publication. doi:10.1155/2022/7463091 PMID:35401731

Chauhan, S., Srivastava, S., Kumar, P., Patel, R. T. M., & Dhillon, P. (2021). Interaction of substance use with physical activity and its effect on depressive symptoms among adolescents. *Journal of Substance Use*, *26*(5), 524–530. doi:10.1080/14659891.2020.1851411

Chen, S. C. (2021). Multimedia in virtual reality and augmented reality. *IEEE MultiMedia*, *28*(2), 5–7. doi:10.1109/MMUL.2021.3086275

Dai, Z., & Yang, J. (2022). A multimedia learning for Chinese character image recognition via human-computer interaction network. *Advances in Multimedia*, *56*(1), 3–17. doi:10.1155/2022/4427091

Dessì, D., Fenu, G., Marras, M., & Reforgiato Recupero, D. (2019). Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections. *Computers in Human Behavior*, *92*, 468–477. doi:10.1016/j.chb.2018.03.004

Feng, Z., Wu, J., & Ni, T. (2021). Research and application of multifeature gesture recognition in human-computer interaction based on virtual reality technology. *Wireless Communications and Mobile Computing, 2021*, Article 3603693. (Retraction published 2023, *Wireless Communications and Mobile Computing, 2023*, Article 9842785)10.1155/2021/3603693

Fong, F. T. K., Imuta, K., Redshaw, J., & Nielsen, M. (2021). The digital social partner: Preschool children display stronger imitative tendency in screen-based than live learning. *Human Behavior and Emerging Technologies*, *3*(4), 585–594. doi:10.1002/hbe2.280

Gao, N. (2018). Construction and implementation of teaching mode for digital mapping based on interactive micro-course technology. [iJET]. *International Journal of Emerging Technologies in Learning*, *13*(2), 21–32. doi:10.3991/ijet.v13i02.8317

Ge, T., & Darcy, O. (2022). Study on the design of interactive distance multimedia teaching system based on VR technology. *International Journal of Continuing Engineering Education and Lifelong Learning*, *32*(1), 65–77. doi:10.1504/IJCEELL.2022.121221

Gurcan, F., Cagiltay, N. E., & Cagiltay, K. (2021). Mapping human-computer interaction research themes and trends from its existence to today: A topic modeling-based review of past 60 years. *International Journal of Human-Computer Interaction*, *37*(3), 267–280. doi:10.1080/10447318.2020.1819668

He, D., Zhou, Z., Gan, C., Li, F., Liu, X., Li, Y., Wang, L., & Wen, S. (2019, July). Stnet: Local and global spatial-temporal modeling for action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 8401–8408. doi:10.1609/aaai.v33i01.33018401

Jiang, S., Wang, L., & Dong, Y. (2021). Application of virtual reality human-computer interaction technology based on the sensor in English teaching. *Journal of Sensors*, *2021*, 2505119. Advance online publication. doi:10.1155/2021/2505119

Kamariotou, V., Kamariotou, M., & Kitsios, F. (2021). Strategic planning for virtual exhibitions and visitors' experience: A multidisciplinary approach for museums in the digital age. *Digital Applications in Archaeology and Cultural Heritage*, *21*, e00183. doi:10.1016/j.daach.2021.e00183

Li, J., Goel, V., Ohanyan, M., Navasardyan, S., Wei, Y., & Shi, H. (2022. Vmformer: End-to-end video matting with transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6678–6687).

Liang, W. (2019). Scene art design based on human-computer interaction and multimedia information system: An interactive perspective. *Multimedia Tools and Applications*, *78*(4), 4767–4785. doi:10.1007/s11042-018-7070-6

Lou, X., Chen, Z., Hansen, P., & Peng, R. (2022). Asymmetric Free-hand interaction on a large display and inspirations for designing natural user interfaces. *Symmetry*, *14*(5), 928. doi:10.3390/sym14050928

Nir, O., Rapoport, G., & Shamir, A. (2022, May). CAST: Character labeling in Animation using Self-supervision by Tracking. *Computer Graphics Forum, 41*(2), 135–145). 10.1111/cgf.14464

O'Dwyer, N., Zerman, E., Young, G. W., Smolic, A., Dunne, S., & Shenton, H. (2021). Volumetric video in augmented reality applications for museological narratives: A User study for the long room in the library of Trinity College Dublin. *Journal on Computing and Cultural Heritage*, *14*(2), 1–20. doi:10.1145/3425400

Sulehu, M., Rimalia, W., & Tamra, T. (2022). Aplikasi Virtual Whiteboard sebagai Media Pembelajaran Daring. *Journal Software. Hardware and Information Technology*, *2*(2), 1–9. doi:10.24252/shift.v2i2.28

Tian, N., & Tsai, S. B. (2021). An empirical study on interactive flipped classroom model based on digital micro-video course by big data analysis and models. *Mathematical Problems in Engineering*, *16*(10), 2–19. doi:10.1155/2021/8789355

Wik, M., & Pettersson, J. S. (2019). Lack of multimedia tools in intervention support for running systems. *International Journal of Web Science*, *3*(2), 148. doi:10.1504/IJWS.2019.102217

Wu, Z., Wang, X., Jiang, Y. G., Ye, H., & Xue, X. (2015, October). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 461–470). doi:10.1145/2733373.2806222

Xiong, F., Shi, X., & Yeung, D. Y. (2017). Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 5151–5159). Association for Computing Machinery. doi:10.1109/ICCV.2017.551

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision* (pp. 4507–4515). IEEE.

Zhao, F., Jiang, Y., Yao, K., Zhang, J., Wang, L., Dai, H., Zhong, Y., Zhang, Y., Wu, M., Xu, L., & Yu, J. (2022). Human performance modeling and rendering via neural animated mesh. [TOG]. *ACM Transactions on Graphics*, *41*(6), 1–17. doi:10.1145/3550454.3555451

Linran Sun was born in Henan, China,in 1989. From 2007 to 2012, she studied in Beihai College of Beihang University and received her bachelor's degree in 2012. From 2012 to 2014, she studied in Guilin university of technology and received her Master's degree in 2014. From 2014, she works in Xinyang Agriculture and Forestry University. Currently, She is attending Ph.D. program at Cheongju University in Korea. She has published a total of 6 papers. Her research interests are included virtual interactive architectural design method and spatial cognition.

Nojun Kwak was born in Seoul, Korea, in 1969. From 1989 to 1996, he studied in Sungkyunkwan University and received bachelor's degree in 1996.From 2001 to 2003, he studied in School of Visual Arts and received Master's degree in 2003.From 2009, he works in Cheongju University.