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

Research on the Construction of Music Animation CAD System Based on Music Form and Emotion Recognition

Dan Shen[1] and Wenjia Zhao[2]

[1]School of Music and Dance, Harbin University, Harbin, Heilongjiang 150086, China, shendan@hrbu.edu.cn
[2]School of Music and Dance, Harbin University, Harbin, Heilongjiang 150086, China, wenjia0207@163.com

Corresponding author: Dan Shen, shendan@hrbu.edu.cn

Abstract. Music, as a unique artistic form of people's emotional expression, focuses on a rendering of the soul of the music audience and a re-experience of the virtual reality world. This experience is all-round, not just limited to hearing. In the process of music visualization, it is difficult to mechanically transform it into vision by a single rule. How to enhance the musical form and emotional characteristics of visual music works through multi-channel mapping mode is the research purpose of this paper. This article proposes a music based emotion recognition model feature, and a musical animation CAD (Computer Aided Design) system is constructed. The system extracts some basic musical features from MIDI (Musical Instrument Digital Interface) files, and then extracts the musical features of the music. Thus, each music segment is sliced into emotional music visualization programs. And by summarizing the obtained segment features as a whole, design an emotional feature program that can reflect the musical form. Through deep learning algorithms, it visualizes and matches the improved segment nodes, improving the expressive form of music emotional animation.

Keywords: Deep Learning; Musical Features; Emotional Recognition; Computer Aided Design; Music Animation; Visualization
DOI: https://doi.org/10.14733/cadaps.2024.S1.204-217

1 INTRODUCTION

Music is favored by people because of its rich and varied characteristics. The melody and rhythm of music can arouse the emotional resonance of the audience. With the advent of the digital age, the ways people experience and consume music are also enriched. Music emotion recognition is an emerging emotional language expression technology. Agarwal and Om [1] developed a framework system for music emotion recognition using the optimized model recognition of automatic encoders. It considers the music emotion classification and recognition dataset related to text. A comparison was made between existing emotion classifiers. When people enjoy music, there are generally two types of audio: natural audio, such as MP3 format and wav format; Structured audio, such as MIDI format. Because music is a special kind of sound information, how to present this information in the form of visual graphics and images to accurately deduce the unique characteristics of music is the key to the research. Chaturvedi et al. [2] attempted to construct a connection between emotional expression and music sensors. Signal acquisition for perceptual emotions was constructed through CAD assisted model training and result visualization. Due to the inability to directly capture the emotional features triggered by the music we listen to, the focus is on the application of sensors for emotion recognition. By combining different detection techniques, the accuracy of music signal parameters was determined and a music model with emotional features was constructed. About natural audio feature extraction, Fourier transform, windowing and framing are generally used for preprocessing; The extraction of structured audio features requires high correlation, so it is necessary to understand, analyze and translate the rules and attributes of audio features, and finally get the corresponding feature quantity. Music can effectively help people relieve emotions and relieve stress. With the continuous development of computers, online music datasets are constantly increasing, and the automation technology of music sentiment databases is constantly improving. Its assistance in automating music emotion classification has become a hot research field [3]. DL is an important branch of machine learning. Its original design is to try to make the machine have the ability to learn and analyze problems like human beings, and to simulate the brain's ability to collect and analyze information by constructing NN (Neural Network). A large number of scholars' research and experiments have proved that this idea is feasible and efficient. In many cases where there are interference feature vectors or miscellaneous feature vectors, DL has been proved to generate good feature sequences. Such as image recognition, natural language processing, etc. Therefore, DL can be applied to music form and emotion recognition.

Chen et al. [4] uses the original algorithmic dataset to incorporate the data that maximizes emotions into the optimal subspace. And decompose the dataset into time-frequency features using relevant algorithms, and use linear discrimination for sentiment recognition and classification. Experiments have shown that under the processing of spatial filtering, it effectively reduces the fluctuation of subjects' daily emotions and improves the accuracy and stability of emotions. The information retrieval of music emotion recognition has always been a challenging issue. Dong et al. [5] analyzed the significant impact, including music sequence information, on the two-dimensional time-frequency acoustic spectrum adaptive features of music emotional signals. This method can effectively retrieve music features of different emotional expressions and perform feature extraction, thereby achieving continuous emotional prediction and reducing computational complexity. Music visualization makes music not only limited to ears, but also can be seen, which strengthens people's perceptual experience and combines it with the rational use of media technology at a higher level. With the maturity of computer graphics technology, music animation, as a brand-new art form combining vision and hearing, has attracted more and more attention. Visual-centered media takes music as the carrier, and with the help of computer information and multimedia technology, music content is interpreted through pictures and images, which can finally achieve the popular communication of audio-visual combination. It can provide more intuitive visual presentation for the vast number of music lovers and listeners.

Ehrlich et al. [6] proposed a feedback measure for participants' emotion induced music regulation. It constructed a music algorithm recognition model for subjective emotions. Through functional simulations of emotional rationality and intelligent behavior, the closed-loop relationship between participants' emotions and musical feelings was analyzed. In terms of subjective emotional expression, the system constructed can not only capture the real-time emotional state of users, but also regulate the mood of music users. Er and Aydlek [7] proposed a new emotion recognition method for music recording. And extracts chromatograms from pre trained network models. A deep learning model was used to extract deep visual features for music emotion recognition for training and deep testing. In addition, in the design of most music animations, the user inputs the action parameters corresponding to the note characteristics through GUI (Graphical User Interface), thus controlling the movement of the three-dimensional animation. Because the production of this kind of music animation is only based on some isolated note characteristics, the music animation made is often not very satisfactory. Computer technology makes music visible, extracts the characteristics of music, and makes the visual effect appear in various forms through the research of image processing, virtual technology, digital and multimedia. At the level of multi-audio feature extraction, it is necessary to collect the digital signals of music and sort them out. The emotional integration of music animation in music form not only preserves the overall structure of music, but also appropriately expresses the connotation of music. Based on the in-depth discussion of DL algorithm, this paper puts forward a musical form and emotion recognition model based on DL algorithm, and builds a musical animation CAD system on this basis. The system extracts multiple features of music, and designs these features comprehensively and visually, which makes the image express more music information.

2 RELATED WORK

Ge et al. [8] conducted extensive feature analysis tests on the algorithm and simulated the management of music classification. Gómez et al. [9] developed a computer-aided algorithm for adaptive dialogue emotion. A model state that can automatically generate movie dialogue emotions was constructed by sharing the inner feelings of emotional music. This algorithm represents the emotional correlation feature value results of automatic music recognition. The development of computing systems can help scholars better process musical psychological signals. Promoting music recognition of emotions to a wider application space has become a research direction for researchers. Through CAD assisted automatic composition and music visualization, better application analysis of music emotion dataset models can be achieved [10]. Hizlisoy et al. [11] analyzed the performance of emotional feature recognition in deep learning networks. Jing and Song [12] used computer automatic image continuous generation technology to simulate the CAD animation parameters of character models in the scene. It conducts an overall analysis and design of the feature differentiation of animation tasks by shaping the overall image of designated tasks. Research has summarized effective methods for constructing new 3D animated characters. López et al. [13] conducted information exchange on emotional states through emotional computer analysis. It uses a brain computer interface between humans for personalized emotional expression needs processing, and constructs a research framework that affects user emotions through emotion recognition.

The emotion mining of emotional design is currently in the development stage, and relying on CAD's user emotion self-mining technology has brought great help to researchers. Mazhar et al. [14] conducted a decision evaluation of algorithms based on machine learning, which analyzed the support of different computer machine learning algorithms on a dataset. By classifying emotion analysis, we have achieved the analysis results of algorithm image recognition tasks with symmetrical characteristics, significantly improving the efficiency of computer-aided emotion recognition. Orjesek et al. [15] conducted information retrieval for detecting changes in music emotional content. By analyzing the emotional characteristics of audio waveform music using deep neural networks, a convolutional layer based automatic encoding bidirectional control unit is proposed. This unit can effectively awaken emotional expression. Compared with other models, the proposed iterative layer can effectively enhance the accuracy of feature discrimination. Pandeya et al. [16] analyzed the emotional communication chain of a large amount of emotional acoustic background information. It proposes an audio video exchange enhancement method that includes computer-aided design. This method is based on CAD sentiment analysis operations and utilizes a large amount of visual acoustic information for separable convolution strategy investigation. The results show that incorporating individual information into the overall emotional information flow can well reduce the analysis cost of neural networks. It can guide the direction of audio-visual emotion search in both single network mode and multi network mode. The feature extraction of human-computer interaction through computers is one of the important links in emotion recognition. Saha et al. [17] conducted a computer-assisted supervised search algorithm. By conducting cosine similarity combination search on relevant subsets, feature associations with lower correlation values and higher facial emotions were identified. The results indicate that its dataset evaluation has effectively optimized the feature vectors and significantly reduced similarity interference.

Music is an advanced form of enhancing emotional expression. Visnu et al. [18] constructed an efficient music recommendation system analysis framework using digital CAD technology. It uses facial technology recognition to improve the accuracy of the algorithm's musical expression of facial emotions. The system can recommend effective songs by recognizing emotions, which saves time and costs. Effective emotional recognition of individual differences can be achieved through computer-aided analysis. Music recommendation and music information retrieval can effectively apply individual emotional models. Xu et al. [19] conducted a machine learning based analysis of the emotional association between individual users and music. It uses personal characteristics as input and compares the degree of impact of model recognition through prediction and validation of perceived emotions. With the continuous development of CAD technology, the support optimization vector machine model for emotion classification using real music emotion datasets can effectively perform emotion optimization recognition. Yang and Li [20] have constructed a music emotion optimization recognition based on vector projection space. A set analysis was conducted on the practical reliability of the music emotion dataset. Combined with previous literature, this paper puts forward a musical form and emotion recognition model based on DL algorithm, and builds a musical animation CAD system on this basis.

3 MUSIC FORM AND EMOTION RECOGNITION BASED ON DL

At present, it is widely used in the field of target recognition and feature extraction, especially in computer vision, image processing and natural language processing. Deep learning technology can extract low-level features and abstraction high-level concepts in the application process. The main challenge in music emotion recognition is that the energy features are difficult to correlate with the deep information in the music, making it difficult to build a "bridge" between the two. This leads to a bottleneck in the performance of emotion recognition, and deep learning can help establish a connection between the two. The combination of music and Qin Gang is difficult to describe in detail with parameters. In order to distinguish the differences between emotional semantics and audio signals, this article conducted differential fusion through deep learning and addressed the challenges posed by different audio signals. On the one hand, DL technology can use massive data sets, and its performance will continue to improve with the increase of data scale. On the other hand, DL technology does not need to manually extract features for different problems, but is similar to black box operation, which can directly obtain features by training massive data sets, and obtain the inherent laws and representation ability in these data. Although DL has made some achievements in SER, the traditional SER is not without any value. The shallow learning model has fast training speed and few parameters, and the extracted features are targeted, while the DL network structure is complex, requiring a lot of training data, and the parameter adjustment is complex. DL does not need manual participation and thus lacks the guidance of prior knowledge. Therefore, shallow learning can be used to guide DL. In music feature extraction, there are short-term average zero-crossing rate, short-term average energy, short-term autocorrelation function, spectrogram, spectral centroid, frequency domain bandwidth and so on. For these digital signals that cannot be read directly from audio signals, preprocessing is needed. The whole process is "framing-pre-emphasis-windowing-silent frame discrimination-ending".

The recognition of music emotions is mainly divided into discrete emotion recognition and continuous emotion recognition. The method of discrete emotion recognition is to attach an emotional label to the entire piece of music. The continuous emotion recognition method involves cutting the entire piece of music into several segments and creating an emotion label value for each segment. And generate corresponding regression values based on the dimensional space model, so the selection of methods determines the process and results of the experiment. Continuous emotion recognition mainly relies on a universal continuous emotion space model. Its advantage lies in its ability to distinguish the subtle differences between different emotions in music and people's subjective feelings towards it. It can also be used to express emotions and experience people's lives. At present, the most widely used feature parameter is spectral correlation feature, and the commonly used spectral correlation features include linear prediction coefficient, Mel cepstrum coefficient, spectrogram and so on. When you have a certain length of music signal, you can draw its corresponding spectrogram. Although the structure of spectrogram is simple, it can express a lot of information.

Because music signals are real and unstable, the quality of feature parameters extracted from music signals will directly affect the results of SER. The spectrogram itself contains all the spectral information of the music signal, without any processing, so the information about music in the spectrogram is lossless. This is also the reason why this paper chooses spectrogram as the feature input of music signal. The gray value set at point (a,b) is:

$$g_i(a,b) = \log_{10}\left|X_i(m,n)\right| \tag{1}$$

$$G_i(a,b) = \frac{g_i(a,b) - g_{\min}(a,b)}{g_{\max}(a,b) - g_{\min}(a,b)} \tag{2}$$

Where: $g_{\max}(a,b)$ and $g_{\min}(a,b)$ are the maximum and minimum values in $G_i(a,b)$ gray scale.

Signal analysis in frequency domain can only explain frequency characteristics. When we want to observe frequency domain information in time stream, we need to use this spectrogram analyzed by short-term signals near each time point. The generation of spectrogram requires framing and windowing, short-time scale conversion to decibel representation of amplitude, and then splicing the processed frequency domain information according to time sequence to form spectrogram, as shown in Figure 1.

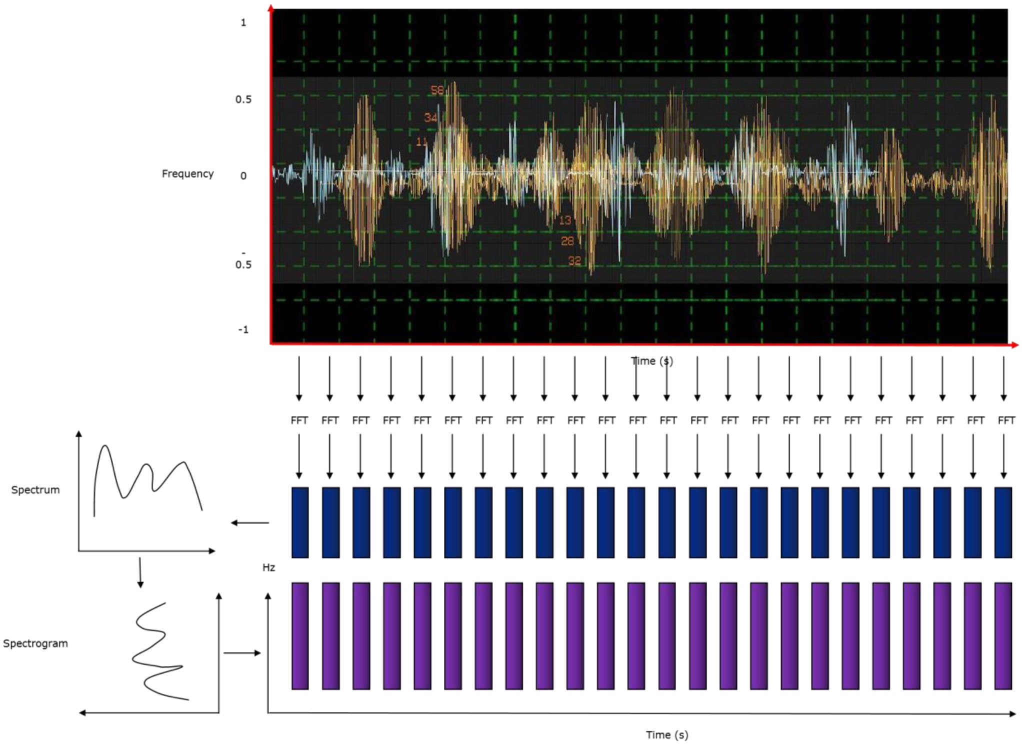

Figure 1: Spectrum of idioms generated by music signals.

Aiming at the low recognition rate of traditional music emotion recognition model, this paper proposes a music emotion recognition model based on DL (see Figure 2). In this model, the spectrogram of music signal features is used as music feature input, and CNN is used to extract features and classify emotions. In the shallow learning feature extraction module, using artificial feature extraction method, and through feature selection, select effective features; In the DL feature extraction module, the gray-scale spectrogram with rich emotional information is obtained as input, and finally the final result is obtained through effective decision fusion.

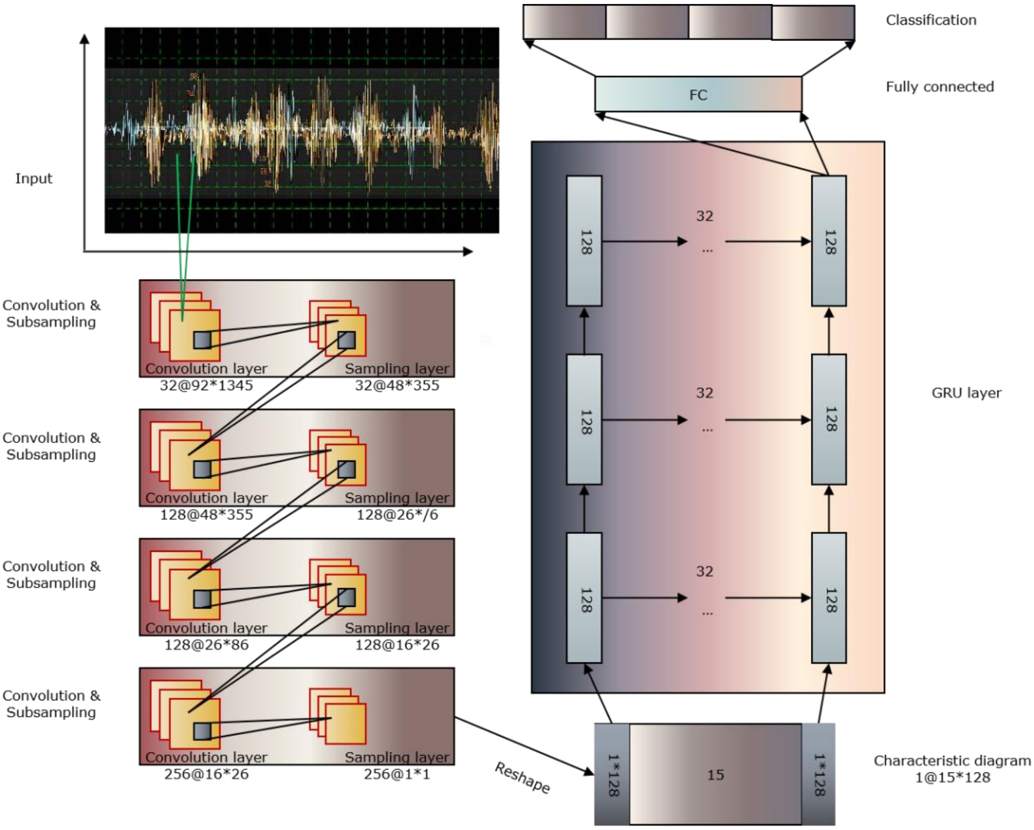

Figure 2: CNN music emotion recognition model.

In CNN model, the input is the spectrogram of audio, and the features are automatically extracted through CNN structure. In this paper, before musical form and emotion recognition, the speech signal is preprocessed and feature extracted, and then the speech features are input into the acoustic model for training. Then use the text corpus to train the language model, and prepare the dictionary connecting the acoustic model and the language model; Finally, the model is used to decode the speech to be recognized and the recognition result is obtained. Suppose a pair of audio samples (X_a, X_b), the generated mixed sample is X_{com}, and the generated X_{com} is only used to calculate KL divergence loss, regardless of the mixing of emotional label dimensions. At the same time, this paper mixes the label dimensions to generate mixed label data Y_{com}, which is used to calculate the cross-entropy loss of classification:

$$X_{com} = (1-\gamma) \times X_a + \gamma \times X_b \tag{3}$$

$$Y_{com} = (1-\gamma)\times Y_a + \gamma \times Y_b \tag{4}$$

Where: γ is the sample mixing coefficient. In order to make the feature expression ability stronger, the eigenvalues are binarized by using the Herveside step function $O(x)$, and the final binarized and weighted results are as follows:

$$S_i^l = \sum_{\omega=1}^{L2} 2^{\omega-1} O\left(I_i^l\right) \quad i = 1,2,3,\ldots,N \tag{5}$$

Divide the output S into B blocks, and make histogram statistics for each block, and get the final features:

$$f_i = \left[B_h\left(S_i^l\right), B_h\left(S_i^2\right), \ldots, B_h\left(S_i^N\right)\right] \in R^{2^{L}2L_1B} \tag{6}$$

Where: f_i represents the feature representation of the i spectrogram obtained through PCANET network; B_h stands for block histogram statistics.

Emotion recognition based on deep learning optimization algorithms requires continuously increasing the trained speech features to update weights. Using audio as the input point and using model feature training for output emotion recognition. Use optimization algorithms for speech feature analysis of emotion recognition. The key of the model lies in the design of convolution kernel, step size, layer number and other parameters in the process of convolution and pooling of CNN, which skillfully reduces the dimension of the frequency direction of the finally obtained feature map to 1. In this way, the characteristics of frequency direction are extracted, and certain time series characteristics are retained.

4 CONSTRUCTION OF MUSIC ANIMATION CAD SYSTEM

In this paper, MIDI file is selected as the input sound source. Because many basic musical features can be directly obtained from it, on this basis, we can also analyze the complex features such as melody, harmony and rhythm, so as to identify the musical form features and divide the whole music into several segments. In the construction of visualization method flow, firstly, MIDI text document is established, that is, a file header block/multiple track blocks; Then there are the music header and MIDI event. The music header contains the block identification and the fast length, while the MIDI event contains the channel information and system information. In the preprocessing stage, the input music file and image file are analyzed respectively, identify the color space and the brightness of the unit pixel in the image file. Strictly control the generation time of each frame, and extend the fixed time after each frame is generated. In this way, the purpose of locking the frame rate can be achieved, and the fixed frame rate is very helpful to control the behavior of the whole animation scene. In the data interaction stage, the random generation effect is established, and the position and timing of particle generation are controlled by music characteristics and image characteristics as dependent variables. In the particle rendering stage of the image, effective feature value positions and particle state analysis were performed. And guide the feature values based on the number of real-time generated examples to achieve the best rendering effect. Guided by the image style of music features, the operation process is shown in Figure 3.

In order to create a sense of immersion in music visualization, it is necessary to present a virtual environment in order to give the audience a high-level experience, which requires the combination of image system and projection system. At the same time, the purpose of music visualization is to express auditory feelings with visual effects, which has real-time and expressive requirements. In this paper, the average value of the data obtained at the current moment is considered as the control parameter.

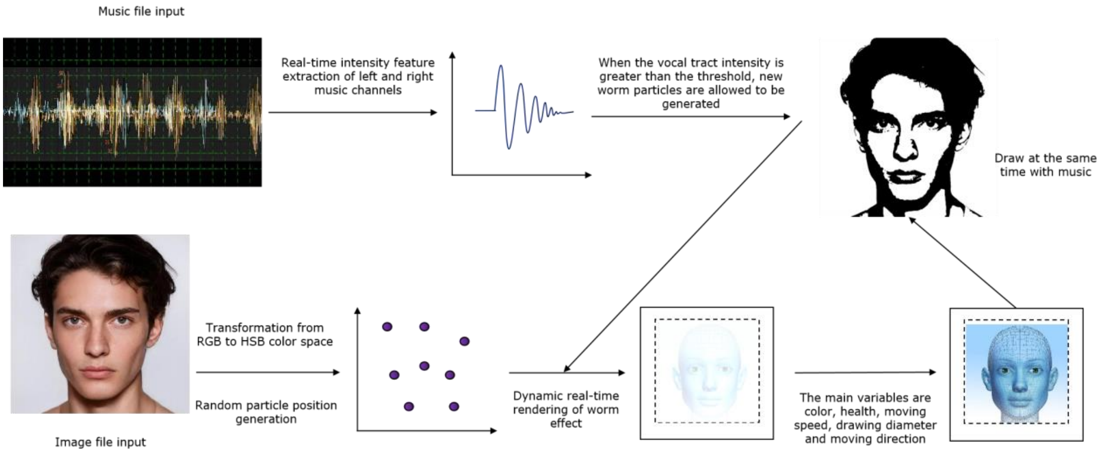
Music file input
Real-time intensity feature extraction of left and right music channels
When the vocal tract intensity is greater than the threshold, new worm particles are allowed to be generated
Draw at the same time with music
Image file input
Transformation from RGB to HSB color space
Random particle position generation
Dynamic real-time rendering of worm effect
The main variables are color, health, moving speed, drawing diameter and moving direction

Figure 3: Operation flow of image style migration guided by music features.

The performance program controller contains six parameters: (1) the controller number, that is, the segment number. (2) The melody tree number of the piece. (3) The opening bar of the passage. (4) The closing bar of the passage. (5) The emotional classification of the passage. (6) The number of the basic performance program matched for the piece. The music visual image is generated. For a hidden unit, use x_t to represent the input of step t. The activation value of the current unit is:

$$s = f\left(U_{xt} + W_{S_{t-1}}\right) \tag{7}$$

Among them, f represents the activation function, and ReLU is used in this article. The output of step t is calculated by the Softmax layer. The value i_t of the input gate unit controls how much of the input at the current time point can enter the memory unit, and its calculation expression is:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{8}$$

Among them, W_{xi}, W_{hi}, W_{ci} are used to control the connection weights related to the input gate, and b_i is the bias term. Framing processing is performed to divide the time-domain discrete signal into overlapping frames:

$$X_{STFT}(k,n) = \sum_{m=0}^{N-1} x(n-m)w(m)e^{-j2k\pi m/N} \tag{9}$$

Strong $X_{STFT}(k,n)$ maps to a twelve-dimensional vector $p(k)$, each dimension representing the intensity of a semitone level. The mapping formula is:

$$p(k) = \left[12\log_2\left(\frac{k}{N \cdot f_{sr}} \cdot f_{ref}\right)\right]\mathrm{mod}12 \tag{10}$$

Among them, f_{ref} is the reference frequency, and f_{sr} is the sampling rate. Accumulate the frequency values of the frequency points corresponding to each sound level to obtain the value of each sound level contour feature component of each time segment. The formula is as follows:

$$PCP(p) = \sum_{k=p(k)=p} \left|X(K)\right|^2 \qquad p = 1,2,3,\ldots,11 \tag{11}$$

The tone contour features represent a tone scale by a twelve-dimensional vector, which reflects the relative strength of notes under the scale of chromatic scale in each twelve-chromatic interval. Segment matching is to find the most suitable basic performance program for each segment according to emotion. Each basic performance program not only records the names and time of all actions, but also the emotional color embodied in the basic performance program. When the speed of music changes, the system should automatically adjust the movement time according to the change of speed, so as to realize the synchronization of music and animation.

5 SYSTEM TESTING AND ANALYSIS

5.1 Emotional Classification Test

The experiment uses embedded setting information with lyrics text as a vector matrix to calibrate text content such as songs. It used word matrix vectors for parameter model training. The input parameters of CNN network are in the form of [batch _ size, height, width, channels]. Considering the computer memory and the complexity of the model, batch_size is 24. Take spectrogram as the input of CNN network. An example of spectrogram is given, and the final size of the image is 96*1366 after down-sampling. Input 24 spectrograms with the size of 96*1366 at a time. The generating parameters of spectrograms are frame length of 256 and frame shift of 128. Each picture is a single channel. CNN network mainly includes three convolution layers, and each convolution layer uses Relu nonlinear activation function. Two-layer pooling mode is maxpool, and there are 512 Cell; connecting the two fully connected layers. Finally, emotion recognition is output through Softmax layer. Set the number of neurons to 256 and dropout to 0.5. In order to make the results more convincing, this paper derives the network convergence trend diagram in the process of solving, as shown in Figure 4.

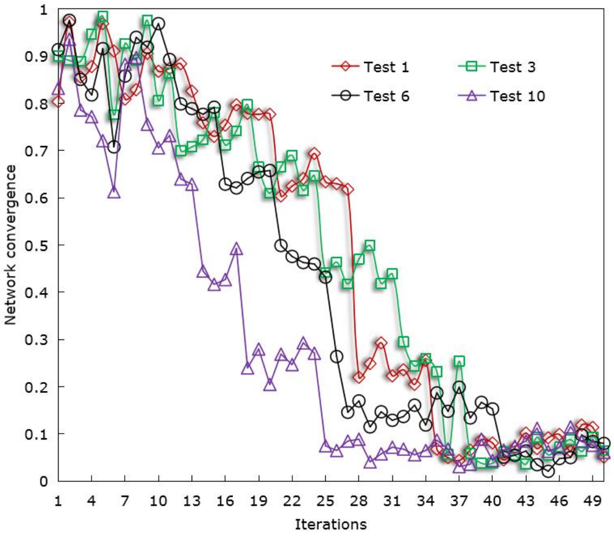

Figure 4: Network convergence trend diagram.

It can be seen from the figure that after about 28 iterations, the output error of the algorithm has converged to a certain extent.

The classification results of this experiment are listed in Table 1 and Table 2. The arithmetic average of the last 10 groups of performance index data after model convergence is taken as the performance index of this training every time.

Training times	*Accuracy*	*Recall*	*Precision*	*F1*
1	0.828	0.755	0.778	0.745
2	0.821	0.749	0.773	0.749
3	0.812	0.768	0.771	0.751
4	0.799	0.725	0.725	0.699
5	0.834	0.799	0.783	0.765
6	0.827	0.763	0.785	0.748
7	0.813	0.752	0.752	0.723
8	0.809	0.777	0.783	0.761
9	0.819	0.783	0.788	0.765
10	0.801	0.744	0.735	0.718
Average value	0.816	0.762	0.767	0.742

Table 1: Cross-validation results of CNN music emotion recognition model.

Training times	*Accuracy*	*Recall*	*Precision*	*F1*
1	0.523	0.398	0.401	0.375
2	0.519	0.372	0.393	0.377
3	0.507	0.369	0.377	0.369
4	0.502	0.388	0.382	0.375
5	0.506	0.381	0.415	0.378
6	0.511	0.383	0.407	0.372
7	0.499	0.373	0.378	0.354
8	0.405	0.385	0.377	0.333
9	0.501	0.382	0.385	0.381
10	0.504	0.394	0.399	0.386
Average value	0.498	0.383	0.391	0.371

Table 2: Cross-validation results of RNN music emotion recognition model.

From the experimental results in the table, it can be seen that compared with using RNN to extract the time series features of spectrogram, the emotion classification results obtained by using this improved CNN network are better. Among them, the average Accuracy of CNN music emotion recognition model is 0.816, the average Recall is 0.762, the average Precision is 0.767, and the average F1 is 0.742. However, the average Accuracy, Recall, Precision and F1 of RNN musical emotion recognition model are 0.498, 0.383, 0.391 and 0.371 respectively. The results show that this method is effective and reliable, because it makes full use of the advantages of the recognition model and enriches the ways to obtain emotional information.

5.2 Emotional Visual Simulation of Music Animation CAD System

In this paper, 100 representative classical music, 120 hip-hop music and 100 country music in wav format are selected as the music library of the experiment, with a total of 320 pieces of music with a sampling rate of 44100Hz and eight bits of storage. The input parameters of the classifier are as follows: strength, speed, timbre, melody line direction and melody note density. These parameters are obtained by calculating the numerical values of some basic features and musical features. Through experiments, the emotion mapping accuracy of the music animation CAD system constructed in this paper is shown in Figure 5. The comparison of emotion mapping accuracy of different methods is shown in Figure 6.

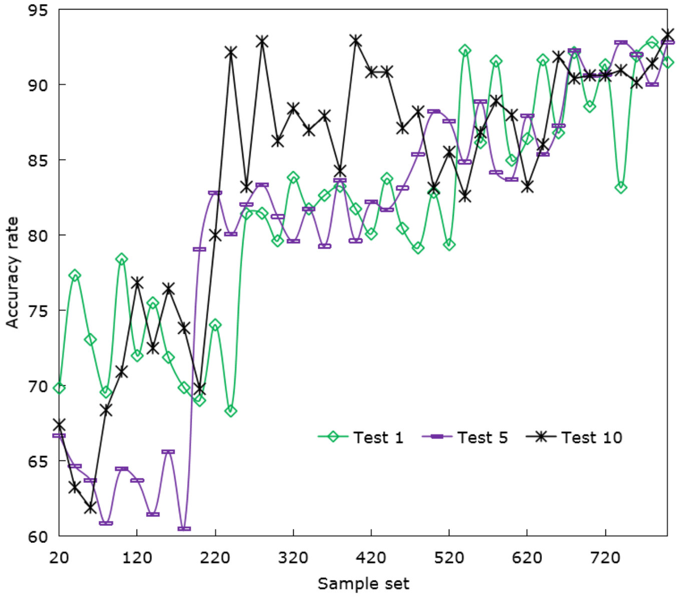

Figure 5: Accuracy of emotion mapping by CNN method.

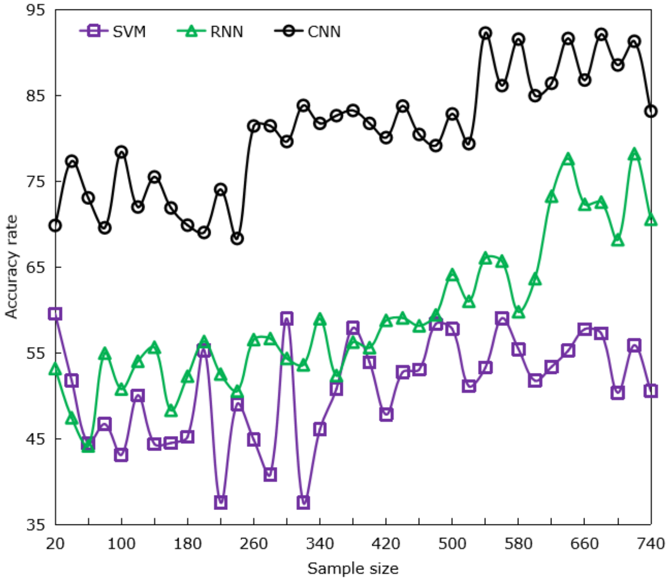

Figure 6: Comparison of accuracy of emotional mapping.

The experiments in Figure 5 and Figure 6 show that the emotion mapping accuracy of the music animation CAD system constructed by CNN method is high, which can reach about 90%, and the accuracy is stable from the results of many experiments. The emotion mapping accuracy of music animation CAD system constructed by RNN method is the second, which can reach about 80%. However, the emotion mapping accuracy of the music animation CAD system constructed by SVM method is the worst, and its accuracy only reaches 60%.

Music files and image files have an impact on the visual expression of image style transfer at the same time, and the visual images and style attributes generated by different collocation methods are unique. The image reconstruction process is dynamic and visible, and the whole process from blank screen to portrait effect presentation is synchronized with the music playback process in real time. According to the speed of the music, the time interval between adjacent actions in the basic performance program is determined, and then some special actions are designed for key notes. On the basis of segment matching, note matching makes the design of music animation combine point with surface. In addition, by simplifying the physical model and indexing the vertices of the model, the memory occupancy rate is greatly reduced. Figure 7 shows the subjective scoring results of animation generated by the music animation CAD system.

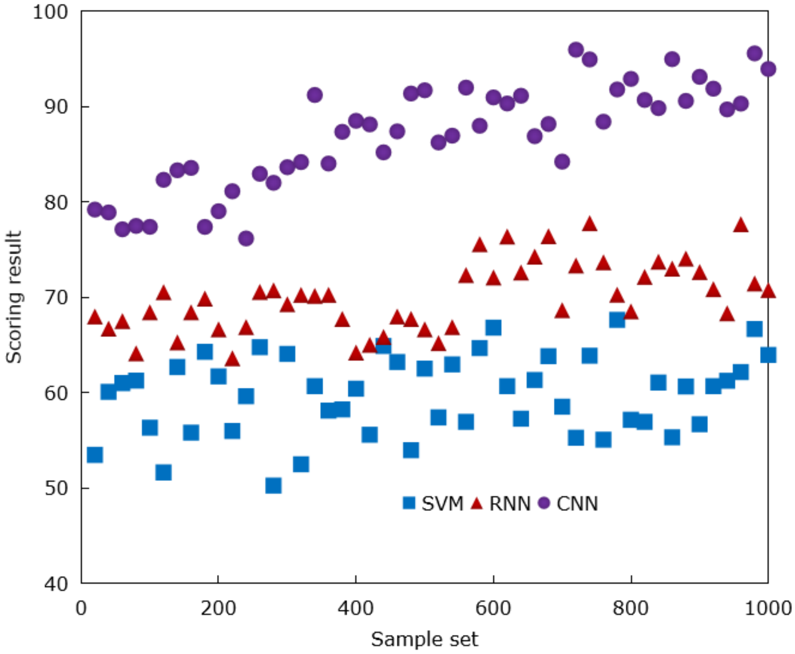

Figure 7: Subjective scoring result.

The comparison results in Figure 7 show that the music animation CAD system constructed by CNN has the highest subjective score, with the highest score of 90%. This value is much better than RNN method and SVM method. A musical animation CAD system is constructed. In this section, satisfactory music visualization effect has been achieved through experiments. At the same time, it has certain reference significance for generating more complex music special effects, as well as virtual reality, scene animation and so on.

6 CONCLUSIONS

The interaction design required for music visualization includes the changes of artificial intelligence driven by data self-interaction, as well as the changes of human-computer interaction interface between designers and users after watching music visualization images. Through the visual modeling selection of basic units, the quantity control of dense arrangement and the selection of arrangement and combination, the customized design of large-scale customers is carried out, and the diversified design results of uncertain interactive changes are achieved. The design of music animation based on music visualization can not only fully reflect the designer's unique ideas, but also greatly save the manpower and material resources for music animation production. And a musical animation CAD system is constructed. Simulation results show that compared with using RNN to extract the time series features of spectrogram, the emotion classification results obtained by using this improved CNN network are better. At the same time, the emotion mapping accuracy of the music animation CAD system constructed by CNN method is high, which can reach about 90%. The data results show that this method is effective and reliable. The experiment has achieved satisfactory visual effect of music. The visual effect generated by the music animation CAD system constructed in this paper can help users to generate unique personal portraits with music. The visual effect is also richer and can be displayed visually with more representative features, which makes people easier to understand and accept. This effect plays a positive role in the audience's better understanding of the theme of music, and it has certain reference significance for generating more complex music special effects, as well as virtual reality, scene animation and so on.

Dan Shen, https://orcid.org/0000-0002-5795-672X
Wenjia Zhao, https://orcid.org/0009-0004-8244-842X

REFERENCES

[1] Agarwal, G.; Om, H.: An efficient supervised framework for music mood recognition using autoencoder - based optimised support vector regression model, IET Signal Processing, 15(2), 2021, 98-121. https://doi.org/10.1049/sil2.12015

[2] Chaturvedi, V.; Kaur, A.-B.; Varshney, V.; Garg, A.; Chhabra, G.-S.; Kumar, M.: Music mood and human emotion recognition based on physiological signals: a systematic review, Multimedia Systems, 28(1), 2022, 21-44. https://doi.org/10.1007/s00530-021-00786-6

[3] Chaudhary, D.; Singh, N.-P.; Singh, S.: A survey on autonomous techniques for music classification based on human emotions recognition, International Journal of Computing and Digital Systems, 9(03), 2020, 1-15. http://dx.doi.org/10.12785/ijcds/090308

[4] Chen, J.; Jiang, D.; Zhang, Y.: A common spatial pattern and wavelet packet decomposition combined method for EEG-based emotion recognition, Journal of Advanced Computational Intelligence and Intelligent Informatics, 23(2), 2019, 274-281. https://doi.org/10.20965/jaciii.2019.p0274

[5] Dong, Y.; Yang, X.; Zhao, X.; Li, J.: Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition, IEEE Transactions on Multimedia, 21(12), 2019, 3150-3163. https://doi.org/10.1109/TMM.2019.2918739

[6] Ehrlich, S.-K.; Agres, K.-R.; Guan, C.; Cheng, G.: A closed-loop, music-based brain-computer interface for emotion mediation, PloS One, 14(3), 2019, e0213516. https://doi.org/10.1371/journal.pone.0213516

[7] Er, M.-B.; Aydilek, I.-B.: Music emotion recognition by using chroma spectrogram and deep visual features, International Journal of Computational Intelligence Systems, 12(2), 2019, 1622-1634. https://doi.org/10.2991/ijcis.d.191216.001

[8] Ge, M.; Tian, Y.; Ge, Y.: Optimization of computer aided design system for music automatic classification based on feature analysis, Computer-Aided Design and Applications, 19(S3), 2021, 153-163. https://doi.org/10.14733/cadaps.2022.S3.153-163

[9] Gómez, C.-J.-S.; Cano, E.; Eerola, T.; Herrera, P.; Hu, X.; Yang, Y.-H.; Gómez, E.: Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications, IEEE Signal Processing Magazine, 38(6), 2021, 106-114. https://doi.org/10.1109/MSP.2021.3106232

[10] Han, D.; Kong, Y.; Han, J.; Wang, G.: A survey of music emotion recognition, Frontiers of Computer Science, 16(6), 2022, 166335. https://doi.org/10.1007/s11704-021-0569-4

[11] Hizlisoy, S.; Yildirim, S.; Tufekci, Z.: Music emotion recognition using convolutional long short-term memory deep neural networks, Engineering Science and Technology, an International Journal, 24(3), 2021, 760-767. https://doi.org/10.1016/j.jestch.2020.10.009

[12] Jing, Y.; Song, Y.: Application of 3D reality technology combined with CAD in animation modeling design, Computer-Aided Design and Applications, 18(S3), 2020, 164-175. https://doi.org/10.14733/cadaps.2021.S3.164-175

[13] López, H.-J.-L.; González, C.-I.; López, C.-J.-L.; Ruiz, M.-B.: Towards the recognition of the emotions of people with visual disabilities through brain–computer interfaces, Sensors, 19(11), 2019, 2620. https://doi.org/10.3390/s19112620

[14] Mazhar, T.; Malik, M.-A.; Nadeem, M.-A.; Mohsan, S.-A.-H.; Haq, I.; Karim, F.-K.; Mostafa, S.-M.: Movie reviews classification through facial image recognition and emotion detection using machine learning methods, Symmetry, 14(12), 2022, 2607. https://doi.org/10.3390/sym14122607

[15] Orjesek, R.; Jarina, R.; Chmulik, M.: End-to-end music emotion variation detection using iteratively reconstructed deep features, Multimedia Tools and Applications, 81(4), 2022, 5017-5031. https://doi.org/10.1007/s11042-021-11584-7

[16] Pandeya, Y.-R.; Bhattarai, B.; Lee, J.: Deep-learning-based multimodal emotion classification for music videos, Sensors, 21(14), 2021, 4927. https://doi.org/10.3390/s21144927

[17] Saha, S.; Ghosh, M.; Ghosh, S.; Sen, S.; Singh, P.-K.; Geem, Z.-W; Sarkar, R.: Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm, Applied Sciences, 10(8), 2020, 2816. https://doi.org/10.3390/app10082816

[18] Visnu, D.-S.; Balaji, B.; Kirubha, H.-K.-S.: Music recommendation system based on facial emotion recognition, Journal of Computational and Theoretical Nanoscience, 17(4), 2020, 1662-1665. https://doi.org/10.1166/jctn.2020.8420

[19] Xu, L.; Wen, X.; Shi, J.; Li, S.; Xiao, Y.; Wan, Q.; Qian, X.: Effects of individual factors on perceived emotion and felt emotion of music: based on machine learning methods, Psychology of Music, 49(5), 2021, 1069-1087. https://doi.org/10.1177/0305735620928422

[20] Yang, C.; Li, Q.: Music emotion feature recognition based on Internet of things and computer-aided technology, Computer-Aided Design & Applications, 19(S6), 2021, 80-90. https://doi.org/10.14733/cadaps.2022.S6.80-90