Hindawi
Scientific Programming
Volume 2021, Article ID 7087588, 10 pages
https://doi.org/10.1155/2021/7087588

Research Article

Research on Multimodal Music Emotion Recognition Method Based on Image Sequence

Zhao Yu

Guangxi Arts University, College of Music Education, Department of Keyboard Instruments, Nanning 530022, Guangxi Province, China

Correspondence should be addressed to Zhao Yu; 20100010@gxau.edu.cn

Received 26 September 2021; Revised 14 October 2021; Accepted 23 October 2021; Published 6 December 2021

Academic Editor: Bai Yuan Ding

Copyright © 2021 Zhao Yu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The work of music performance system is to control the light change by identifying the emotional elements of music. Therefore, once the identification error occurs, it will not be able to create a good stage effect. Therefore, a multimodal music emotion recognition method based on image sequence is studied. The emotional characteristics of music are analyzed, including acoustic characteristics, melody characteristics, and audio characteristics, and the feature vector is constructed. The recognition and classification model based on neural network is trained, the weight and threshold of each layer are adjusted, and then the feature vector is input into the trained model to realize the intelligent recognition and classification of multimodal music emotion. The threshold of the starting point range of a specific humming note is given by the center clipping method, which is used to eliminate the low amplitude part of the humming note signal, extract the short-time spectral structure features and envelope features of the pitch, and complete the multimodal music emotion recognition. The results show that the calculated kappa coefficient k is greater than 0.75, which shows that the recognition and classification results are in good agreement with the actual results, and the classification and recognition accuracy is high.

1. Introduction

Music is an art form that takes sound as a means of communication and then produces emotional experience. Music can communicate emotion directly in the form of sound movement. The essence of music is emotion. The specific form of music sound wave vibration is directly related to human emotion. According to this connection, music can be used to describe people's emotional activities in detail. All music activities obey and reflect the fluctuations of people's inner world, whether the creators and performers vent their emotions or the listeners accept the emotional connotation of the music. Nowadays, digital music technology has brought great changes to music, a traditional and classic way of emotional communication. The development of computer science has brought revolutionary progress to the creation, communication, storage, and release of music works. Especially with the continuous enrichment of computer music materials, it has become an urgent scientific research topic to study the emotional information of music works by using intelligent information analysis and processing methods, so as to make the computer have the ability to recognize and express music emotions like people. Music appeared earlier than language. When human beings did not use language to express their feelings, they had learned to use music [1, 2]. It can be said that music plays an important role in human history, and music has been integrated into all aspects of human life [3]. With the continuous development of science and technology, the creation, storage, and dissemination of music have been greatly changed. Music is an art form that takes sound as a means of communication and then produces emotional experience. Music can directly carry out emotional communication in the form of sound movement [4, 5]. It can be said that the essence of music is emotion. The specific form of music acoustic vibration is directly related to human emotion. According to this connection, music can be used to describe people's emotional activities in detail [6]. All music activities obey and reflect the fluctuations of people's inner world, whether the creators and performers vent their emotions or the listeners accept the emotional

connotation of the music. Nowadays, digital music technology has brought great changes to music, a traditional and classic way of emotional communication. The development of image sequence has brought revolutionary progress to the creation, communication, storage, and release of music works. Generally, image sequence noise is an unpredictable random signal. Noise is very important for image sequence processing. It affects all links of input, acquisition and processing of image processing, and the whole process of output results [7, 8]. In particular, the input of image and the suppression of acquisition noise are very key problems. If the input is accompanied by large noise, it will inevitably affect the whole process and output results. Therefore, a good image sequence processing system, whether analog processing or digital processing by computer, takes reducing the noise of the first level as the main target [9, 10]. In particular, with the continuous enrichment of computer music materials, it has become an important research content to use the image sequence intelligent information analysis and processing method to study the emotional information of music works, so as to make the computer have the ability to recognize and express multimodal music emotions like people.

In this regard, relevant scholars have proposed many studies. Reference [11] proposed the common neural mechanism of emotion processing in music and vocalization and compared the neural mechanisms involved in vocalization and music processing, so as to observe their possible similarities in emotional content coding. Positive and negative emotional sounds (such as laughter and crying) and violin music stimuli extracted by numbers are used as stimuli, which have common melody contour and main pitch/frequency characteristics. Reference [12] proposed that the semantic and episodic memory of music are provided by different neural networks, and the extraction of brain semantic memory and episodic memory is completed by different neural networks. It is basically obtained through language and visual space materials. Two delay identification tasks are constructed, one containing only familiar items and the other only unfamiliar items. For each recognition task, the general extraction target is presented in the previous semantic task. By comparing two perceptual control tasks with another perceptual control task, the situational task and semantic task are compared. Based on the above analysis, a multimodal music emotion recognition method based on image sequence is proposed. The music emotion features including acoustic features, melody features, and audio features are analyzed, and the feature vector is constructed. The recognition and classification model based on neural network is trained, the weight and threshold of each layer are adjusted, and the feature vector is input into the trained model to realize the intelligent recognition and classification of multimodal music emotion. The threshold of the starting point range of a specific humming note is given by the center clipping method, which is used to eliminate the low amplitude part of the humming note signal, extract the short-time spectral structure features and envelope features of the pitch, and complete the multimodal music emotion recognition. The recognition and expression of multimodal music emotion enable users to realize emotional human-computer interaction through music, which enriches the research content of human-computer interaction technology.

2. Multimodal Music Emotion Recognition and Classification Based on Image Sequence

In addition to the necessary music itself, a perfect music performance is a complementary live atmosphere. In music performance, the contrast of the on-site atmosphere is mainly realized by lighting, which is often changed with the emotional factors expressed in the music to assist the music to create a good stage effect. In this context, in order to better control the light, multimodal music emotion recognition is very important [13–15]. Therefore, aiming at multimodal music emotion, a classification and recognition model is constructed to complete the research on intelligent recognition and classification of multimodal music emotion in music performance system.

2.1. Analysis of Emotional Characteristics of Multimodal Music. The realization of multimodal music emotion recognition is based on multimodal music emotion features, so multimodal music emotion feature extraction is the first link of this research [16, 17]. In the previous multimodal music emotion classification, most of them take a music feature as the classification basis. Although they can also complete the classification task, their accuracy cannot be guaranteed. In order to solve the above problems, in this study, a variety of music features are extracted and fused based on image sequences and then classified and recognized based on fusion features. The principle of image sequence is shown in Figure 1.

In order to identify the emotional characteristics in music [18], it is necessary to understand the composition of music. Among them, the music related factors that can obviously show emotional characteristics include acoustic characteristics, melody characteristics, and audio characteristics.

2.1.1. Acoustic Characteristics. Acoustic feature refers to the physical quantity that represents the acoustic characteristics of multimodal music speech. It is also a general term for the acoustic performance of many elements of sound, for example, the energy concentration area, formant frequency, formant intensity, and bandwidth representing the timbre of multimodal music, as well as the duration, fundamental frequency, and average voice power representing the prosodic characteristics of multimodal music speech. For the classification of multimodal music speech, the traditional method is to study the characteristics of pronunciation organs, such as the tongue position of vowels, front and back, and the pronunciation position of consonants. Now, with the progress of science and technology, further fine research can be made according to the acoustic characteristics.

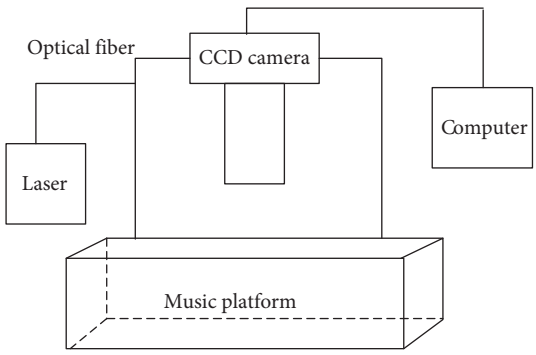

Figure 1: Schematic diagram of image sequence.

Acoustic factor is the most basic component of music [19, 20]. Music with different emotions shows different acoustic characteristics, and the basic corresponding relationship is shown in Table 1.

2.1.2. Melody Characteristics. Melody features are also called melody features; that is, the lines composed of high and low tones with different lengths are the soul of music and the melody of music. The tones are organized according to certain laws [21–23]. The extracted features include five aspects.

(1) Balance parameter Y_1: Balance refers to the proportional value of the volume in the left and right channels. The calculation formula is as follows:

$$Y_1 = P_{\text{an}}(k), k = 1, 2, \ldots, 16. \tag{1}$$

(2) Volume parameter Y_2: Volume refers to the loudness of the sound that can be heard by the human ear. The calculation formula is as follows:

$$Y_2 = \frac{V_{\text{olume}}}{127}. \tag{2}$$

(3) Pitch parameter Y_3: Pitch refers to the vibration frequency of the fundamental frequency of a note. Fast paced music has fast vibration frequency; on the contrary, it has slow vibration frequency. The calculation formula is as follows:

$$Y_3 = \frac{1}{n}\sum_{i=1}^{n} P_{\text{itch}}. \tag{3}$$

(4) Average strength parameter Y_4: Strength refers to the strength of the power generated by music. Soothing music has weak strength, while more shocking music has strong strength [24, 25]. The calculation formula is as follows:

$$Y_4 = \sum_{i=1}^{N} \frac{V_{el(k,i)}}{N}, k = 1, 2, \ldots, 16. \tag{4}$$

(5) Note energy parameters Y_5: Note energy refers to the sum of the product of note pitch and length. The calculation formula is as follows:

$$Y_5 = \sum_{i=1}^{n}\left(P_{ij} \times D_{ij}\right), j = 1, 2, \ldots, 16. \tag{5}$$

In the formula, $P_{\text{an}}(k)$ represents the balance value of left and right channels, and its value range is 0–127; V_{olume} represents the volume of the track, with a range of 0–127; P_{itch} stands for note pitch; n represents the number of notes in the track; $V_{el(k,i)}$ represents the intensity value of the i note in the k track; k indicates track number; N represents the number of notes in the k track; P_{ij} and D_{ij} represent the pitch and length of i notes in the j track channel.

2.1.3. Audio Features. Audio feature is an important condition for recognizing and identifying multimodal music emotion. Different music emotion is expressed through different audio features. Audio is one of the important influencing factors in music, which affects the rhythm of music. The faster the rhythm, the more obvious the audio, and the happier the multimodal music emotion expressed. On the contrary, multimodal music emotion is more dull or depressing [26, 27]. The description of audio features based on image sequences can be carried out from two aspects, real-time domain features and frequency domain features [28].

(1) Time domain characteristics

The time domain characteristics of audio refer to the time domain parameters of each frame calculated from the music signal, mainly including zero crossing rate and amplitude [29–31]. The following is a specific analysis.

(1) Zero crossing rate Z_n: Zero crossing rate refers to the frequency at which the audio signal waveform passes through the zero level. Generally speaking, the zero crossing rate in the high-frequency band of a piece of music will be relatively high; on the contrary, the zero crossing rate will be relatively low. Through this parameter, we can well distinguish between voiced and unvoiced sounds in music. Generally, unvoiced sounds are mostly used in cheerful music, while voiced sounds are often used in slow and deep music. The calculation formula of zero crossing rate is as follows:

$$Z_n = \frac{\sum_{m=1}^{N} \text{sgn}\left[s_n x(m)\right] - \text{sgn}\left[s_n x(m-1)\right]}{2N}. \tag{6}$$

In the formula, $s_n x(m)$ represents the symbol function of the audio signal $x(m)$; N represents the effective width of the window; n represents the time position of the window.

(2) Range M_n: Amplitude refers to the width expanded by the waveform vibration of audio signal [32–34]. The more passionate the music, the greater the audio amplitude. The more soothing the music, the smoother the audio amplitude. The audio amplitude is described as follows:

Table 1: Corresponding relationship between acoustic characteristics and multimodal music emotion.

Acoustic characteristics	Happy	Hate	Anger	Sadness	Fear
Pronunciation	Normal	Normal	Tighten	Vague	Clear
Pitch mean	Very high	Very low	Very high	Slightly lower	Very high
Pitch range	Very wide	Slightly wider	Very wide	Slightly narrow	Very wide
Pitch change	Smooth, curved up	Wide, downward bending	Stress mutation	Bend down	Normal
Tone quality	Breathing sound, singing sound	Mumble, chest sound	Breathing sound	Resonance sound	Sharp voice
Speed of speech	Fast or slow	Very fast	Slightly faster	Slightly slower	Soon
Strength	High	Low	High	Low	Normal

$$M_n = \sum_{m=n-(N-1)}^{n} |x(m)w(n-m)| = \sum_{m=n-(N-1)}^{n} |x(m)|w(n-m). \quad (7)$$

In the formula, $w(n-m)$ represents the moving window function.

(3) Frequency domain characteristics: The frequency domain characteristics of audio include two: spectral centroid C_t and spectral flux F_t. The calculation formula is as follows:

$$\text{Spectrum centroid } C_t = \frac{\sum_{n=1}^{N} M_t[n] \times n}{\sum_{n=1}^{N} M_t[n]}, \quad \text{Spectral flux } F_t = \sum_{n=1}^{N} \left(N_t[n] - N_{t-1}[n]\right)^2. \quad (8)$$

In the formula, $M_t[n]$ represents the amplitude of the short-time spectrum of the t frame at the frequency point n; $N_t[n]$ and $N_t[n-1]$ represent the normalized amplitude of the spectrum of the t frame and the $t-1$ frame at the frequency point n, respectively.

Based on the above three categories and 14 multimodal music emotional features, a feature vector is formed, which is used to describe the emotional factors of a piece of music. It is described as follows:

$$U = \{U_1, U_2, U_3\}. \quad (9)$$

In the formula, U_1 represents acoustic characteristics; U_2 represents melody characteristics; U_3 represents audio characteristics. The audio feature structure is shown in Figure 2.

2.2. Construction of Multimodal Music Emotion Recognition Classification Model. Based on the emotional features contained in the above music, a classification and recognition model is established to realize multimodal music emotion recognition and classification, and a neural network is used to construct the model [35, 36]. BP neural network is an intelligent algorithm invented by simulating the working principle of human brain neural network. The neural network mainly includes three

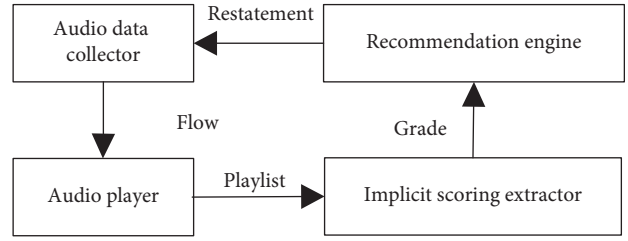

Figure 2: Audio feature structure.

layers, and the classification processing is realized through the operation of each layer. The classification and recognition model constructed by this algorithm is shown in Figure 3.

In Figure 3, training is the key in model construction, and the specific process is as follows. First, enter the choice of training samples, and after operation of hidden layer and output layer, you will get results, and then to compare the results with the expected results, when the difference between them is less than the set threshold, the training is completed; otherwise, there will be back propagation, difference from the output to the input, and repetitive process, until you reach the optimal weight and threshold. The purpose of BP neural network training is to adjust and optimize the weights and thresholds connected at every two levels in the model. Therefore, the formula is given as follows.

(1) Adjustment formula of connection weight w_{ij} and threshold θ_j between input layer and hidden layer:

$$\begin{cases} w_{ij}(N+1) = w_{ij}(N) + \beta \cdot \mu_j^k \cdot c_i \\ \theta_j(N+1) = \theta_j(N) + \beta \cdot \mu_j^k \\ i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, p \\ 0 < \beta < 1 \end{cases}. \quad (10)$$

In the formula, μ_j^k represents the error value in the hidden layer; c_i represents the input eigenvector; N represents the number of iterations; k represents the number of training samples; n represents the number of neurons in the input layer; p represents the number of neurons in the hidden layer.

(2) Adjustment formula of connection weight v_{jt} and threshold γ_t between hidden layer and output layer:

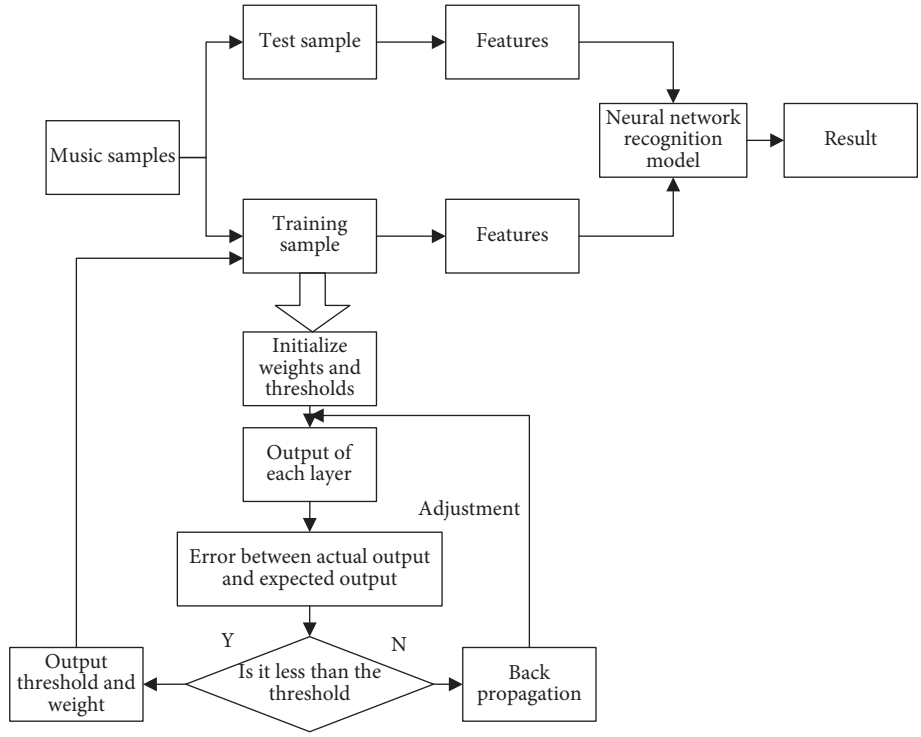

Figure 3: Classification and recognition model based on BP neural network.

$$\begin{cases} v_{jt}(N+1) = v_{jt}(N) + \alpha \cdot d_t^k \cdot y_j \\ \gamma_t(N+1) = \gamma_t(N) + \alpha \cdot d_t^k \\ j = 1, 2, \cdots, p \\ t = 1, 2, \cdots, q \\ 0 < \alpha < 1 \end{cases} . \quad (11)$$

In the formula, d_t^k represents the error value between the target eigenvector and the actual output vector; y_j represents the output of the hidden layer.

The trained model based on BP neural network can realize multimodal music emotion classification by inputting test music samples.

2.3. Intelligent Recognition of Note Starting Point Based on Clipping

2.3.1. Calculation of Correlation Function between Note Signals. In the process of intelligent optimization and recognition of the note starting point of feature tone retrieval, the initial note signal is preprocessed based on the image sequence to filter the noise of the high-frequency part. The random note signal is divided into short-term stationary signals based on the image sequence, the similarity between different phonetic waveform signals is calculated, and the cross-correlation function between each note signal is obtained. The design of recognition framework based on image sequence is shown in Figure 4.

Pervasive environment combines network technology and mobile technology and designs a customer-oriented adaptive recommendation structure. Pervasive environment is composed of network devices, including computers, mobile phones, and various network connected appliances, and network services including computing, management, and control. In this environment, the network can collect query, configuration, and management information from

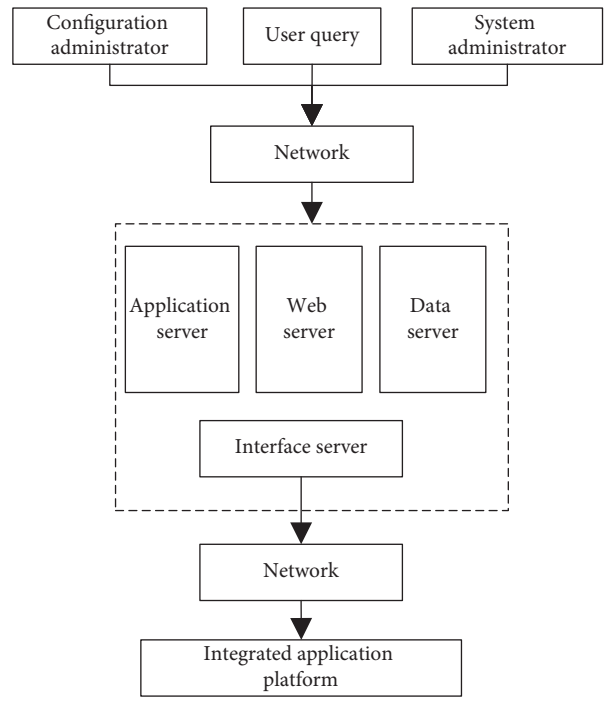

Figure 4: Frame diagram of recognition based on image sequence.

users and administrators, transfer these pieces of information to each server port, and then apply them to the comprehensive platform through the network to provide the basis for the design of the recommendation system.

The specific steps are detailed as follows.

Assuming that n represents the note frame length and N represents the sampling points in the frame, each humming note signal in the feature tone retrieval is windowed and framing processed by formula (11), so as to make each humming note signal short-term stable:

$$W(n) = \frac{x(n) \times E(n)}{N \times n}. \tag{12}$$

In the formula, $x(n)$ represents any humming note signal and $E(n)$ represents the short-time energy of $x(n)$.

Assuming that $s(k)$ represents the current sampling value of short-time humming note signal, $s(k)$ is defined as the linear combination of historical sampling value and excitation signal, which is expressed by the following formula:

$$s(k) = \frac{s(n) \times e(n) \times v(n)}{a_i \times p} \times G. \tag{13}$$

In the formula, a_i represents the prediction coefficient of the image sequence, p represents the prediction order of the image sequence, G represents the gain factor of the image sequence, $e(n)$ represents the glottic pulse excitation of the image sequence, and $v(n)$ represents the channel response value of the image sequence.

$x(n)$ is judged as the result of glottic pulse excitation $e(n)$ filtered by channel response $v(n)$, and $e(n)$ is a short-time humming note signal with periodic characteristics.

Assuming that $R_{\text{corss}}(t)$ represents a function with the same period, the similarity between waveform signals of different humming notes is calculated by the following formula:

$$R_{\text{auto}}(t) = \frac{1}{N} \sum_{n=1}^{N} x(n)x(n+t). \tag{14}$$

The similarity between the waveform signals of different humming notes mainly has two states: Cross and Jiugong grid, as shown in Figures 5 and 6.

Regular squares are used to represent the similarity between different humming note waveform signals. Generally, the image sequence value is 0 or 1. The two-dimensional space is formed by a large number of image sequences. The adjacent elements are the subelements to be studied, and their shape is mainly square. $y(n)$ represents $x(n)$ and signals with the same period T, and the mathematical expression for discrete-time signals is given by the following formula:

$$R_{\text{cross}}(t) = \frac{1}{N} \sum_{n=1}^{N} x(n) \times y(n)^T. \tag{15}$$

Based on the image sequence, the center clipping method is used to give the threshold of the starting point range of specific humming notes, which is expressed as follows:

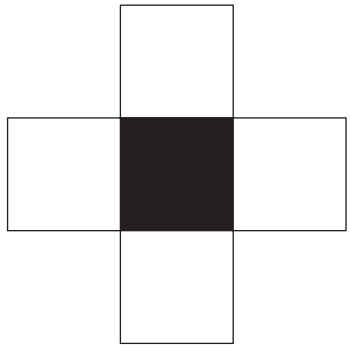

Figure 5: Cross structure.

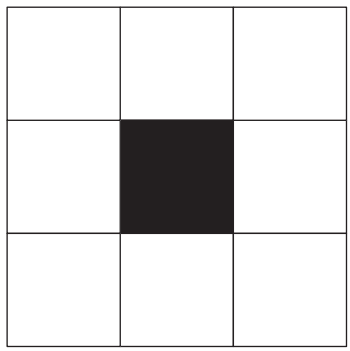

Figure 6: Jiugong lattice structure.

$$x''(n) = \frac{s(n) + g(n)}{y(n) \times x(n)}. \tag{16}$$

In the formula, represents Additive White Noise Gaussian independent of $s(n)$. $y(n)$ is the third-order level signal of humming note searched by $x(n)$ through clipping method, eliminates the low amplitude part of humming note signal, and calculates the correlation function between humming note starting point signals.

To sum up, it can be explained that, in the process of intelligent optimization and recognition of note starting point of feature tone retrieval, the initial note signal is preprocessed, the similarity between different note waveform signals is calculated, and the cross-correlation function between each note signal is obtained, which lays a foundation for intelligent optimization and recognition of note starting point of feature tone retrieval.

2.3.2. Intelligent Optimization Recognition of Note Starting Point Based on Starting Point Feature. Because the melody pitch feature extraction is a key link in the intelligent optimization and recognition of the note starting point of the feature tone retrieval and directly affects the quality of the feature tone retrieval, in the process of recognition, the short-term spectral structure features and envelope features of the melody pitch are extracted based on the correlation function between the obtained note starting point signals based on the image sequence. Based on the feature transformation and fusion of each melody pitch starting point, the intelligent optimization recognition of note starting

point is completed. The flowchart of image sequence feature extraction is shown in Figure 7.

According to Figure 7, firstly, the input multimodal music audio signal is prefiltered to convert the input analog audio into a digital audio signal within the sound frequency range that can be received by the human ear. Secondly, according to the short-time stability of the audio signal, the preweighted audio signal is processed into frames, and the Hamming window is used to window the signal of each frame to reduce the influence of Gibbs effect. The short-time Fourier transform converts the time domain signal into the frequency domain signal, which is convenient for the triangular window filtering of the subsequent Mel filter. Then, the logarithm of the filtered signal is taken, and the discrete cosine transform is carried out to remove the correlation between the signals of various dimensions, and the signal is mapped to the low dimensional space. Finally, the Mel cepstrum coefficient is obtained by spectral weighting, cepstrum mean subtraction, and difference processing. Because the lower order parameters of cepstrum are easily affected by the characteristics of speaker and channel, the recognition ability is improved.

The specific steps of intelligent optimization identification are detailed as follows.

Assuming that $\partial(o)$ represents the smoothing parameters of the pitch trajectory, based on the obtained $R'_{\text{cross}}(t)$, the short-time spectral structure features of the extracted humming melody pitch represented by BN and the envelope features represented by MFCC are extracted by the following formulae:

$$\text{BN} = \frac{R'_{\text{cross}}(t) \times \partial(o)}{\varpi_{(j)} \times \varepsilon(h)}, \tag{17}$$

$$\text{MFCC} = \frac{R'_{\text{cross}}(t) \times \partial(o)}{\varpi_{(j)} \times BN}. \tag{18}$$

In the formula, $\varpi_{(j)}$ represents the number of starting points of humming notes and $\varepsilon(h)$ represents the offset vector.

A set of transformation matrices for the starting points of humming melody pitch is obtained by discrimination training. Based on the image sequence, each transformation matrix in the set corresponds to a region in the feature space division of the starting points of humming notes, which is transformed with the transformation matrix corresponding to the region to which the feature vector belongs. It is assumed that $o(t)$ represents the input feature of time t, A_i represents the transformation matrix corresponding to the i domain, and the characteristic transformation of the s melody pitch segment is described by the following formula:

$$o'_s(t) = R \sum_{i=1}^{A_i} S \times o(t) \times x_{i,s}. \tag{19}$$

In the formula, R represents the starting paragraph of melody pitch after domain division and $x_{i,s}$ represents the weight coefficient corresponding to the selected feature transformation matrix A_i.

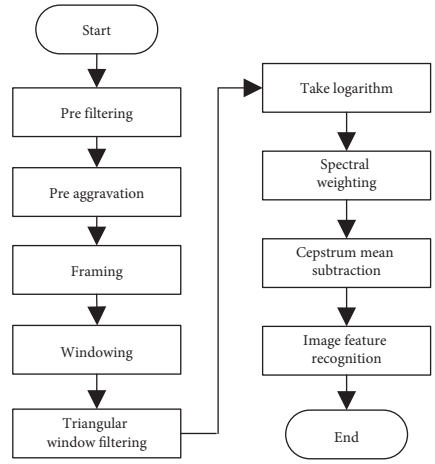

Figure 7: Flowchart of image sequence feature extraction.

Assuming that h represents the excitation signal of the BN layer humming melody pitch node, the transformation matrix features represented by M^{BN} and M^{RDLT} are fused by the following formula:

$$y_{\text{con}}(t) = \frac{\left[M^{\text{BN}}, M^{\text{RDLT}}\right]}{h \times xo(t)}. \tag{20}$$

In the formula, $xo(t)$ represents the regularization function.

Assuming that the estimated value of β noise spectrum is used, the parameters of the fused transformation matrix feature $y_{\text{con}}(t)$ are optimized by the following formula:

$$H_{\text{MPE}}(Y_{\text{con}}) = \frac{y_{\text{con}}(t) \times M^{\text{BN}} \times M^{\text{RDLT}}}{M^O}. \tag{21}$$

In the formula, M^O represents the transformation matrix corresponding to the nonzero coefficient term. Based on the results calculated by formula (21), the intelligent recognition of note starting point in feature tone retrieval can be effectively completed, so as to complete the research of multimodal music emotion recognition method based on image sequence.

3. Experimental Analysis

In order to test the application effect of the multimodal music emotion recognition method based on image sequence, MATLAB software is used as the algorithm operation platform, and a specific example is selected for simulation test and analysis. The experimental environment settings are shown in Table 2.

The samples selected in the test are from the emotional corpus. According to the selected samples and the emotions to be expressed, they are divided into five categories. The specific distribution of samples is shown in Table 3.

Kappa coefficient is selected as the index to evaluate the intelligent recognition and classification of music emotion. Kappa coefficient is used for consistency test and classification accuracy. Its calculation formula is as follows.

Table 2: Experimental parameter setting.

Parameter	Numerical value
Node	10
CPU	2
Core frequency	1.9 GHz
Memory	8 GB

Table 3: Distribution of test samples.

Emotion type	Music clip name	Characteristic dimension	Sample properties
Happy	Red head rope fragment	455	Training sample
	Love in the rain	355	Training sample
	Carmen fragment	784	Test sample
	Trout fragment	232	Test sample
Sadness	Schindler list theme song clip	534	Training sample
	Liang Zhu fragment	454	Training sample
	Pathetique fragment	234	Test sample
	Parting fragment	545	Test sample
Tender	Lullaby fragment	215	Training sample
	Blue Danube segment	313	Training sample
	Little star clip	534	Test sample
	To Alice	341	Test sample
Anger	Polish dance pieces	132	Training sample
	Destiny fragment 1	431	Training sample
	International song clip	453	Test sample
	Empty madness	315	Test sample
Fear	Gloomy Sunday clip	341	Training sample
	Ghost call clip	345	Training sample
	Thirteen pairs of eyes	422	Test sample
	Step by step press clip	244	Test sample

Table 4: Recognition results of some music clips.

Music clip name	Emotion type	Paper method	Reference [11] method	Reference [12] method
Little star clip	Tender	Tender	Sadness	Sadness
Blue Danube segment	Tender	Tender	Sadness	Happy
Liang Zhu fragment	Sadness	Sadness	Tender	Tender
Love in the rain	Happy	Happy	Tender	Tender
Trout fragment	Happy	Happy	Happy	Tender
Ghost call film	Fear	Fear	Fear	Sadness
Segment 2	Sadness	Sadness	Sadness	Fear

$$k = \frac{p_o - p_e}{1 - p_e}. \tag{22}$$

In the formula, p_o is the observation consistency rate and p_e represents the expected consistency rate. The larger the values of $k \in [-1, 1]$ and the larger the k value, the more consistent the two results. When $k \geq 0.75$, the results are consistent and the classification recognition is more accurate. If $k < 0.4$, it indicates lack of consistency and poor classification and recognition accuracy.

Input the test samples in Table 3 into the trained neural network model, test the samples, count the sample test results, and calculate the kappa coefficient. The results are as follows:

$$\begin{aligned} k_{\text{Happy}} &= 0.855, \\ k_{\text{Sadness}} &= 0.870, \\ k_{\text{Tender}} &= 0.912, \\ k_{\text{Anger}} &= 0.825, \\ k_{\text{Fear}} &= 0.811. \end{aligned} \tag{23}$$

The kappa coefficient k values calculated above are greater than 0.75, indicating that the recognition and classification results are in good agreement with the actual results, and the classification and recognition accuracy is high, which has achieved the research purpose. The

multimodal music emotion recognition method is used to identify the music fragments in the test set, and some test samples and their discrimination results are intercepted, as shown in Table 4.

It can be seen from Table 4 that the identification results of the same test sample by different methods are different. Rhythm and melody characteristics have a great influence on the recognition of music emotion. On the premise that the image sequence is unchanged, selecting the appropriate music feature input vector will improve the accuracy of multimodal music emotion recognition to a certain extent.

4. Conclusion and Prospect

4.1. Conclusion. Multimodal musical emotion is a breakthrough in the field of artificial intelligence. It has become a new research feature of computer science, cognitive science, neuroscience, brain science, psychology, behavioral science, and other interdisciplinary fields. Multimodal musical emotion understanding is an important branch of emotion computing and has a broad development prospect. The multimodal music emotion recognition method based on image sequence verifies the performance of the algorithm through an example. The kappa coefficient proves that the classification recognition accuracy of the algorithm is high, which achieves the research goal. Meanwhile, the rhythm and melody characteristics have a great influence on the recognition of music emotion.

4.2. Prospect. The possible future research direction is to apply deep learning method to music emotion recognition. Deep learning is a kind of based on feature hierarchical structure, characteristics of unsupervised learning learning method, has a lot of the hidden layer of all the excellent characteristics of artificial neural network learning ability, learning and to the characteristics of the characterization of the nature of the data more through millions of music is used to study characteristics. Thus, let the machine independently choose better music features to describe the relationship between the music and the emotion.

Data Availability

The raw data supporting the conclusions of this article will be made available by the author, without undue reservation.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding this work.

References

[1] I. Peretz, J. Ayotte, R. J. Zatorre et al., "Effects of vocal training in a musicophile with congenital amusia," *Neuron*, vol. 33, no. 2, pp. 185–191, 2020.

[2] A. P. Montgomery, A. Mousavi, M. Carbonaro, and D. Hayward, "Using learning analytics to explore self-regulated learning in flipped blended learning music teacher education," *British Journal of Educational Technology*, vol. 9, no. 10, pp. 114–127, 2019.

[3] S. Hawkins, "Situational influences on rhythmicity in speech, music, and their interaction," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 369, no. 1658, Article ID 20130398, 2019.

[4] M. Bigliassi, C. I. Karageorghis, G. K. Hoy, and G. S. Layne, "The Way You Make Me Feel: Psychological and cerebral responses to music during real-life physical activity," *Psychology of Sport and Exercise*, vol. 41, no. 3, pp. 211–217, 2019.

[5] L. Bochen, X. Liu, K. Dinesh, and Z. Duan, "Creating a Multitrack classical music performance Dataset for multimodal music analysis: Challenges, Insights, and applications," *IEEE Transactions on Multimedia*, vol. 12, no. 14, pp. 159–164, 2019.

[6] S. Nag, S. Sanyal, A. Banerjee, R. Sengupta, and D. Ghosh, "Music of brain and music on brain: a Novel EEG Sonification approach," *Cognitive Neurodynamics*, vol. 13, no. 4, pp. 13–31, 2019.

[7] X. Wang, G. Soumitra, and G. Sun-Wei, "Quantitative quality control in microarray image processing and data acquisition," *Nucleic acids research*, vol. 29, no. 15, pp. 75–80, 2019.

[8] L. Reichel and U. O. Ugwu, "Tensor Krylov subspace methods with an invertible linear transform product applied to image processing," *Applied Numerical Mathematics*, vol. 166, no. 8, pp. 186–207, 2021.

[9] W. Z. Liang, I. Possignolo, X. Qiao, and K. DeJonge, "Utilizing digital image processing and two-source energy balance model for the estimation of evapotranspiration of dry edible beans in western Nebraska," *Irrigation Science*, vol. 204, no. 39, pp. 617–631, 2021.

[10] M. Talaat, M. Tayseer, and A. El-Zein, "Digital image processing for physical basis analysis of electrical failure forecasting in XLPE power cables based on field simulation using finite-element method," *IET Generation, Transmission & Distribution*, vol. 14, no. 26, pp. 6703–6714, 2020.

[11] A. M. Proverbio, F. Benedetto, and M. Guazzone, "Shared neural mechanisms for processing emotions in music and vocalizations," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1987–2007, 2020.

[12] H. Platel, B. Jean-Claude, B. . Desgranges, F. Bernard, and F. Eustache, "Semantic and episodic memory of music are subserved by distinct neural networks," *NeuroImage*, vol. 20, no. 1, pp. 244–256, 2019.

[13] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology an International Journal*, vol. 24, no. 3, pp. 760–767, 2020.

[14] K. W. Cheuk, Y. J. Luo, B. Balamurali, and G. Roig, "Regression-based music emotion prediction using triplet neural networks," in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 21, no. 1, pp. 15–21, 2020.

[15] S. Chowdhury, V. Praher, and G. Widmer, "Tracing back music emotion predictions to sound Sources and Intuitive perceptual Qualities," vol. 14, no. 6, pp. 45–52, 2021.

[16] D. . Zheng, "Music emotion recognition classification algorithm based on forward neural network," *Information & Technology*, vol. 43, no. 12, pp. 57–61, 2019.

[17] X. Tang, C. X. Zhang, and L. I. Jiang-Feng, "Music emotion recognition based on deep learning," *Computer Knowledge and Technology*, vol. 15, no. 11, pp. 232–237, 2019.

[18] F. Pan, L. Zhang, Y. Ou, and X. Zhang, "The audio-visual integration effect on music emotion: behavioral and

physiological evidence," *PLoS One*, vol. 14, no. 5, Article ID 0217040, 2019.

[19] T. L. Nguyen, B. L. Trieu, Y. Hiraguri, M. Morinaga, T. Morihara, and T. Yano, "Effects of changes in acoustic and non-acoustic factors on Public Health and Reactions: Follow-up Surveys in the Vicinity of the Hanoi Noi Bai International Airport," *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, pp. 2597–2530, 2020.

[20] S. H. Park and P. J. Lee, "Reaction to floor impact noise in multi-storey residential buildings: the effects of acoustic and non-acoustic factors," *Applied Acoustics*, vol. 150, no. 7, pp. 268–278, 2019.

[21] J. L. Tian and S. N. University, "Dynamic Visualization method of music melody based on MATLAB," *Modern Computer*, no. 33, pp. 3–6+10, 2019.

[22] I. Goienetxea, I. Mendialdua, I. Rodríguez, and B. Sierra, "Statistics-based music generation approach Considering Both rhythm and melody Coherence," *IEEE Access*, vol. 7, no. 2, Article ID 183365, 2019.

[23] Y. Yu and S. Canales, "Conditional LSTM-GAN for melody generation from Lyrics," *Computer Science*, vol. 15, no. 8, pp. 26–35, 2019.

[24] M. Farzaneh and R. Mahdian Toroghi, "Music generation using an interactive Evolutionary algorithm," *Pattern Recognition and Artificial Intelligence*, vol. 18, no. 12, pp. 207–217, 2019.

[25] S. Swaminathan and E. G. Schellenberg, "Musical ability, music training, and language ability in childhood," *Journal of Experimental Psychology Learning Memory and Cognition*, vol. 46, no. 12, pp. 2340–2348, 2019.

[26] D. Herremans and T. Bergmans, "Hit Song prediction based on early Adopter data and audio features," *Sound*, vol. 16, no. 10, pp. 148–153, 2020.

[27] B. Gong, M. Kaya, and N. Tintarev, "Contextual Personalized Re-Ranking of music recommendations through audio features," *Information Retrieval*, vol. 6, no. 9, pp. 89–95, 2020.

[28] A. Kmb, A. Tb, A. Ds, and Z. Zhao, "Contributions of MIR to soundscape ecology. Part 3: Tagging and classifying audio features using a multi-labeling k -nearest neighbor approach," *Ecological Informatics*, vol. 51, no. 5, pp. 103–111, 2019.

[29] Z. Wang, L. Wang, and H. Huang, "Joint low rank embedded multiple features learning for audio-visual emotion recognition," *Neurocomputing*, vol. 388, no. 5, pp. 324–333, 2020.

[30] X. Guo, W. Zhong, L. Ye, L. Fang, and Q. Zhang, "Affective Video content analysis based on two Compact audio-visual features," *Communications in Computer and Information Science*, vol. 16, no. 2, pp. 355–364, 2020.

[31] E. Han and H. Cha, "Audio feature extraction for effective emotion classification," *IEIE Transactions on Smart Processing & Computing*, vol. 8, no. 4, pp. 100–107, 2019.

[32] Z. H. Wang, G. J. Horng, T. H. Hsu, and A. Aripriharta, "Heart sound signal recovery based on time series signal prediction using a recurrent neural network in the long short-term memory model," *The Journal of Supercomputing*, vol. 76, no. 1, pp. 8373–8390, 2019.

[33] F. Huang, J. Zeng, Y. Zhang, and W. Xu, "Convolutional recurrent neural networks with multi-sized convolution filters for sound-event recognition," *Modern Physics Letters B*, vol. 34, no. 23, Article ID 2050235, 2020.

[34] R. Haeb-Umbach, S. Watanabe, T. Nakatani et al., "Speech processing for digital Home Assistants: Combining signal processing with deep-learning Techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[35] R. S. Da, Özbey, Muzaffer, Çatlı, and A. Burak, "A transfer-learning approach for Accelerated MRI using deep neural networks," *Magnetic Resonance in Medicine*, vol. 84, no. 3, pp. 663–685, 2020, https://pubmed.ncbi.nlm.nih.gov/?term=%C3%87ukur+T&cauthor_id=31898840.

[36] Y. Zhang, T. S. Lee, M. Li, F. Liu, and S. Tang, "Convolutional neural network models of V1 responses to complex patterns," *Journal of Computational Neuroscience*, vol. 46, no. 1, pp. 33–54, 2019.