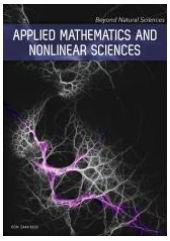
sciendo Applied Mathematics and Nonlinear Sciences, 9(1) (2024) 1-15



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Application and Effectiveness Analysis of Multimodal Emotion Recognition Technology in Music Education

Bing Yan[1,†]

1. Xuzhou Kindergarten Teachers College, Xuzhou, Jiangsu, 221000, China.

Submission Info

Communicated by Z. Sabir
Received May 4, 2024
Accepted August 15, 2024
Available online October 4, 2024

Abstract

Emotions in music education affect learners' cognitive activities, and failure to capture learners' emotional changes in a timely manner can lead to music teachers' inability to adjust their teaching strategies in a timely manner. In this paper, a convolutional neural network is utilized to extract speech and visual emotion features of students during the process of music education. The spatial plane fusion method is used to fuse the speech and visual emotion modalities. A cross-modal interactive attention mechanism is introduced to optimize the fusion effect of the multimodal emotion features. Then, a support vector machine is used to identify and classify the emotion features. The study shows that the multimodal emotion recognition model proposed in this paper can achieve an emotion recognition accuracy of 88.78%, can accurately recognize the emotional state of students, and can assist teachers in effectively intervening in the negative emotions of students. In the music classroom applying this technology, the average test score of the student's music education program is 93.70, and their will to learn music education is 95.09% on average. This paper's multimodal emotion recognition model helps teachers implement effective interventions in music education and establishes the foundation for improving students' interest in music learning.

Keywords: Feature extraction; Feature fusion; Interactive attention; Support vector machine; Music education.
AMS 2010 codes: 00A35

†Corresponding author.
Email address: yanbing1472582024@163.com

ISSN 2444-8656
https://doi.org/10.2478/amns-2024-2716

1 Introduction

Social-emotional competence is crucial to a person's growth and development [1]. CASEL, based on more than two decades of research and practical experience, states that the best practice for social-emotional competence is classroom teaching [2]. Music classroom teaching has the universal value and significance of ordinary classroom teaching [3]. At the same time, music is a special subject that involves the cultivation of students' creative ability, cooperative ability, and aesthetic ability [4]. The process of music education is the process of cultivating social-emotional ability, and music education is in a unique position to promote the development of social-emotional ability.

Multimodal theory can be traced back to the multimodal discourse analysis that emerged in the West in the 1990s [5]. Under the influence of the international trend of rapid development of modern information technology, the multimodal model of teaching and learning emerged and was applied in different professional education fields. The mode focuses on the form of acceptance and perception, which can achieve comprehensive cultivation of students' skills and abilities by actively mobilizing their multimodal cognition and is conducive to injecting new vitality into modern education [6-7].

As a new teaching concept and mode, multimodality is still in the early stage of development when applied to the teaching of musicology courses. In the reform of higher education, the reform of the college curriculum is one of the important contents and requirements of connotative development [8]. Exploring and integrating information technology in the curriculum has increasingly become a hot topic in the reform of higher education. Try to apply the multimodal teaching mode will be the traditional musicology professional course group combined with the characteristics of the times to be an all-round reform on the basis of active exploration, to promote the diversity of practice, and further applied to teaching practice, applied to the teaching system, deepen the curriculum reform, will enhance the quality of music education in colleges and universities will have a positive impact on the quality of teaching [9-10].

Papadogianni, M et al. conducted a teaching practice to evaluate the bimodal auditory-tactile music teaching methodology, noting that this multimodal music teaching model promotes students' perceptual ability to perceive beats [11]. Hoyos, A. A. C et al. contributed to the development of research on instructional analytics by discussing key points of instructional analytics, including the integration of teaching-learning analytics, multimodal analytics, and the analysis of teachers' numerical competencies [12]. Wang, C. H. et al. developed an effective design tutoring model that has positive effects on student motivation and learning outcomes [13]. Martin-Gutierrez, D. et al. built the SpotGenTrack popular dataset to enable efficient retrieval of music-related data and designed an innovative multimodal end-to-end deep learning architecture to evaluate the recognition of music recordings, which was confirmed based on simulation-based experiments [14]. Roberto, M. M et al. attempted to model teachers' use of instructional space using the Moodoo analytical framework, which helps to understand how teachers incorporate spatial elements into instructional strategies and instructional design [15]. Liu, M introduced the concept and significance of deep learning and concluded that music learning under the combination of this deep learning concept is beneficial to the improvement of students' music literacy and music comprehensive ability [16]. Ma, X. designed a multimedia-assisted music teaching model based on AI technology, and after feedback from teaching experiments, it was learned that this AI technology-enabled assisted music teaching model improved the quality of modern music teaching to a certain extent [17]. Tong, G. et al. conceptualized a multimodal music emotion recognition strategy with knowledge sublimation and music style migration as the core logic, which was verified to have the ability of accurate music emotion recognition through simulation experiments [18].

In this paper, a convolutional neural network is used to extract emotional features from students' speech during music education teaching, and then two features, lens length, and motion, are extracted as visual emotional features. The interaction features in the speech affective features and visual affective features are extracted by the cross-modal interaction attention module, and then subsequently, supervised and unsupervised latent spatial plane fusion methods compute the affective feature values of the two modalities. Class labels are also used to optimize the intra-class tightness and inter-class separability of the fused emotional feature vectors, which improves the accuracy of the multimodal emotional feature recognition model. Finally, the support vector machine method is used to recognize and classify students' emotions in music education, and the specific process of its application in music education is described. To analyze the effectiveness of the constructed multimodal emotion recognition method, validation experiments are conducted in the dataset. The model is then applied to the practice of music education to identify students' emotions in the classroom and assist teachers in learning interventions. The effectiveness of the model is explored through comparative analysis.

2 Multimodal Emotion Recognition in Music Education

2.1 Multimodal Emotion Feature Extraction Methods

2.1.1 Audio Emotion Feature Extraction

In the speech feature extraction of music education teaching video [19], the Mel frequency cepstrum coefficient (MFCC) is first used to extract the initial speech features. The main principle of MFCC speech feature extraction is that the spectrum of sound can reflect the relationship between sound frequency and energy, the peaks in the spectrum contain the discriminative information of the sound, and the cepstrum analysis can find out these peaks and their change process. When extracting MFCC features, the window size is set to 20ms, the moving step is set to 10ms, and the output dimension is set to 25.

MFCC features contain speech recognition information but cannot further extract advanced semantic features, so MFCC features are not suitable for direct fusion with face features. Therefore, on the basis of MFCC feature extraction, the convolutional neural network is utilized to further extract its features. In the article, an audio feature extractor based on a convolutional neural network is designed for further extraction of advanced semantic features. Audio modal features using feature preprocessing are used as input, denoted as X_A. The features are first put through a convolutional operation to extract the local features of neighboring audio elements. After that, max-pooling is used to perform downsampling to remove redundant information:

$$\hat{X}_A = Conv1D\left(X_A, k_A\right) \tag{1}$$

$$\hat{X}_A = Dropout\left(BN\left(MaxPool\left(\hat{X}_A\right)\right)\right) \tag{2}$$

Where k_A is the size of the convolution kernel for the audio modality and $\hat{X}_A$ denotes the learned semantic features. Next, the learned features are then fed into the 1D temporal convolution to obtain the higher-order semantic features of the audio. The formula is as follows:

$$\hat{X}_A = BN\left(\mathrm{Re}\,LU\left(conv1D\left(\hat{X}_A, k_A\right)\right)\right) \tag{3}$$

2.1.2 Visual Emotion Feature Extraction

The two visual features extracted in the Visual Feature Extraction in Music Instructional Videos [20] module are shot length and motion.

1) Lens length is defined as the frames and images captured continuously by the camera without significant color changes. To obtain the lens length, the frames are displayed in HSV color space such that F^t is the frame at moment t, and then the color histograms for the three channels of color h_H, saturation h_S, and luminance h_V are calculated from each frame. For the frame of moment $t > N$, the feature matrix of construction X^t is:

$$X^t = \begin{bmatrix} x^t \\ x^{t-1} \\ \vdots \\ x^{t-N+1} \end{bmatrix} \tag{4}$$

Where $t = N, \cdots, T$ is the window length and the number of windows for all frames, respectively. Then, the X^t matrix is decomposed using singular value decomposition (SVD). $s_1, s_2 \cdots, s_N$ is the eigenvalue and s_1 is its maximum value. The rank of X^t is determined by multiplying the number of eigenvalues greater than the threshold τ by s_1. In other words. X^t the rank r^t is the count of those s_i where $s_i / s_1 > \tau$. This computed rank has two important properties: first, if $r^t > r^{(t-1)}$, then the image content of the current frame is very different from the previous one. Second, if $r^t < r^{(t-1)}$, then the image content is stable enough to cover the previous frame. Therefore, it can be concluded that the frame with the highest rank is the beginning of a shot frame. Similarly, frames $r^t > r^{(t-1)}$ and $r^{(t-1)} = 1$ are the last frames of that shot, and thus, the difference between these two frames indicates the length of the shot.

2) Movement. Psychological and physiological studies have shown a dichotomous correlation between the intensity of the emotion evoked in the observer and the motion observed in the video. In this study, motion vectors were used according to the Block Matching Algorithm (BMA). In this paper, 16 pixels are considered as the macroblock size pixel $p = 7$. The output of the cost function is used for matching between two macroblocks. The most suitable macroblock is the one that is most compatible with the current macroblock. Mean Square Error (MSE) and Mean Absolute Difference (MAD) are the most commonly used cost functions in the algorithm:

$$MSE = \frac{1}{S^2}\sum_{i=0}^{S-1}\sum_{j=0}^{S-1}\left(C_{ij} - R_{ij}\right)^2 \tag{5}$$

$$MAD = \frac{1}{S^2}\sum_{i=0}^{S-1}\sum_{j=0}^{S-1}\left|C_{ij} - R_{ij}\right| \tag{6}$$

Where S is the macroblock size, C_{ij} and R_{ij} are the pixel points compared in the current macroblock and the reference macroblock. Peak Signal Noise Ratio (PSNR) represents the property of the motion frame generated from the motion vector and the reference frame macroblock:

$$PSNR = 10\lg\left[\frac{MAX_I^2}{MSE}\right] \tag{7}$$

Where MAX_I is the maximum value of the pixel in the image. In this paper, the Exhaustive Search (EX) algorithm is used to obtain the motion vectors. This algorithm is able to find the optimal match with the highest PSNR among all pattern-matching algorithms.

2.2 Multimodal Emotional Feature Fusion Approach

2.2.1 Cross-modal interaction attention module

The visual modality has rich facial expression information, while the audio modality carries feature information corresponding to facial expression features. The higher-order feature representations of the audio and video modalities are obtained previously by the audio feature extraction module and the video feature extraction module respectively $\hat{X}_A = \left(x_a^l\right)_{l=1}^L$, $\hat{X}_A = \left(x_v^l\right)_{l=1}^L$, L denotes the number of subsequences of the video sequence X, and x_a^l and x_v^l denote the audio and video feature vectors of the lth subsequence of the video sequence X respectively.

In order to better obtain the interaction features of audio and video modalities in music teaching videos and extract the joint representation of audio and video features, this paper constructs a cross-modal interaction attention module [21] for extracting the interaction features between audio and video modalities. The inter-modal weights of sub-sequence audio and video features are computed from a given video sequence, respectively, for obtaining the correlation between audio and video modalities. The inter-correlation of A and V features is computed through a learnable weight matrix W:

$$Z = \hat{X}_A^T W \hat{X}_V \tag{8}$$

Where $Z \in R^{L\times L}, W \in R^{K\times K}$ denotes the interaction weight between the audio and video features and K denotes the feature dimension of the audio and video features. The inter-correlation matrix Z gives a measure of the correlation between audio and video features, and a larger correlation coefficient in matrix Z indicates a higher correlation between the corresponding audio and video sequence features. Here, the weights W are obtained based on the inter-correlation learning of audio and video features, and the attention weights of each modality are learned interactively based on the other modality, which can effectively utilize the complementarity between different modalities.

2.2.2 Affective Feature Fusion Recognition Methods

In the feature fusion module [22], both supervised and unsupervised latent spatial plane fusion methods are used in this paper. Specifically, CCA and CFA are used as unsupervised methods, and MFA is used as a supervised method to analyze the features of audio and video. To apply CCA, the

average feature vector of the audio and video signals is set to (a,v) and the value of (a,v) is made zero as shown in (9):

$$(a,v)=\left\{\left(a_1,v_1\right),\left(a_2,v_2\right),\cdots,\left(a_n,v_n\right)\right\} \tag{9}$$

a_i and v_i are the original features extracted from the modalities in s and t dimensions, respectively, and n is the number of samples (i.e., the number of frames). The goal of the CCA is to form two transformation matrices W_a and W_v with dimensions $s\times r$ and $t\times r$, where $r\le\min(s,t)$. Maps the original features of the audio and video modalities into the latent spatial plane through W_a and W_v, maximizing the correlation between $\hat{a}=aW_a$ and $\hat{v}=vW_v$. The correlation coefficient ρ between the corresponding projected feature vectors $\hat{a}$ and $\hat{v}$ is maximized as follows:

$$\begin{aligned}\rho&=\max_{W_a,W_v}\frac{E\left[\hat{a}^T\hat{v}\right]}{\sqrt{E\left[\hat{a}^2\right]E\left[\hat{v}^2\right]}}\\&=\max_{W_a,W_v}\frac{E\left[W_a^Ta^TvW_v\right]}{\sqrt{E\left[W_a^Ta^TaW_a\right]E\left[W_v^Tv^TvW_v\right]}}\\&=\max_{W_a,W_v}\frac{{W_a}^TC_{av}W_v}{\sqrt{{W_a}^TC_{aa}W_a{W_v}^TC_{vv}W_v}}\end{aligned} \tag{10}$$

Where C_{av},C_{aa},C_w is the cross-covariance matrix of (a,v), the covariance matrix of a, and the covariance matrix of v, respectively. Eq. (11) approximates an eigenvalue problem:

$$C_{aa}^{-1}C_{av}C_{vv}^{-1}C_{va}W_a=\rho^2W_a \tag{11}$$

$$C_{vv}^{-1}C_{va}C_{aa}^{-1}C_{av}W_v=\rho^2W_v \tag{12}$$

For the feature-level fusion module of the proposed framework, CFA can also be used. When applying the CFA method using two transformation matrices W_a and W_v to (a,v), the following criterion needs to be minimized:

$$\min_{W_a,W_v}=\left\|aW_a-vW_v\right\|_F^2 \tag{13}$$

Where $W_a^TW_a$ and $W_v^TW_v$ are unit matrices. Paradigm F is calculated as follows:

$$\left\|W\right\|_F=\sqrt{\sum_{ij}w_{ij}^2} \tag{14}$$

In order to obtain the optimal transformation matrices W_a and W_v, it is necessary to solve Eq. (14). Then, the singular value decomposition is used to decompose the mutual covariance matrix C_{av}:

$$C_{av}=S_{av}\Lambda_{av}D_{av} \tag{15}$$

So there:

$$W_a = S_{av}, W_v = D_{av} \tag{16}$$

Class labeling can be used to generate shared potential space planes more efficiently. The use of class labels enables feature-level fusion methods to create shared spaces, resulting in feature vectors with higher intraclass tightness and interclass separability. In order to improve the performance of the feature-level fusion method, the MFA algorithm is used in this paper.MFA uses class labeling information in the process of generating the potential space plane. In MFA, intra-class tightness is defined as:

$$\begin{aligned} S_C &= \sum_i \sum_{i \in N_{k_1}^+(j)} \left\| W^T x_i - W^T x_j \right\|^2 \\ &= 2w^T X (D - S) X^T w \end{aligned} \tag{17}$$

Where $X = \left[x_1, x_2, \cdots, x_N \right]$ is the set of samples (i.e., frames), N is the number of the sample, $N_{k_1}^+$ is x_i the k_i neighbors in the same class, and S and D are calculated as:

$$S_{ij} = \begin{cases} 1, if \ \ i \in N_{k_1}^+(j) \\ 0, otherwise \end{cases} \tag{18}$$

$$D_{ij} = \sum_j S_{ij} \tag{19}$$

In addition, the concrete expression for interclass separability is:

$$\begin{aligned} S_P &= \sum_i \sum_{(i,j) \in P_{k_2}(c_i)} \left\| W^T x_i - W^T x_j \right\|^2 \\ &= 2w^T X \left(D^P - S^P \right) X^T w \end{aligned} \tag{20}$$

Where c_i is the ind sentiment category. $P_{k_2}(c)$ is the set of k_2 closest matches. The formula for S is:

$$S_{ij}^P = \begin{cases} 1, if \ (i, j) \in P_{k_2}\left(c_i\right) \\ 0, otherwise \end{cases} \tag{21}$$

The objective function can now be expressed as:

$$\hat{w} = \arg\min_w \frac{W^T X (D - S) X^T w}{W^T X \left(D^P - S^P \right) X^T w} \tag{22}$$

By solving the generalized eigenvalue problem, the optimal solution $Y = X^T w$ can be calculated. $Ly = \lambda L^P y$ where $L = D - S$ and $L^P = D^P - S^P$ are the Laplace matrices of W and W^P, respectively. After the multimodal emotional features are fused, this paper carries out emotion recognition and classification of the fused emotional features by means of a support vector machine.

2.3 Music Education with Application of Multimodal Emotion Recognition

Only by promoting the coordinated development of emotion and cognition in the process of music teaching can we cultivate high-quality talents in the new era and achieve the harmonious unity of emotion and cognition in teaching. While fully considering the cognitive factors in teaching, it is also necessary to pay full attention to the emotional factors in teaching and strive to play a positive role in order to improve the objectives of music teaching, improve all aspects of teaching, optimize the teaching effect, and promote the overall development of students' quality. For the classroom teaching scene of music education, teachers and students in the same environment can feel each other's emotions, and the multimodal emotion recognition constructed in this paper is mainly used for the detection and prediction of teachers' and students' emotions and the assessment of the overall atmosphere of the classroom. Firstly, it can track the emotional state of teachers and students during the teaching process of music education classrooms and provide timely feedback to teachers as a reference for dynamic and thus adjust the teaching. Different emotion statistics and classification outputs can be used to analyze the classroom atmosphere, which can also be used as an indicator for teaching evaluation.

3 Results and discussion

3.1 Multimodal emotion recognition performance

3.1.1 Experimental data set

The Ryerson Audiovisual Database of Emotional Speech and Song (RAVDESS) is a multimodal emotion recognition dataset of 1,440 short speech video clips from 24 performers (12 males and 12 females). Eight emotions were included in the dataset: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised.

The English-language audio-visual emotion dataset e NTERFACE'05 contains 1166 video files and gathers emotion data from 42 adults across 14 countries. It includes six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. A 5-fold cross-validation is executed on both of the above datasets, dividing the samples into training and test sets in a ratio of 5:1, and finally, the cross-validation results are averaged as the experimental results.

3.1.2 Experimental setup

This experiment was done on the system platform Ubuntu 18.02, CPU@3.0GHZ, CUDA 11.2, and NVIDIA RTX 2080 to train the whole model. The overall experimental process uses the Adam optimizer. The total epoch is set to 70; that is, the complete training set passes through the network 70 times, and the batch size is set to 64. The model training oscillations are reduced, and it can converge to a better level. The learning rate is 0.001, and as the number of iterations increases, the model loss value continues to decrease, the model parameters will gradually converge to a relatively stable state, and the final accuracy rate is the average accuracy rate of 5-fold cross-validation. In the training phase, the face images in the dataset are normalized so that the range is restricted to between [-1,1] to ensure that the training data have the same distribution, thus accelerating the convergence speed of the model. In the model containing dropout, setting p to 0.2, i.e., each node does not work with a 20% probability, can effectively prevent the model from generating overfitting problems.

3.1.3 Experimental comparison results

The experiments in this section verify the effectiveness of the proposed multimodal emotion recognition model, using the RAVDESS and e NTERFACE'05 datasets as experimental objects for data preprocessing, feature extraction, and feature fusion of videos. Comparing the model of this paper with several latest multimodal fusion algorithms, the comparison results of the emotion recognition accuracy of this paper's method with other fusion algorithms in the RAVDESS dataset are shown in Table 1, and the comparison results of the emotion recognition accuracy in the e NTERFACE'05 dataset are shown in Table 2. The RAVDESS dataset shows that this paper's model has a model recognition accuracy of 88.78% with a total of 25.41M parameters. The NTERFACE'05 dataset has reached 89.26%. Compared to the unimodal Multiplicative model, the recognition accuracy of the emotion recognition model in this paper is about 24% higher. This indicates that the multimodal fusion method makes the emotion recognition model get more emotional feature inputs, and the feature information among multiple modalities complements each other, which helps the network to obtain, and verifies the importance of the multimodal fusion method. In addition, the model in this paper brings significant performance improvement compared to Multiplicative and Multiplication models, although it increases 1.76-2M parameters. Furthermore, the proposed emotion recognition model enhances CFN-SR's accuracy by 10.68-12.32% compared to the CFN-SR with superior performance. The results above suggest that the multimodal emotion fusion model presented in this paper is superior for recognizing character emotions in videos and can be utilized in music education.

Table 1. Analysis of emotional recognition Accuracy (RAVDESS)

Model	Fusion Stage	Accuracy	#Params
Multiplicative	Late	65.51%	23.65M
Multiplication	Late	65.7%	23.65M
Concat + FC	Early	69.52%	27.58M
MCBP	Early	71.36%	49.52M
MMTM	Model	72.08%	32.56M
CFN-SR	Model	78.10%	26.94M
This algorithm	Model	88.78%	25.41M

Table 2. Analysis of emotional recognition accuracy (e NTERFACE'05)

Model	Fusion Stage	Accuracy	#Params
Multiplicative	Late	65.92%	24.59M
Multiplication	Late	66.58%	24.59M
Concat + FC	Early	70.42%	27.84M
MCBP	Early	71.52%	50.42M
MMTM	Model	72.69%	33.52M
CFN-SR	Model	76.94%	26.85M
This algorithm	Model	89.26%	26.95M

3.2 Analysis of the effect of applying the emotion recognition model

Students' cognitive activities in the learning process of music education are affected by their emotions, and studies have shown that positive emotions can enhance students' learning efficiency in music

education, while negative emotions will restrict the benign development of music learning. In this study, the multimodal emotion recognition model was applied to the teaching of music education, and the model was used to recognize the emotions of each student during the process of music teaching. Teachers can improve the learning outcomes of music learners by providing instructional interventions that are based on each student's emotional state.

3.2.1 Experimental design for teaching music

1) Experimental subjects

The experimental subjects selected for this study were students from School H, all of whom had studied the basic curriculum of music education during normal music teaching activities. The students were divided into an experimental group (Group A) and a control group (Group B). Both groups consisted of 5 students. Group A used the multimodal emotion recognition model to monitor the learning process during music teaching and learning. The teacher took certain learning interventions when negative emotions were found to arise in the subjects. Group B used the multimodal emotion recognition model to monitor the learning process during music learning but did not take any interventions for the subjects throughout the process. The Multimodal Emotion Recognition Model was used to monitor the students in Group B while they were learning music, but no interventions were taken during the process. After the music education program ended, a follow-up test was administered to both groups. The relevant data will be analyzed at the end of the experiment to determine if the multimodal emotion recognition model's application in music education has an impact on the learning effect.

2) Teaching content

The music education course selected for this study is "Chinese Music History", which is 45 minutes long and involves 6 knowledge points, each of which not only introduces basic concepts but also includes examples and applications. The corresponding moments of each knowledge point are shown in Table 3, including the origin of music, ancient songs and dances, ancient musical instruments, schools of thought and theories of music, musical achievements, and appreciation of collections of works. In music education teaching, when the multimodal emotion recognition model recognizes that students have negative learning emotions, the teacher conducts corresponding learning interventions to improve their learning. At the end of the music education learning program, each student will receive an accompanying quiz.

Table 3. The corresponding moments of the knowledge

Course time	Current knowledge	
00:00~05:45	Knowledge 1	The origin of music
05:46~15:20	Knowledge 2	Ancient songs and ancient music
15:21~25:00	Knowledge 3	Cologne
25:01~31:15	Knowledge 4	Music school of thought and theory
31:15~40:30	Knowledge 5	Music achievement
40:31~45:00	Knowledge 6	Appreciation of the portfolio

3.2.2 Analysis of the effectiveness of music education teaching

In this paper, the teaching experiment will comprehensively analyze the learning effect of music education from three quantitative indexes: learning interest, learning will, and learning achievement. The formula for calculating learners' learning interests is as follows:

$$I = \frac{t_0 + t_1}{T} \tag{23}$$

Where I denotes the learner's interest in learning, t_0 is the total length of time in which the learner's neutral emotion is recognized, t_1 is the total length of time in which the emotion recognition model recognizes the learner's positive emotion, and T is the total length of time in which the course is studied.

The formula for the learner's will to learn is as follows:

$$W = \frac{T - t}{T} \tag{24}$$

Where W denotes the learner's will to learn, T denotes the total duration of music education learning, and t denotes the time when the emotion recognition model did not capture the learner's facial expression.

1) Overall student analysis

Since the multimodal emotion recognition model uses every 10 seconds to record the subjects' learning emotions, the theoretical maximum total number of emotion recognition for each subject is 270, and the time points at which the students' emotions are not recognized are recorded as missing recognition. The statistical results of the data of some of the subject students in Group A are shown in Table 4, and the statistical results of the data of some of the subject students in Group B are shown in Table 5. By comparing the test scores of the students in Group A and Group B at the end of the music education program, it was found that the test scores of five students in Group A and Group B were between 91.48-96.24, and 65.92-72.59, respectively, which shows that the average test scores of the students in Group B were much lower than those of the students in Group A. Compared to the students in Group B, the average test scores of the students in Group A and Group B were much lower. The five students in both Group A and Group B had an average learning interest of 92.80% and 64.56%, respectively. Also, the willingness to learn about music education was much higher among students in group A (95.09%) than among students in group B (67.93%). This suggests that teachers are able to effectively intervene in students' emotional states identified by the multimodal emotion recognition model, which in turn enhances their interest in learning music, their will to learn, and their achievement.

Table 4. Group A students' emotional identification results

Numbering	Student 1	Student 2	Student 3	Student 4	Student 5
Pleasure	19	25	34	35	48
Wonder	0	5	9	5	11
Quietness	217	189	198	206	170
Revulsion	2	5	4	1	3
Get angry	5	0	1	2	0
Be afraid	1	0	0	0	1
Heartbreak	4	2	8	5	6
Deletion	11	22	8	16	31
Total amount	259	248	262	254	239
Study interest	89.52%	92.56%	93.48%	95.62%	94.68%
Learning will	90.52%	95.67%	98.24%	96.24%	94.78%
Test score	95.26	91.48	92.54	96.42	92.82

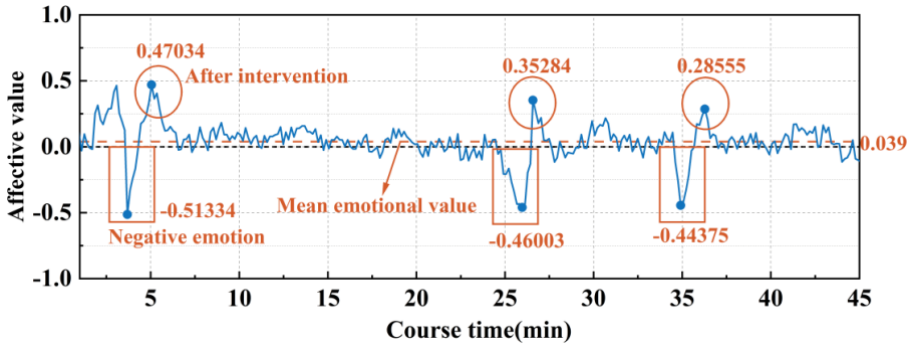
Table 5. Group B students' emotional identification results

Numbering	Student 6	Student 7	Student 8	Student 9	Student 10
Pleasure	11	5	9	10	8
Wonder	1	2	5	4	7
Quietness	198	177	194	159	131
Revulsion	21	34	28	35	36
Get angry	5	2	9	8	9
Be afraid	11	8	2	3	9
Heartbreak	2	8	11	15	18
Deletion	21	34	12	36	52
Total amount	249	236	258	234	218
Study interest	65.29%	62.17%	69.52%	65.23%	60.59%
Learning will	71.52%	68.29%	75.96%	62.39%	61.48%
Test score	69.52	72.59	70.28	65.92	68.42

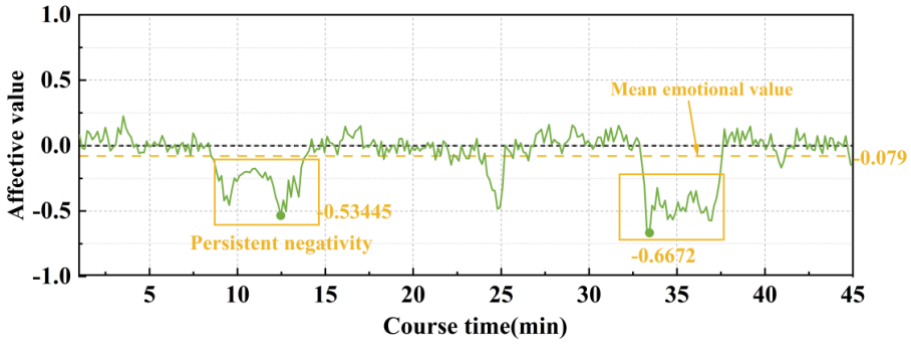
2) Analysis of individual students

This paper analyzes the specific changes in the emotions of Student 1 in Group A and Student 6 in Group B in the music education course, and the results of the emotion analysis of the two students are shown in Figure 1, with (a) and (b) representing the results of the analysis of the changes in the emotions of Student 1 in Group A and Student 2 in Group B, respectively. Student 1's average affective value in the music education program was 0.039, which is positive. At 3.68 minutes into the class, the emotion recognition model recognized that the student had an affective value of -0.513, and the teacher made a timely learning intervention to increase his affective value to 0.470. Similarly, at 25.8 minutes and 34.9 minutes into the class. In contrast, student #6 originated from the control group, so even if negative emotions are recognized, the teacher does not implement the strategy of learning intervention. The student's emotions were consistently negative during the 9.2-13.5 minutes and 33.4-37.0 minutes of the music education classroom, with an average emotion value of only -0.079

throughout the entire music education classroom. Comparing the emotions of these two students, the use of the music education model based on multimodal emotion recognition allows for timely identification of the student's classroom emotions and timely interventions to enhance the students' music classroom effectiveness.



(a) Student 1 (Group A)



(b) Student 6 (Group B)

Figure 1. The emotional change of middle school students in music education

4 Conclusion

This paper constructs a multimodal emotion recognition model to recognize students' emotions in a music education classroom. When there are 25.41M parameters, this paper achieves a model recognition accuracy of 88.78%. The accuracy rate in the e-NTERFACE'05 dataset is 89.26%, which is much higher than the accuracy rate of other emotion recognition models. Teaching comparison experiments show that the emotion recognition model proposed in this paper, applied in music education, is effective in recognizing students' emotional changes. After the teachers carried out the corresponding teaching interventions for the students based on the emotional values recognized by the model in this paper, the average test score of the five students in the experimental group in the music education course was 93.70, while the average test score of the students in the control group was only 69.35. It was discovered that the experimental group students had a higher desire to learn music education (95.09%) than the Group B students (67.93%).

In conclusion, the multimodal emotion recognition model proposed in this paper can effectively capture the emotional state of learners when applied to music education, providing technical support for teachers to intervene in students' learning based on their emotions, and at the same time, the widespread use of this model in colleges and universities can improve the learning effect of students' music education.

References

[1] Bellocchi, A. (2018). Early career science teacher experiences of social bonds and emotion management. Journal of Research in Science Teaching.

[2] Information, Studies, Faculty, of, Humanities, & and, et al. (2017). Development of the instructional model by integrating information literacy in the class learning and teaching processes. Education for Information, 28(2-4), 137-150.

[3] Liang, Y. (2019). Intelligent emotion evaluation method of classroom teaching based on expression recognition. International Journal of Emerging Technologies in Learning (iJET), 14(04).

[4] He, J. (2017). Research on the design and experiment of music teaching based on network technology. Boletin Tecnico/Technical Bulletin, 55(10), 101-107.

[5] Li, H. (2017). Study on the innovation path of music teaching mode in multimedia flipped classroom under the internet background. Revista de la Facultad de Ingenieria, 32(12), 913-919.

[6] Yang, L. (2020). Comprehensive evaluation of music course teaching level based on improved multi-attribute fuzzy evaluation model. International Journal of Emerging Technologies in Learning (iJET)(19).

[7] María Alfaro-Contreras, Valero-Mas, J. J., José M. Iesta, & Calvo-Zaragoza, J. (2023). Late multimodal fusion for image and audio music transcription. Expert Systems with Applications, 216, 119491-.

[8] Gong, W., Yu, Q., Sun, H., Huang, W., Cheng, P., & Gonzalez, J. (2024). Mclemcd: multimodal collaborative learning encoder for enhanced music classification from dances. Multimedia systems(1), 30.

[9] Li, N., Peng, Y., & Fan, J. (2023). Analysis of the application of college popular music education relying on the elite teaching optimization algorithm. Applied Artificial Intelligence.

[10] Song, B. (2024). Multimodal interactive classroom teaching strategies based on social network analysis. International journal of networking and virtual organisations(1), 30.

[11] Papadogianni, M., Altinsoy, E., & Andreopoulou, A. (2024). Multimodal exploration in elementary music classroom. Journal on multimodal user interfaces(1), 18.

[12] Hoyos, A. A. C., & Velasquez, J. D. (2020). Teaching analytics: current challenges and future development. Revista Iberoamericana de Tecnologias del Aprendizaje, PP(99), 1-1.

[13] Wang, C. H., & Lin, H. C. K. (2018). Emotional design tutoring system based on multimodal affective computing techniques. International Journal of Distance Education Technologies, 16(1), 103-117.

[14] Martin-Gutierrez, D., Penaloza, G. H., Belmonte-Hernandez, A., & Alvarez, F. (2020). A multimodal end-to-end deep learning architecture for music popularity prediction. IEEE Access, PP(99), 1-1.

[15] Roberto, M. M., Vanessa, E., Katerina, M., Antonette, S., Gloria, F. N., & Jurgen, S., et al. (2022). Moodoo the tracker: spatial classroom analytics for characterising teachers' pedagogical approaches. International Journal of Artificial Intelligence in Education.

[16] Liu, M. (2021). Research on music teaching and creation based on deep learning. Mobile information systems.

[17] Ma, X. (2021). Analysis on the application of multimedia-assisted music teaching based on ai technology. Advances in multimedia(Pt.1), 2021.

[18] Tong, G. (2022). Multimodal music emotion recognition method based on the combination of knowledge distillation and transfer learning. Scientific Programming.

[19] Boulal Hossam,Hamidi Mohamed,Abarkan Mustapha & Barkani Jamal.(2024).Amazigh CNN speech recognition system based on Mel spectrogram feature extraction method.International Journal of Speech Technology(1),287-296.

[20] Vikas Khullar,Isha Kansal,Jyoti Verma,Rajeev Kumar,Karuna Salgotra & Gurpreet Singh Saini.(2024). Deep trained features extraction and dense layer classification of sensitive and normal documents for robotic vision-based segregation.Paladyn(1),

[21] Chen Zhi,Zou Beiji,Dai Yulan,Zhu Chengzhang,Kong Guilan & Zhang Wensheng.(2023).Medical visual question answering with symmetric interaction attention and cross-modal gating.Biomedical Signal Processing and Control

[22] Lianghong Wu,Yujie Zou,Cili Zuo,Liang Chen,Bowen Zhou & Hongqiang Zhang.(2024).A lightweight white blood cells detection network based on CenterNet and feature fusion modules.Measurement Science and Technology(7).