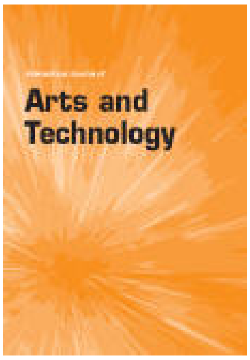


International Journal of Arts and Technology

ISSN online: 1754-8861 - ISSN print: 1754-8853
https://www.inderscience.com/ijart

Multimodal music emotion recognition method based on multi data fusion

Fanguang Zeng

DOI: 10.1504/IJART.2023.10055835

Article History:

Received:	28 December 2022
Last revised:	10 February 2023
Accepted:	01 March 2023
Published online:	28 September 2023

Multimodal music emotion recognition method based on multi data fusion

Fanguang Zeng

Academy of Music,
Pingdingshan University,
Pingdingshan, 467002, China
Email: zfg4425@163.com

Abstract: In order to overcome the problems of low recognition accuracy and long recognition time in traditional multimodal music emotion recognition methods, a multimodal music emotion recognition method based on multiple data fusion is proposed. The multi-modal music emotion is decomposed by the non-negative matrix decomposition method to obtain the multi-modal data of audio and lyrics, and extract the audio modal emotional features and text modal emotional features respectively. After the multi-modal data of the two modal emotional features are weighted and fused through the linear prediction residual, the normalised multi-modal data is used as the training sample and input into the classification model based on support vector machine, so as to identify multimodal music emotion. The experimental results show that the proposed method takes the shortest time for multimodal music emotion recognition and improves the recognition accuracy.

Keywords: multi data fusion; multimodal music; emotional recognition; non-negative matrix decomposition method; support vector machines; SVMs.

Reference to this paper should be made as follows: Zeng, F. (2023) ‘Multimodal music emotion recognition method based on multi data fusion’, *Int. J. Arts and Technology*, Vol. 14, No. 4, pp.271–282.

Biographical notes: Fanguang Zeng is with education background of postgraduate student, and currently is working in Pingdingshan University as an Associate Professor. He is also a member of the Chinese Musicians Association. His main research contents includes vocal performance and teaching, Chinese traditional music. He had got more than ten of his papers published in core journals nationally and participated in different provincial-level projects for three times.

1 Introduction

With the popularisation of information technology and multimedia technology, there are more and more music works. Due to the variety of music, the retrieval efficiency has declined. Music recognition based on emotion can effectively improve the efficiency of music retrieval, and has gradually become a research hotspot. Music emotion recognition belongs to the cross field of music psychology and computer science, and relevant research is of great significance (Zhao et al., 2019; Fan et al., 2020). Music includes two modes: audio and lyric text. For multimodal music, it is inefficient to mark music emotions in a purely manual way, and the quality cannot be guaranteed. It is difficult to meet the emotional annotation requirements of a large number of multimodal music works. Therefore, more and more experts begin to study multimodal music emotion recognition technology (Jian-Wei et al., 2022; Wang et al., 2021).

So far, the research of multimodal music emotion recognition has a history of more than ten years. Many scholars at home and abroad have done in-depth research in this field, and have made certain achievements, such as transfer learning, deep learning, etc. Zhao et al. (2021) studies multimodal music emotion recognition by combining knowledge distillation and transfer learning. By acquiring multimodal music label data, the emotion recognition model is established by combining knowledge distillation and transfer learning, and the acquired multimodal music label data is input as samples into the constructed emotion recognition model for training, and the recognition results are output. Li et al. (2020) carries out modal extraction on the acquired music audio, classifies multimodal music emotion according to the modal extraction results, and establishes multimodal music emotion recognition model by optimising the depth residual network based on the classification results to obtain the recognition results. However, the accuracy of the above two methods for multimodal music emotion recognition is low, resulting in poor recognition effect. Wang et al. (2022) obtains the emotion dictionary, including lyric mode and audio mode, and reduces the dimension of the audio signal by constructing the music emotion model. According to the processing results, the lyric modal feature and audio modal feature are extracted and fused to identify multimodal music emotion. However, the above methods take a long time to identify multimodal music emotion, resulting in low recognition efficiency.

In view of the shortcomings of the above methods, this paper proposes a multimodal music emotion recognition method based on multiple data fusion, which not only solves the problems in traditional methods, but also lays a foundation for digital music resource retrieval. The specific research route of this method is as follows:

1 Multi modal music emotion data acquisition based on matrix decomposition: The audio and lyrics are mapped into the emotional space, and a multi-modal music emotional decomposition matrix is constructed by using the non-negative matrix decomposition method to obtain the multi-modal data of audio mode and lyrics text mode.

2 Emotional feature extraction of multimodal music: After pre operation, emotional features are extracted from the acquired multiple data of audio mode and lyrics text mode. Audio mode emotional features are extracted through pitch, intensity, sound speed, notes and melody, and lyrics modal features are extracted using Doc2Vec.

3 Multi data fusion of emotional characteristics of multimodal music: The linear prediction (LP) residuals are used to fuse the extracted audio modal emotional features and lyrics modal emotional features with multiple data weights, and the fused multi-modal music emotional features' multiple data are normalised.

4 Multimodal music emotion recognition based on classification function: Take the normalised multimodal music emotion feature multivariate data as training samples, calculate the minimum geometric interval from all sample points in the training set to the hyperplane, construct the optimal multimodal music emotion classification function, input the training samples into the optimal multimodal music emotion

classification function, and output the classification results, so as to identify multimodal music emotion.

5 The accuracy and efficiency of multimodal music emotion recognition are used as the evaluation criteria to test the effectiveness of this method.

2 Emotion recognition of multimodal music

2.1 Multi modal music emotion data acquisition based on matrix decomposition

Since multimodal music contains two modes, audio and lyric text, the emotion of multimodal music is decomposed by the method of non-negative matrix decomposition to obtain the multivariate data of the two modes.

Non-negative matrix decomposition [non-negative matrix factorisation (NMF)] technology has become a popular method for data representation. It is to cope with the problem of too high dimension of input data in the real world. It maps the characteristics of two dimensions to a hidden space at the same time (Ni, 2010; Ma et al., 2019; Ferguson, 2022).

Build an audio Lyrics matrix, using NMF to map audio and lyrics to the emotional space at the same time, to obtain the representation or distribution of audio and lyrics in the emotional space. The corresponding data representation is as follows:

The co occurrence information of audio and lyrics is represented as a two-dimensional matrix, where each row represents a song list and each column represents a song. The value in the matrix is 0 or 1, indicating whether the corresponding audio appears in the music (1 if it appears, 0 if it does not appear). Due to the large number of audio and lyrics in music, there will be problems such as excessive resource consumption and time-consuming in data calculation, and excessive frequency of some audio and lyrics will interfere with the decomposition effect. Therefore, it is necessary to preprocess the original music data, select some music as the characteristics of audio and lyrics, and then obtain the original matrix X through matrix characterisation. Assume that the size of matrix X is *m* row *n* column, *m* represents the number of audio, *n* represents the number of lyrics, and the matrix X is as follows:

$$A = \left\{ \begin{matrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{matrix} \right\} \tag{1}$$

NMF aims to find two non-negative decomposition matrices of the original matrix X to replace *U* and *V*, and make the original matrix X as close as possible to the decomposed result UV^T, then the multi-modal music emotional decomposition matrix is:

$$O = \left\| X - UV^T \right\|_F^2 \tag{2}$$

The above decomposition matrix is used to obtain the multivariate data of audio mode and lyrics text mode (Jia et al., 2022).

2.2 Emotional feature extraction of multimodal music

The emotional features of multimodal music are extracted, including audio modal emotional features and lyrics modal emotional features, as shown below.

2.2.1 Audio modal emotional feature extraction

1 Track name

In the process of MIDI music creation, some habitually give names to some audio channels. Generally, those marked ‘MELODIES’, ‘VOCAL’, ‘SING’, ‘SOLO’, ‘LEAD’, and ‘VOICE’ are the track channels where the main melody is located, while those marked ‘ACCU’, ‘DRUM’, ‘BASS’, ‘PERCUSSION’, ‘COMPANION’, and ‘BACK’ do not contain the main melody, and can be directly removed (Lee et al., 2021; Zhang et al., 2021).

2 Channel No

MIDI files typically contain 16 audio channels, one for each channel. Among them, channels 1–9 and 11–16 are the channels of this theme. Channel 10 is a percussion channel, which is generally only used as accompaniment. Therefore, in the pre operation stage, the track of channel 10 can be directly removed.

3 Number of notes

The number of notes in the track where the main melody is located is generally large, and the number of notes in the track where the main melody is located should not be less than half of its average number. Therefore, before extracting audio modal emotional features, the number of notes in the track with half of the average number of notes should be removed.

4 Track length

Pre operation removes redundant tracks. After that, it is necessary to use the theme extraction algorithm for the remaining tracks to determine the theme of the song. This section studies the theme extraction algorithm, finds the MIDI file theme, and prepares for the extraction of audio modal emotional features (Wang et al., 2020).

After pre operation, extract audio modal emotional features through pitch, intensity, sound speed, notes and melody:

1 Average pitch and average intensity of music

In music, a single note cannot effectively express the emotion of the music, so the average pitch and average tone intensity of the music segment are selected to express the development trend of music emotion (Dong et al., 2020). Generally speaking, music with high average pitch gives people a light and loud feeling; The music with low average pitch gives people a sense of heaviness and steadfastness. Music with higher average sound intensity is more penetrating, generally expressing excited and warm emotions; The music with lower average sound intensity is softer, generally expressing soothing and quiet emotions.

The formula for calculating the average pitch $\overline{Pitch}$ of a song is:

$$\overline{Pitch} = \frac{1}{N}\sum_{i=1}^{N} Pitch_i \tag{3}$$

The calculation formula of average sound intensity $\overline{Velocity}$ is:

$$\overline{Velocity} = \frac{1}{N}\sum_{i=1}^{N} Velocity_i \tag{4}$$

where *N* represents the number of notes of a song, $Pitch_i$ represents the pitch of the *i* note of a song, and $Velocity_i$ represents the pitch of the *i* note of a song.

2 Pitch stability and intensity stability

In music creation, the stability of melody often corresponds to the ups and downs of emotion. The stability of melody of music fragments is related to the stability of pitch and intensity. For music with high stability, it indicates that the mood fluctuates less, and usually creates a calm and tranquil atmosphere; For the music with low stability, it indicates that the mood fluctuates greatly and usually creates an active and tense atmosphere.

Then the expressions of pitch stability *Sta_Pitch* and intensity stability *Sta_Velocity* are:

$$Sta\_Pitch = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Pitch_i - \overline{Pitch}\right)^2} \tag{5}$$

$$Sta\_Velocity = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Velocity_i - \overline{Velocity}\right)^2} \tag{6}$$

In the formula, *Pitch* represents the note pitch, $\overline{Pitch}$ represents the average pitch, *Velocity* represents the note pitch, and $\overline{Velocity}$ represents the average pitch.

3 Song speed

The speed of music will also affect the expression of music emotion. Generally, fast music expresses light and lively emotions, while slow music expresses slow and serious emotions. The average sound time of notes can be taken as the song speed *tempo*. If the average sound time of notes is longer, the speed is slower, otherwise, the speed is faster.

The calculation formula of song speed *tempo* is:

$$tempo = \frac{1}{N}\sum_{i=1}^{N} Duration_i \tag{7}$$

In the formula, $Duration_i$ represents the sound time of the *i* note of the song.

4 Music note density

Note density refers to the number of notes arranged horizontally in each bar of the music, that is, the number of notes after the main melody is extracted. Note density can reflect the emotion expressed by music from the side. Generally speaking, music with high note density often expresses strong emotions, such as excitement, tension, etc; Music with low note density often expresses moderate emotions, such as calm and sadness.

The expression of note density *Note_Bar* is:

$$Note\_Bar = \frac{Note\_Number}{Bar\_Number} \tag{8}$$

where *Note_Number* represents the number of musical notes and *Bar_Number* represents the number of musical syllables.

5 Melody direction

Melody is a sequence of note information, including note pitch, sound time and other information. Different melodies have different trends, but similar emotional melodies have certain rules to follow. The influence of melody trend on emotion is generally divided into three types: the upward trend melody usually gives people positive emotions; Downward melodies usually give people negative emotions; The parallel melody needs to be judged in combination with other emotional characteristics.

The melodic direction *Dir_Pitch* of the song is calculated by the note pitch *Pitch* and the sound production time *Duration*, and the expression is:

$$Dir\_Pitch = \sum\nolimits_{i=1}^{N-1} \frac{Pithc_{i+1} - Pitch_i}{Duration_i} \tag{9}$$

The above features constitute an audio modal emotional feature vector *x*, which is prepared for the subsequent research.

$$x\left\{\overline{Pitch}, \overline{Velocity}, Sta + Pitch, Sta\_Velocity, tempo\ , Note\_Bar, Dir\_Pitch\right\} \tag{10}$$

2.2.2 Emotional feature extraction of lyrics

In the lyrics mode, the lyrics text of each song is regarded as a document, and Doc2Vec is used to represent the lyrics text, mapping the text from natural language to mathematical vector form. Most of the previous song lyrics feature extraction methods use BOW (bag) at the document level-of-Words) means that it ignores the finer grained relationship between words and sentences.

Before using the Doc2Vec tool, first clean the lyrics data, remove irrelevant lyrics lines such as ‘lyrics’, ‘composition’, and ‘mixing’, and remove the timestamp. Next, directly conduct model training to obtain the representation of the lyrics text corresponding to each song in the vector space. The vector dimension obtained from the training is determined by the input parameters. Suppose that the lyrics are mapped to the vector space *w* after the model training, and each paragraph or document is mapped to the *k* dimension vector p^k. The expression is:

$$p \to p^k \tag{11}$$

where *p* is the lyrics text corresponding to a song.

X_L is used to represent the emotional characteristics of the lyric mode corresponding to a song, and then the lyric mode is represented as a matrix with the size of 1 * k which is expressed as:

$$X_L \leftarrow p^{1*k} \tag{12}$$

2.3 Multi data fusion of emotional characteristics of multimodal music

By extracting audio modal emotional features and lyrics modal emotional features, these features are fused together, and residual phase (RP) is added to audio modal emotional features to complement audio specific information. RP is defined as the cosine obtained from the analysis of LP residuals of music signals. Set the prediction order as p, the prediction coefficient as a_k, and the prediction error as $e(n)$. For specific music, n and $s(n)$ samples correspond to the linear combination determined by estimation. Under this condition, the predicted samples are:

$$s'(n) = \sum\nolimits_{k=1}^{p} a_k s(n-k) \tag{13}$$

$$e(n) = s(n) - s'(n) \tag{14}$$

LPCs are obtained by minimising the mean square prediction error on the analysis framework. This error $e(n)$ is called the LP residual $r(n)$ of music signals. LP residual contains a lot of information about music emotion, while the phase of the analysis signal derived from LP residual contains better speaker specific information. The analytic signal $r_a(n)$ can be obtained from $r(n)$, $r_h(n)$ is the Hilbert transform of $r(n)$, $R(w)$ is the Fourier transform of $r(n)$, and the calculation formula is:

$$R_h(w) = \begin{cases} -jR(w), o \le w < \pi \\ jR(w), -\pi \le w < 0 \end{cases} \tag{15}$$

Through the above steps, the final analytic signal is obtained. The LP residual contains a lot of information about music emotion. During the research process, specific emotion information can be extracted based on the cosine of the signal phase. Finally, the audio modal emotion feature and the lyrics modal emotion feature are weighted and fused to obtain the following:

$$F_{MF\_RP} = \left(F_{MF}^{T} + F_{RP}^{T}\right)^{T} \tag{16}$$

To avoid the influence of dimensions, the feature data on each dimension is mapped to the [0,1] interval. Use maximum minimum standardisation (Min – Max Normalisation) method, normalise the emotional characteristics of the above fused multimodal music, and set the value of the i position of mode X as X_i, with the function as follows:

$$X_i - \frac{X_i - \min(X)}{\max(X) - \min(X)} \tag{17}$$

2.4 Multimodal music emotion recognition based on classification function

According to the pretreated multimodal music emotion feature multivariate data, a multimodal music emotion classification model is constructed. The purpose of establishing the classification model is to map the multi-modal music emotion feature multivariate data to be classified into four known basic emotions through the classification model, and realise the automatic recognition of multi-modal music emotion. The classification process generally has two stages: the first stage is the training stage, that is, the labeled training set is input into the emotion multi-classifier, and the relevant parameters of the classification model are determined by predicting the given target data; The second stage is the test stage, that is, select the new music file as the test set, use the trained classification model to predict it, and take the recognition accuracy as one of the indicators to evaluate the performance of the model.

Take the pretreated multi-modal music emotional feature multivariate data as the training sample, and make the training sample set $(\overline{x}_i, y_i)$, where $\overline{x}_i$ is the feature vector of the i^{th} *i* sample, y_i is the category to which $\overline{x}_i$ belongs, and the positive category is $y_i = 1$, and the negative category is $y_i = -1$. Assuming that the training set is strictly linearly separable, there is a classification hyperplane as shown in the following formula, which can correctly divide the positive and negative categories in the training set to both sides of the hyperplane.

$$\overline{\omega} * \overline{x} + b = 0 \tag{18}$$

where $\overline{\omega}$ represents the weight vector and b represents the offset.

Calculate the minimum geometric interval from all sample points in the training set to the hyperplane:

$$\delta = \min(\delta_i) \tag{19}$$

where δ_i represents the distance from the sample to the hyperplane H.

Therefore, the relationship between the number N and δ of training samples with wrong classification can be expressed as follows:

$$N \leq \left(\frac{2R}{\delta}\right)^2 \tag{20}$$

Set the optimal weight vector as $\overline{\omega}'$ and the optimal offset as b', and according to the minimum geometric interval from all sample points in the training set to the hyperplane, the optimal multimodal music emotion classification function is constructed:

$$f(x) = \operatorname{sgn}\left(\sum\nolimits_{i=1}^{l} y_i \overline{x}_i + b'\right) \tag{21}$$

- Input: input the preprocessed multi-modal music emotion feature data as a sample into the optimal multi-modal music emotion classification function;
- Output: output the classification results through the optimal multimodal music emotion classification function to identify multimodal music emotion.

Because the emotional features of multimodal music extracted are complex and nonlinear, and the support vector machine (SVM) classification algorithm has great advantages in dealing with small samples and nonlinear problems, this paper selects SVM classification algorithm to build a multimodal music emotional classification model for music emotion recognition. First of all, take the pretreated multimodal music emotion feature multivariate data as training samples, calculate the minimum geometric interval from all sample points of the training set to the hyperplane, and according to the calculation results, input the pretreated multimodal music emotion feature multivariate data into the optimal multimodal music emotion classification function, and output the classification results, so as to identify multimodal music emotion.

3 Simulation experiment analysis

3.1 Experimental scheme

In order to verify the effectiveness of the multimodal music emotion recognition method based on multiple data fusion proposed in this paper in practical application, the accuracy and efficiency of multimodal music emotion recognition are taken as experimental indicators, and the Li et al. (2020) method and the Wang et al. (2022) method are selected as comparison methods to conduct comparative experimental tests with this method.

3.2 Experimental data

In the million song dataset (MSD), there are audio and lyrics data, but lyrics are stored in a word bag, and audio is stored in metadata. In order to obtain the data set corresponding to audio and lyrics, use the song and song name information of mertrolyrics lyrics data set on Kaggle to retrieve and download songs from music websites. Strict search conditions are set to ensure the accuracy of the songs retrieved. Finally, according to the number of the song in mertrolyrics, the corresponding lyrics data and the audio data of the song are combined to obtain a multimodal music style dataset corresponding to the audio and lyrics. Because the less training samples, the easier it is to over-fit, and the poor generalisation of the model, it is necessary to consider the number of training samples to improve the generalisation performance. This paper selects 5,000 pieces of music as the sample data set, 3,000 pieces of music as the test sample, and 2,000 pieces of music as the training sample. The data set includes five style categories, hip-hop, metal, country, folk, jazz, and each category has 1,000 songs. Each song contains audio and lyrics.

3.3 Experimental indicators

The experiment uses multi-modal music emotion recognition accuracy and recognition efficiency as the evaluation criteria:

$$Precisin = \frac{TP}{TP + FP} \tag{22}$$

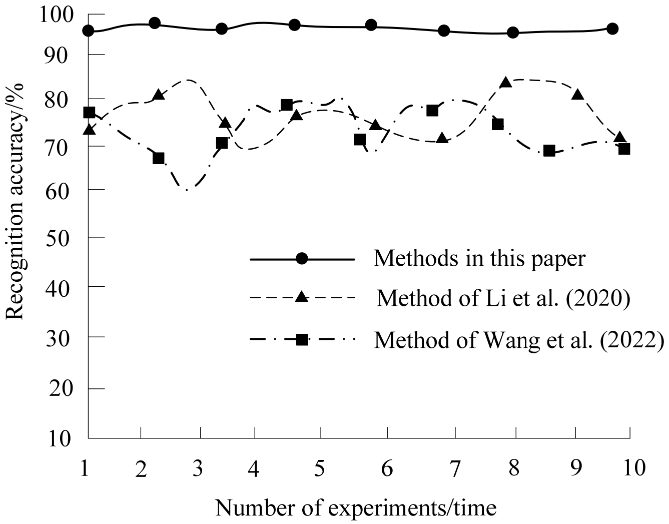
$$S = \frac{Q}{t} \tag{23}$$

In the formula, *TP* represents the number of positive samples correctly divided into positive ones, *FP* represents the number of negative samples wrongly divided into positive ones, *Q* represents the workload of multimodal music emotion recognition, and *t* represents the time taken for recognition.

3.4 Experimental comparison results

The multi-modal music emotion recognition method based on multivariate data fusion, the Li et al. (2020) method and the Wang et al. (2022) method proposed in this paper are used to compare and analyse the accuracy of multi-modal music emotion recognition. The comparison results are shown in Figure 1.

Figure 1 Comparison results of multi-modal music emotion recognition accuracy of three methods



According to Figure 1, the accuracy of multimodal music emotion recognition in this method can reach 99% at most, the accuracy of multimodal music emotion recognition in Li et al. (2020) method is only 85% at most, and the accuracy of multimodal music emotion recognition in Wang et al. (2022) method is only 80% at most. The accuracy of multimodal music emotion recognition in this method is the highest, and the recognition effect is the best.

In order to further verify the effectiveness of the method proposed in this paper, the multi-modal music emotion recognition method based on multiple data fusion, the Li et al. (2020) method and the Wang et al. (2022) method proposed in this paper are used to conduct a comparative analysis of the time used for multi-modal music emotion recognition. The comparative results are shown in Table 1.

Table 1 Time comparison results of multimodal music emotion recognition by three methods/s

Number of experiments/time	*Methods in this paper*	*Li et al. (2020) method*	*Wang et al. (2022) method*
10	11.23	16.77	25.21
20	11.34	16.92	25.84
30	12.26	17.87	26.11
40	12.54	18.36	26.86
50	12.76	18.94	27.13
60	12.98	19.23	27.99
70	13.26	19.47	28.54
80	13.73	19.95	28.92

According to Table 1, it can be seen that the time used for this method to identify multimodal music emotion is within 13.73s, the time used for Li et al. (2020) method to identify multimodal music emotion is within 19.95s, and the time used for Wang et al. (2022) method to identify multimodal music emotion is within 28.92s. The time used for this method to identify multimodal music emotion is the shortest and the identification efficiency is the highest.

4 Conclusions

As the traditional method of multimodal music emotion recognition consumes a long time, resulting in poor recognition effect, this paper studies multimodal music emotion recognition method based on multiple data fusion. The non-negative matrix decomposition method is used to decompose the multimodal music emotion to obtain the multimodal data of audio mode and lyrics mode, and the emotional features of the multimodal data are extracted. The emotional features of the audio mode are extracted through pitch, intensity, etc. The modal features of the lyrics are extracted through Doc2Vec, the extracted emotional features of the two modes are weighted and fused, and the fused multimodal music emotional features are preprocessed and taken as samples, calculate the minimum geometric interval from the sample point to the hyperplane, construct the classification function, and complete the purpose of multi-modal music emotion recognition. Experiments show that the method in this paper takes the shortest time to identify multimodal music emotions and has the highest recognition accuracy. The advantages of this method are:

1 the accuracy of multimodal music emotion recognition is up to 99% and the recognition effect is the best

2 the time of multimodal music emotion recognition in this method is 13.73s, and the recognition efficiency is the highest.

References

Dong, Y.C., Kim, D.H. and Song, B.C. (2020) 'Multimodal attention network for continuous-time emotion recognition using video and EEG signals', *IEEE Access*, Vol. 8, pp.203814–203826.

Fan, T., Wu, P. and Cao, Q. (2020) 'Research on multi-mode fusion netizen emotion recognition based on deep learning', *Journal of Information Resource Management*, Vol. 17, No. 1, pp.39–48.

Ferguson, N. (2022) 'Music emotion recognition based on segment 1evel two stage 1eaming', *International Journal of Multimedia Information Retrieval*, Vol. 11, No. 3, pp.383–394.

Jia, N., Zheng, C. and Sun, W. (2022) 'A multimodal emotion recognition model integrating speech, video and MoCAP', *Multimedia Tools and Applications*, Vol. 81, No. 22, pp.32265–32286.

Jian-Wei, N., Yue-Qi, A.N., Jie, N.I., et al. (2022) 'Multimodal emotion recognition based on facial expression and ECG signal', *Packaging Engineering*, Vol. 43, No. 4, pp.71–79.

Lee, S., Han, D.K. and Ko, H. (2021) 'Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification', *IEEE Access*, Vol. PP, No. 99, pp.1–1.

Li, X., Han, L., Li, J., et al. (2020) 'Multi-modal music emotion classification based on optimized residual network', *Computer and Modernization*, Vol. 29, No. 12, pp.83–89.

Ma, J., Sun, Y. and Zhang, X. (2019) 'Multimodal emotion recognition for the fusion of speech and EEG signals', *Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University*, Vol. 46, No. 1, pp.143–150.

Ni, L. (2020) 'Design of music emotion classification method based on dual modes of audio and lyrics', *Techniques of Automation and Applications*, Vol. 39, No. 5, pp.166–169.

Wang, C., Li, W. and Chen, Z. (2021) 'Multimodal emotion recognition based on speech and video images', *Computer Engineering and Application*, Vol. 57, No. 23, pp.163–170.

Wang, H., Liu, Y., Zhao, M., et al. (2022) 'Research on multi-task music emotion recognition based on multi-modal features', *Journal of Modern Information*, Vol. 42, No. 11, pp.61–75.

Wang, Z., Zhou, X., Wang, W., et al. (2020) 'Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video', *International journal of machine Learning and Cybernetics*, Vol. 11, No. 4, pp.923–934.

Zhang, Y., Cheng, C. and Zhang, Y. (2021) 'Multimodal emotion recognition using a hierarchical fusion convolutional neural network', *IEEE Access*, Vol. PP, No. 99, pp.1–1.

Zhao, J., Liu, H., Liang, X., et al. (2021) 'Multimodal music emotion recognition based on knowledge distillation and transfer learning', *Journal of Fudan University: Natural Science Edition*, Vol. 60, No. 3, pp.309–314, p.322.

Zhao, Y., Wang, Y., Zhou, Y., et al. (2019) 'Research on multi-modal music emotion classification based on DBN', *Information Technology*, Vol. 43, No. 2, pp.102–106, p.110.