International Journal of Information System Modeling and Design
Volume 15 • Issue 1

Construction and Implementation of Content-Based National Music Retrieval Model Under Deep Learning

Jing Shi, Nanchang Institute of Technology, China

Lei Liu, Luxun School of the Arts, Yan'an University, China*

ABSTRACT

This research mainly studies the construction and implementation of the content-based folk music retrieval model. Firstly, it studies the music automatic annotation method based on deep learning, and then proposes the tag conditional random field music automatic annotation method, and then constructs the music annotation depth neural network model combining a variety of music representation and attention mechanism. Finally, it analyzes the proposed folk music retrieval model the effectiveness of the cable model is verified and its performance is evaluated. The results show that in Glu module, Glu blocks had better performance in music annotation, and the music annotation results of each index in music hierarchical sequence modeling are better, which ensures the effectiveness of music annotation. Compared with other algorithms, the AUC tag score of the proposed method is the highest, which is 0.913; it can better model the mapping relationship between the audio features of music input to the text tag and has higher scores on all evaluation indicators.

KEYWORDS

Conditional Random Field, Deep Learning, Folk Music, Multi Label Classification, Neural Network, Retrieval, Self-Attention

In recent years, with the rapid development of internet technology and multimedia applications, many musical works are being uploaded to online digital music libraries (Müller et al., 2019). In the face of huge online music data, it seems like people could find any music suitable for them, but for ordinary users, it is becoming more and more difficult to conveniently find the music they want (Wang et al., 2020). Major digital music providers are now faced with the challenge of providing users with effective music recommendation and retrieval services.

The application of supervised machine-learning methods enables the automatic addition of descriptive labels to music. This technology can efficiently use large-scale music data to automatically add descriptive text labels to music based on the content of the music data, which improves the user's experiences with services such as retrieval and recommendation (Shen et al., 2019). However, the labels generated by automatic music-annotation technology often provide high-level semantic descriptions, which creates a significant disparity with low-level audio representation and poses challenges in the

DOI: 10.4018/IJISMD.343631

*Corresponding Author

method's design, including feature representation and label association. This study aims to investigate effective models and algorithms for automatic music annotation, proposing a music-annotation method based on conditional random fields of music fragments. It aims to aggregate labels from all music fragments to achieve comprehensive music annotation.

Some scholars have used the general summarization algorithm previously applied to text and voice to summarize the projects in the music dataset and evaluated the summarization process of category II and multi-category music-classification tasks by comparing the accuracy of the summary dataset with the complete songs using human-oriented summaries, continuous segments, and original datasets. The results show that compared with the selected baseline, Grasshopper, LexRank, latent semantic analysis (LSA), maximal marginal relevance (MMR), and a centrality model based on a support set all improve the classification performance (Raposo et al., 2016). A local merging method of song feature vectors based on a general background model was proposed, which includes two local activation modes of feature vectors: histogram representation and binary vector representation. Experiments on three open music datasets show that the proposed method is effective in music similarity computation (Seo, 2018).

Other scholars applied MusicMixer to propose a topic modeling method for retrieving similar music clips (Hirai et al., 2018). The MusicMixer method mixes songs according to the similarity of audio by beat-frequency analysis and potential theme analysis of chromatic signal in audio. Furthermore, a method to represent audio signal is proposed to construct a topic model to obtain audio latent semantics. Experimental results show the effectiveness of the proposed latent semantic analysis method. Users can select a song from the list of songs suggested by the system to perform DJ mixing. An expert team developed a two-level accurate and fast query-by-example–based music information retrieval system by using feature-fusion technology and decision-fusion technology. In the first stage, a variety of recognizer sets will automatically identify the type of query; in the second stage, the similarity between the query and other content of the same query-type dataset is measured to find the target song, a genre-adaptive feature-extraction method is proposed, and the feature-fusion technology is used to fuse the features. The results show that the accuracy and retrieval time have been significantly improved (Borjian et al., 2018). Based on the above research results, this research will focus on the optimization of automatic music annotation and build a content-based folk-music retrieval model. The steps are:

1. Literature review: conduct a comprehensive review of existing research on content-based music retrieval and automatic music annotation using deep-learning techniques. Identify the strengths and limitations of current methods.
2. Data collection and preprocessing: collect a suitable dataset of folk-music recordings with associated tags or annotations. Preprocess the dataset by removing noise, normalizing audio levels, and segmenting songs into individual units.
3. Automatic music-annotation method: study and develop a deep learning–based automatic music-annotation method specifically tailored for folk music. Explore different techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms, to effectively annotate music based on its content.
4. Tag conditional random field (CRF) method: propose and implement a tag conditional random field automatic music-annotation method. This method combines the power of CRF models with deep-learning techniques to improve the accuracy and reliability of music annotation.
5. Music annotation deep neural network model: construct a music-annotation deep neural network model that incorporates various music representations and attention mechanisms. The model should be designed to capture complex features and relationships between music audio input and text tags.
6. Experimental analysis: analyze and evaluate the effectiveness of the proposed folk-music retrieval model. Verify the performance of the gated linear unit (GLU) module and GLU blocks in music

annotation. Assess the quality of music-annotation results in terms of hierarchical sequence modeling. Compare the proposed method's area under the curve (AUC) tag score with other algorithms and evaluate its performance using various evaluation indicators.

7. Results analysis and discussion: interpret and analyze the results obtained from the experiments. Discuss the strengths and weaknesses of the proposed model compared to existing methods. Determine the effectiveness of the model in accurately annotating folk music and retrieving relevant content.
8. Application and future work: demonstrate the practical application and potential applications of the developed content-based folk-music retrieval model. Discuss possible areas of improvement and future research directions to enhance the model's performance and expand its capabilities.

This research aims to contribute to the field of content-based folk-music retrieval by developing an effective and accurate model for music annotation and retrieval. The proposed methods should provide valuable insights into the mapping relationship between audio features and text tags, facilitating the exploration and discovery of folk music from diverse traditions and regions.

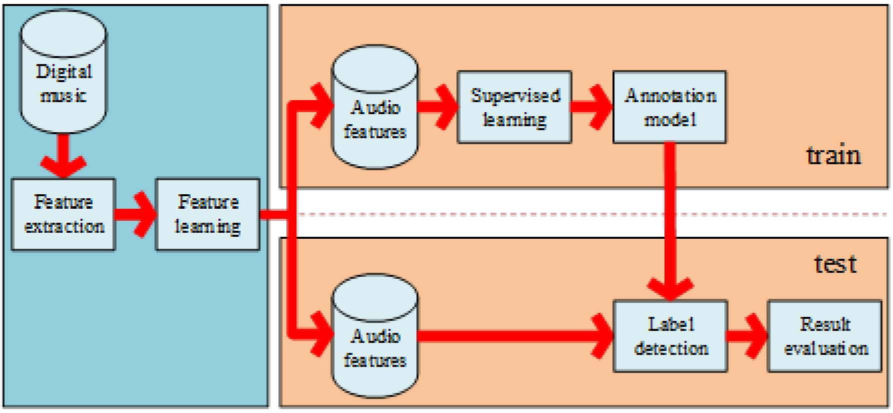
RELATED WORK

Sotiropoulos et al. (2008) explore the use of objective audio-signal features to model the individualized (subjective) perception of similarity between music files. They present MUSIPER, a content-based music retrieval system that constructs music-similarity perception models of its users by associating different music similarity measures to different users. Feature learning and deep learning have drawn great attention in recent years as a way of transforming input data into more effective representations using learning algorithms. Nam et al. (2015) present a two-stage learning model to effectively predict multiple labels from music audio. Yu et al. (2019) propose a deep cross-modal correlation learning architecture involving two-branch deep neural networks for audio modality and text modality (lyrics). A pretrained Doc2Vec model followed by fully connected layers (fully connected deep neural network) is used to represent lyrics. Ghosal and Kolekar (2018) propose a novel approach for music-genre recognition using an ensemble of convolutional long short-term memory based neural networks (CNN LSTM) and a transfer-learning model. The neural network models are trained on a diverse set of spectral and rhythmic features, whereas the transfer-learning model was originally trained on the task of music tagging.

Xie et al. (2018) propose a CNN-based hard-hat detection algorithm. In this algorithm, the detection of construction workers and hard hats is assisted by a computer-vision technique where deep-learning models are trained to identify the proper wearing of hard hats. Based on characteristics of the knowledge expression of construction procedural constraints in Chinese regulations, Zhong et al. (2020) explore a hybrid deep neural network, combining bidirectional LSTM and CRF for the automatic extraction of the qualitative construction procedural constraints. The model-implementation results demonstrate the good performance of the end-to-end deep neural network in the extraction of construction procedural constraints. In this research work, a deep learning–based model has been discussed for content-based image retrieval (CBIR). Singh et al. (2020) study CBIR-CNN, content-based image retrieval on celebrity data using deep convolution neural network. For classification purposes, a four convolution layer model has been proposed.

Content-based music information retrieval has seen rapid progress with the adoption of deep learning. Manco et al. (2021) propose to address music description via audio captioning, defined as the task of generating a natural language description of music audio content in a human-like manner. In order to study the application of the deep-learning method in music-genre recognition, Xu (2022) proposes the parameter-extraction feature and the recognition-classification method of an ethnic music genre based on the deep beliefs network (DBN) with five kinds of ethnic musical instruments as the experimental objects. The DBN is the best way for softmax to identify and classify national musical

Figure 1. Schematic Diagram of Basic Method Framework of Automatic Music Annotation



instruments, and the accuracy rate is 99.2%. The deep CNN model in the field of deep learning has achieved good results in the fields of image and voice. Miao and Cheng (2023) study construction of a multimodal automatic music-annotation model based on a neural network algorithm. The construction of a multimodal automatic music-labeling model based on a neural network algorithm is launched.

CONSTRUCTION OF CONTENT-BASED FOLK-MUSIC RETRIEVAL MODEL

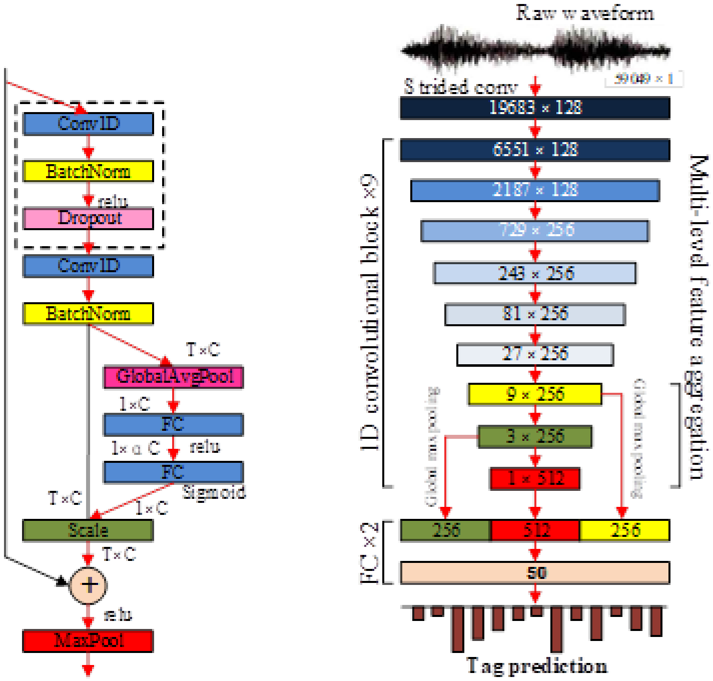
Automatic Music Annotation Based on Deep Learning

In previous research, numerous algorithms have been proposed by researchers to address the issue of automatic music annotation. The overall framework is shown in Fig. 1. In the process of automatically labeling music, the music is typically categorized or associated with various types of labels. First, audio features are used to represent music samples, and then unsupervised feature learning or feature selection and combination are used to represent music features. For the existing audio-feature representation, the input of the algorithm generally combines the set of audio text tags and the training set of feature representation and outputs the trained automatic-annotation model. The model can be applied to each test set and predict the music samples and text tags. At the same time, the prediction results can be compared with the real annotation. After the relevant evaluation indexes are calculated, the calculation results are estimated to evaluate the annotation performance of the model.

In the above process, the audio-feature representation of feature learning and music-label prediction are the main tasks of automatic music annotation. For audio-feature representation, it is mainly divided into artificial design and feature learning. The artificial audio-feature representation is based mainly on the inherent sampling rate of the original audio waveform signal storage, such as WAV and MP3, and then by autoregressive modeling, Fourier transform signal-processing methods to obtain the audio-feature representation. However, the process of manual design is very time-consuming, and it is necessary to try a variety of schemes before choosing or combining a better scheme. Feature-learning audio-feature representation is to automatically find suitable feature representation for corresponding tasks through machine learning. According to adaptive feature learning, it is not necessary to have more prior knowledge of music signal when discussing music-related automatic-annotation problems so as to better study the annotation model. Generally, feature-learning methods include sparse coding, K-means, and restricted Boltzmann machine. These methods are based on deep learning and integrate label prediction and feature learning in a deep neural network.

Deep neural network models are widely used in various fields. At the same time, some music-annotation methods are also applied to neural network structure, such as multilayer perceptron, CNN,

Figure 2. Architecture of Music-Annotation Methods Based on Deep Learning with Primitive Feature Representation



(a) A stack of convolution modules combining resnets and senets

(b) Overall network framework

and RNN. Through data modeling and processing, these neural network structures can be divided into three kinds of music-annotation methods, which are multilayer perceptron, convolutional neural network, and a music-annotation method combined with sequence modeling. At present, the music-annotation method based on deep learning is widely used and usually uses relatively primitive feature representation. This kind of network model can better intersect with the maximum pooling layer so that the high-level abstract-feature representation can also be learned and absorbed; then, according to the output prediction of the subsequent full connection layer, the confidence of the corresponding label can be obtained. In natural language processing and other sequence-modeling problems, a one-dimensional maximum pooling layer and convolution operation are often used to capture the characteristics of the audio signal. Generally, one-dimensional convolution is represented by a two-dimensional time-frequency signal, and the convolution kernel slides on the time axis. At this time, the input feature and convolution kernel have the same frequency dimension, as shown in Fig. 2.

Label Conditional Random Field Automatic Music Annotation

Most music-annotation methods divide a piece of music into a certain-length segment or directly regard it as a whole in order to model the internal relationship of music. In this study, a hierarchical adaptive music-representation method is applied. First, the audio signal sequence of music is determined and divided into multiple segments with the same audio characteristics, in which the length of the music segment is determined by its signal characteristics. Therefore, this study combines the previous research results, that is, using the decomposition algorithm based on self-similarity to segment music. In order to represent the characteristics of audio signals on music clips, a set of low-level audio features, such as energy, intensity and its ratio, spectrum attenuation and flux, spectrum contrast, pitch level profile, and MFCCs, is also applied in this study. In order to find the relationship between local music clips and text tags more accurately, we need to infer the tags of each music clip. The semantic tags of clips in the same music are different, but the tags of adjacent clips are more similar. According to the understanding of the image semantic segmentation method, this paper proposes an automatic music-annotation method based on conditional random fields.

In general, music text tags do not exist independently, but have certain relevance. The correlation of music tags is calculated by the Jaccard similarity coefficient, as shown in equation (1) (Bag et al., 2019).

$$C(i,j) = \frac{n_{ij}}{n_i + n_j - n_{ij}} \tag{1}$$

In equation (1), n_{ij} and n_j are the number of samples marked i and j, n_{ij} is the number of samples marked at the same time, and the Jaccard similarity coefficient is $\left[0,1\right]$. On the fragment of music tag i, a mutually exclusive tag group G_i conditional random field is constructed. Music X can be expressed as $X = \left\{x_p\right\}_{p=1}^{N}$, with p as music clips, N as the number of music clips, and x_p as the audio feature vector of p. If y is the tag set corresponding to X, then the tag in G_i corresponding to y is $y^i \subseteq y$. Therefore, the energy function of the conditional random field is shown in equation (2).

$$E\left(X,G_i,f_i\right) = \sum_p D\left(x_p,f_p^i\right) + \sum_{(p,q)\in N} V\left(f_p^i,f_q^i\right) + \tau\left(y^i,f^i\right) \tag{2}$$

In equation (2), $f^i = \left[f_p^i\right]_{p=1\ldots N}$ is the subset of text tags corresponding to each segment of X in G_i. $D\left(x_p,f_p^i\right)$ measures the relationship between x_p and f_p^i, $V\left(f_p^i,f_q^i\right)$ constrains the similarity of f_p^i and f_q^i, τ also constrains the consistency of the music-level label and the clip-level label, and N is the adjacent relationship between clips. Due to the high similarity of adjacent clips in the same music, it is necessary to provide corresponding constraints for $V\left(f_p^i,f_q^i\right)$, so that the text label can move smoothly on the time axis, as shown in equation (3).

$$V\left(f_p^i,f_q^i\right) = \left(1 - C\left(f_p^i,f_q^i\right)\right) S\left(x_p,x_q;w^i\right) \tag{3}$$

In equation (3), f_p^i and f_q^i are the comparison signature subsets of music tags corresponding to p and q in G_i; p, q is two adjacent music clips; and $S\left(x_p,x_q;w^i\right)$ measures the similarity between x_p, x_q of two music clips, which is calculated based on the regression parameter w^i of tag i. The audio-feature similarity calculation of the p, q self-test is shown in equation (4).

$$S\left(x_p,x_q;w^i\right) = \begin{cases} \exp\left(-\dfrac{1}{2\sigma^2} Dist\left(x_p w^i, x_q w^i\right)\right), & \left(p,q\right) \in N \\ 0, & Otherwise \end{cases} \tag{4}$$

In equation (4), $x_p w^i$, $x_q w^i$ means that x_p, x_q of two adjacent music clips is obtained from the weighting of corresponding elements of regression parameter of tag i, σ is scale super parameter, and $Dist()$ is Euclidean distance. In order to increase the flexibility of labeling waveforms, local global consistency is labeled in the last item after equation (2), and inconsistency penalty terms of fragment-level labeling f^i and level labeling y^i are defined, as shown in equation (5).

$$\tau\left(y_i, f_i\right) = c\sum_{l,p} \eta\left(y_l^i = 0 \wedge f_p^i = l\right) \tag{5}$$

In equation (5), $\eta(\cdot)$ is the indicator function, which is 1 when the condition is satisfied, otherwise it is 0; c is the consistency-strength super parameter that controls the level labeling of music clips.

With tag-specific feature learning, sparsity, sharing, and discrimination of music features should be fully considered. After these three characteristics are combined, the objective function of feature learning is shown in equation (6).

$$\arg\min_{w} \frac{1}{2}\left\|\bar{X} W - F\right\|^2 + \frac{\alpha}{2} Tr\left(R W^T W\right) + \beta \left\|W\right\|_l \tag{6}$$

In equation (6), $\bar{X} = \left[x_1, x_2, ..., x_M\right]^T$ is the audio feature of M music clips, $F = \left[f_1, f_2, ..., f_M\right]^T$ is the annotation vector corresponding to the clips, $W = \left[W^1, W^2, ..., W^L\right]$ is the regression vector corresponding to the tags that need to be learned, R is the distance matrix between tags, α, β is the weight super parameter of each item in the control objective function, and the values are not less than 0. In regards to the minimization problem, the optimization method selected in this study refers to the previous research results, namely, the accelerated proximal gradient method, to solve the non-smooth problem (Shimizu & Kanno, 2018). The general form of this method is shown in equation (7).

$$\arg\min_{W \in \mathrm{H}} \left\{G\left(W\right) = g\left(W\right) + h\left(W\right)\right\} \tag{7}$$

In equation (7), H is a real Hilbert space, $h\left(W\right)$, $g\left(W\right)$ is a convex function, and $h\left(W\right)$ is Lipschitz continuous, where $\left\|\nabla h\left(W_1\right) - \nabla h\left(W_2\right)\right\| \le L_f \left\|\Delta W\right\|$, $\Delta W = W_1 - W_2$, and L_f are Lipschitz constants. The update process of the proximal gradient is shown in equation (8).

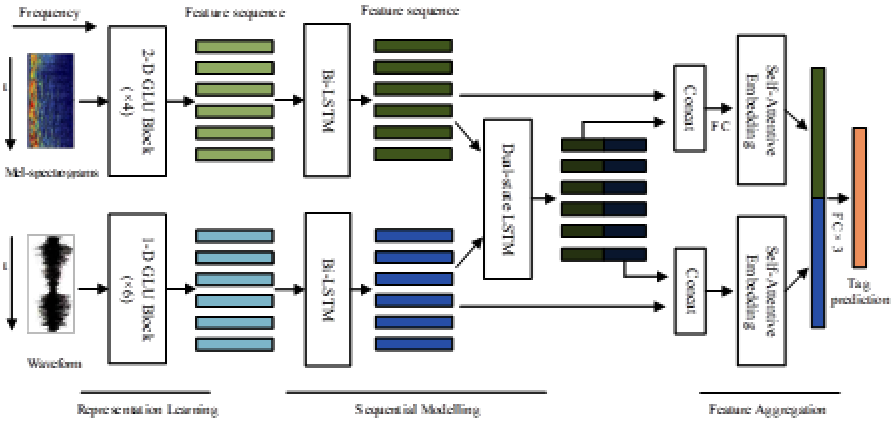
$$W^{t+1} = prox_{\lambda h}\left(W^t - \lambda \nabla g\left(W^t\right)\right) \tag{8}$$

In equation (8), λ is the step parameter, and the value is $1 / L_f$; W^t is the value of the t iteration of W; and $prox_{\lambda h}$ is the proximal operator of the function $h(\cdot)$ at λ.

Music-Annotation Deep Neural Network Model Based on Combination of Multiple Music Representation and Attention Mechanism

In order to solve the problem of automatic music annotation, this study proposes an automatic music-annotation model based on deep learning, as shown in Fig. 3. In this model, first, multiple GLUs are

Figure 3. Structure Diagram of Hierarchical Attention Deep Neural Network for Automatic Music Annotation



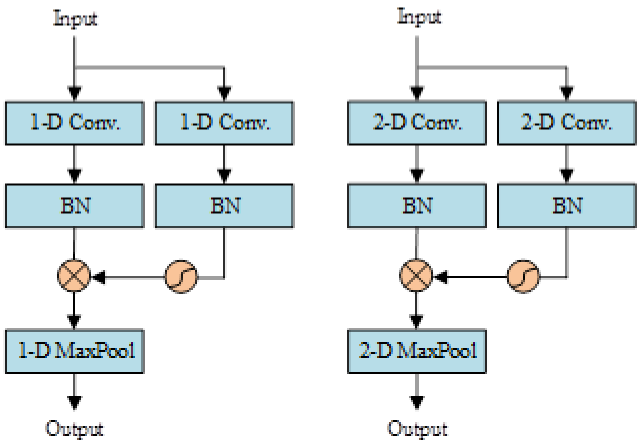
combined, and then music-feature representation is learned from one-dimensional raw waveform and two-dimensional Mel spectrogram, so more-abundant music information can be obtained. Then, a two-way long-term and short-term memory network is used to encode the hierarchical sequence structure of the two kinds of music, and double-memory LSTM is used to encode the temporal correlation between the two kinds of music. Finally, the music representation of each moment is aggregated into the overall feature representation through the self-attention weight mechanism, and finally the most suitable label for music prediction is achieved.

In order to make the music-tagging task better learn the music features related to text tags, this study introduces convolutional neural network layer in volume to process music data and introduces attention mechanism in the volume layer, that is, using gated linear unit to replace the ReLU function in convolution. Fig. 4 is a schematic diagram of the one-dimensional and two-dimensional GLU module structure.

The forward calculation method of the GLU module is shown in equation (9).

$$Y = \left(W * X + b\right) \odot \sigma\left(V * X + d\right) \tag{9}$$

Figure 4. Schematic Diagram of Module Structure of One-Dimensional GLU and Two-Dimensional GLU



In equation (9), W , V is the convolution kernel, b , d is two cheap terms, $*$ is convolution operation, $\odot$ is the product of corresponding elements of matrix, and σ is the activation function of sigmoid. From the perspective of learning, GLU calculates the importance of each element in the feature representation and then allocates the weight to achieve the purpose of attention mechanism on each time-frequency representation element. In order to learn and extract effective audio-feature representation in Mel spectrum, four cascaded 2D GLU modules are used to represent learning branches, and the convolution kernel size is 3×3 , the corresponding convolution kernels of the four modules are 32, 48, 64, and 64, and the maximum pooling operation step 2×4 of the first three modules and 2×2 of the last one are set. At the same time, six one-dimensional GLU modules are proposed to correspond to the amplitude signal of one-dimensional time series in the original waveform. The corresponding convolution cores are 16, 16, 24, 32, 48, and 64. The convolution core size is 5, and the pooling layer size is 4. The two-dimensional convolution feature-learning branch corresponding to Mel spectrum input and the one-dimensional convolution feature-learning branch corresponding to the original waveform signal input will finally output a feature sequence with the size of 56×64 and input the feature sequence into RNN structure to achieve the long-term time-sequence structure of learning music.

LSTM is a variant function applied in RNN. It has three control gates, input, forgetting, and output, and a unit state variable. If t exists at each time, the calculation method is shown in equation (10).

$$\begin{cases} i_t = \sigma\left(W_{ih}h_{t-1} + W_{ix}x_t + b_i\right) \\ f_t = \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right) \\ o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right) \\ c_t = f_t \odot c_{t-1} + i_t \odot \varphi\left(W_{ch}h_{t-1} + W_{cx}x_t + b_c\right) \\ h_t = o_t \odot \varphi\left(c_t\right) \end{cases} \tag{10}$$

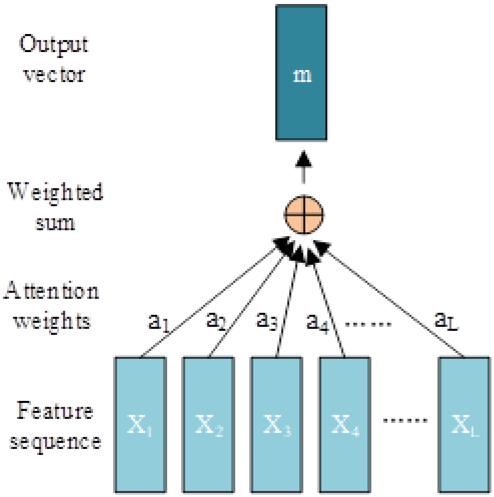
In equation (10), x_t , h_t is the input and output at time t ; i_t , f_t , o_t is the corresponding value of input, forgetting, and output gates; c_t is the unit state value at the time; W_{*h} , W_{*x} is the corresponding weight matrix; b_* is the bias term; and σ , φ is the activation function of sigmoid and tanh, respectively. By introducing an additional network branch, the standard LSTM model with a single memory state is extended to two-unit memory states and combined with the output of RNN, which represents the channel order modeling, to represent the potential temporal correlation, as shown in equation (11).

$$\begin{cases} i_t = \sigma\left(W_{ih}h_{t-1}^m + U_{ih}h_{t-1}^w + W_{ix}x_t^m + U_{ix}x_t^w\right) \\ f_t = \sigma\left(W_{fh}h_{t-1}^m + U_{fh}h_{t-1}^w + W_{fx}x_t^m + U_{fx}x_t^w\right) \\ o_t = \sigma\left(W_{oh}h_{t-1}^m + U_{oh}h_{t-1}^w + W_{ox}x_t^m + U_{ox}x_t^w\right) \end{cases} \tag{11}$$

In equation (11), U_{*h} , U_{*x} is the hidden state h_{t-1}^w corresponding to the previous time and the input characteristic x_t^w corresponding to the current time. Furthermore, the memory and hidden state of two groups of units representing branches can be obtained, as shown in equation (12).

$$\begin{cases} c_t^m = f_t \odot c_{t-1}^m + i_t \odot \varphi\left(W_{ch}h_{t-1}^m + W_{cx}x_t^m\right) \\ h_t^m = o_t \odot \varphi\left(c_t^m\right) \\ c_t^w = f_t \odot c_{t-1}^w + i_t \odot \varphi\left(U_{ch}h_{t-1}^w + U_{cx}x_t^w\right) \\ h_t^w = o_t \odot \varphi\left(c_t^w\right) \end{cases} \tag{12}$$

Figure 5. Schematic Diagram of Single-Attention Weight Mechanism



In equation (12), c_t^m, h_t^m is the Mel spectrum, which represents the corresponding unit memory state and hidden state; c_t^w, h_t^w is the original waveform signal, which represents the corresponding unit memory state and hidden state. The above-mentioned hierarchical sequence modeling is combined with a self-attention mechanism to realize the prediction of music tags.

Figure 5 shows a typical self-attention weight module. Given an input sequence $X = \left[x_1 .x_2, \ldots, x_{L_x}\right]^T$ and L_x as the length of the feature vector, the attention weight represented by the feature at each moment is shown in equation (13).

$$a = soft\max\left(w_2\varphi\left(W_1X^T\right)\right) \tag{13}$$

In equation (13), D_x represents the dimension of the vector x_i in the feature sequence; $W_1 \in R^{D_x \times D_x}$ and $w_2 \in R^{D_x}$ are the weight matrices to be learned; $soft\max\left(\cdot\right)$ ensures that the sum of attention weights is 1.The overall feature representation m is obtained by computing the weighted sum of each representation in the feature sequence.

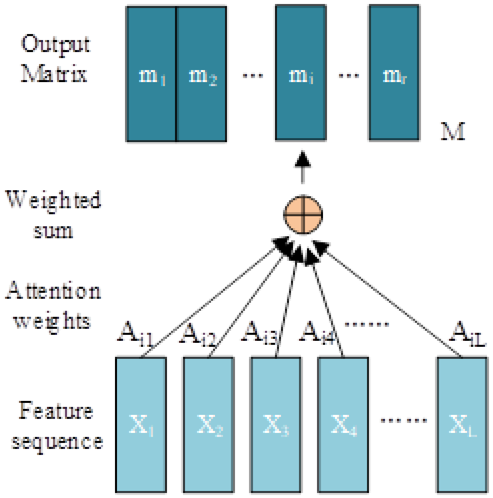
See equation (14).

$$m = \sum_{t=1}^{L_x} a_t x_t = aX \tag{14}$$

First, the matrix of single-attention weight is calculated by the feature sequence $X = \left[x_1 .x_2, \ldots, x_{L_x}\right]^T$ output in the previous stage, $w_2 \in R^{D_x}$ is extended to $W_2 \in R^{r \times D_x}$, and r is the super parameter indicating the number of attention weight vectors. Fig. 6 shows how to calculate the weight of multiple attention.

The calculation method of attention weight matrix A is consistent with equation (13). Then calculate the feature sequence of two-dimensional embedding matrix M aggregate music; see equation (15).

Figure 6. Schematic Diagram of Single-Attention Weight Mechanism



$$M = AX \tag{15}$$

Among them, the i th row vector m_i in M is the weighted sum $m_i = \sum_j A_{i,j} x_j$ of the feature vectors corresponding to each time in the feature sequence, and each individual vector actually pays attention to different parts of the whole feature sequence so as to retain the music features of different aspects of the sequence features.

The proposed annotation model has three full connection layers to predict music tags. The first two dimensions are 1,536, and the activation method is ReLU; the third dimension is 188, which ensures that the output value is within $[0,1]$ through sigmoid and is used as the confidence value of the text tag.

EXPERIMENTAL ANALYSIS OF CONTENT-BASED FOLK-MUSIC RETRIEVAL MODEL

Validation of the Model

This experiment used the commonly used MTAT dataset to analyze and verify the performance of the music annotation method model (Bisharad & Laskar, 2019). MTAT stands for MagnaTagATune, which is a popular dataset used for music information retrieval tasks, specifically music auto-tagging. The MagnaTagATune dataset consists of audio excerpts from various genres and contains associated tags that describe the musical characteristics or content of the audio. It is widely used in research to develop machine-learning models for automatically assigning relevant tags to music based on its acoustic features. The music samples in the first 12 subdirectories of the MTAT dataset are used as the training set, the 13th as the verification set, and the 14th–16th as the test set. At the same time, 128-dimensional logarithmic Mel spectrum features are extracted before the experiment, the sampling rate is 16 KHz, and the analysis window of 512 is applied. The samples of the original waveform signal are sampled at 8 KHz, and each dimension of all Mel spectrum features is normalized through the training data to obtain a single variance or zero mean; that is, the original waveform signal is shrunk to within $[-1,1]$. The retrieval model will be realized through TensorFlow, and the specific parameter setting of TensorFlow refers to the previous research results (Sanchez et al., 2020).

Table 1. Performance Comparison of Music Representation Learning Annotation Based on Different Convolutional Structures

Performance	Regular Conv.		SE Blocks		GLU Blocks	
Top *N* tags	50	188 (All)	50	188 (All)	50	188 (All)
AUC-tag	0.902	0.889	0.902	0.890	0.913	0.900
AUC-clip	0.934	0.955	0.935	0.955	0.945	0.962
MAP-tag	0.427	0.194	0.428	0.196	0.458	0.212
MAP-clip	0.692	0.603	0.695	0.608	0.720	0.635

In the compression excitation network (SENet) network structure, the labeling performance of the proposed model (GLU blocks) and GLU module replaced by other forms of convolution module is compared, and the comparison results are shown in Table 1. Among them, regular conv. refers to the maximum pool replacement of GLU module by the combination of ReLU activation function and conventional 1D/2D convolution; SE blocks refers to the replacement of a 1D/2D GLU module in the music representation learning part of the model with a 1D/2D SE module (Foote, 1997).

It can be seen from Table 1 that under SE module, GLU blocks music representation learning annotation performance is significantly better than that of SE blocks and regular conv., and conventional convolution operation annotation performance is lower than music representation learning annotation performance. It shows that the attention weight calculation method in GLU module is more fine-grained and more suitable for the overall annotation model and the music annotation performance is better (Li & Ogihara, 2004).

The effectiveness of the proposed method is further verified by the results of several variations of the music sequence model (Ajoodha et al., 2015). Among them, WAV and MEL refer to the music annotation that uses a single music representation branch of Mel spectrum and original waveform signal; MEL + WAV refers to the realization of music annotation that uses only two music representation forms; and CORR refers to the annotation that uses only the feature sequence based on the correlation between two music representation branches that is calculated by the extended two-state LSTM (Aigrain et al., 1996).

It can be seen from Table 2 that the music-annotation results of each index of the model proposed in this study are good, which ensures the effectiveness of this music hierarchical sequence modeling. We need to verify the effectiveness of the model's self-attention label prediction and compare it with the method proposed in this paper (multi-attention weighting) against conventional single-attention weight schemes and maximum pooling schemes. The results are presented in Table 3.Among them, max-pooling refers to the maximum pooling of sequence features obtained from two music-representation branches along the time axis to obtain the overall music feature vector used

Table 2. Performance Comparison of Annotation Using Different Feature Combinations in Music Sequence Modeling

Performance	MEL		WAV		MEL+WAV		CORR		Method of This Paper	
Top *N* Tags	50	188 (All)	50	188 (All)	50	188 (All)	50	188 (All)	50	188 (All)
AUC-tag	0.909	0.898	0.894	0.878	0.910	0.898	0.907	0.893	0.913	0.900
AUC-clip	0.942	0.960	0.929	0.951	0.943	0.961	0.939	0.957	0.945	0.962
MAP-tag	0.449	0.203	0.416	0.184	0.451	0.210	0.445	0.202	0.458	0.212
MAP-clip	0.710	0.624	0.684	0.597	0.715	0.631	0.708	0.622	0.720	0.635

Table 3. Comparison of Annotation Performance Based on Different Sequential Feature Aggregation Mechanisms

Performance	Max-Pooling		Single Atten. Weighting		Multi. Atten. Weighting	
Top *N* tags	50	188 (All)	50	188 (All)	50	188 (All)
AUC-tag	0.876	0.848	0.910	0.895	0.913	0.900
AUC-clip	0.919	0.944	0.942	0.960	0.945	0.962
MAP-tag	0.370	0.151	0.449	0.206	0.458	0.212
MAP-clip	0.644	0.555	0.712	0.627	0.720	0.635

to predict text tags; single atten. weighting (Shen et al., 2006) refers to the calculation of only one attention weight vector.

It can be seen from Table 3 that the method proposed in this study applies the feature aggregation method of multigroup attention weight, which improves the music-annotation performance of all evaluation indexes, reflects that the multi-weight attention method can obtain more feature representation of music, and ensures the effectiveness of the research method in music annotation (Casey et al., 2008).

Comparing the Performance of Related Research Methods

This study compares the proposed music-annotation method with some other methods on the MTAT dataset; the comparison results are shown in Table 4 and Fig. 7.

According to the results in Table 4 and Fig. 7, the AUC tag score of the method proposed in this study is the highest, which is 0.913. It reflects that the overall model architecture combining two kinds of music representation and music representation learning hierarchy sequence modeling is effective for the automatic music-annotation task.

Among the comparison methods, the event loc method is the closest to this study in research. Therefore, the method proposed in this study is more deeply compared with event LOC. In multi-index, the results are shown in Table 5 and Fig. 8.

It can be seen from the results in Table 5 and Fig. 8 that the proposed automatic music-annotation method can better model the mapping relationship between the audio feature input of music and the text label, and the scores of all evaluation indexes are higher.

Table 4. Comparison of the Labeling Performance of the Proposed Method and Other Methods on the MTAT Dataset

Method	AUC-tag
PSMC-MTSL (Hamel et al., 2011)	0.872
Transfer Learning (He et al., 2020)	0.879
FCN-4 (Ortego et al., 2020)	0.883
Event Loc. (Wang & Wang, 2014)	0.885
MSFL-MLP (Rathbun et al., 1997)	0.887
Time-Frequency CNN (Li et al., 2020)	0.897
SampleCNN (Steppa & Holch, 2019)	0.905
ReSE (Kim et al., 2018)	0.909
Method of this paper	0.913

Figure 7. Labeling Performance Comparison of Proposed and Latest Methods on MTAT Dataset

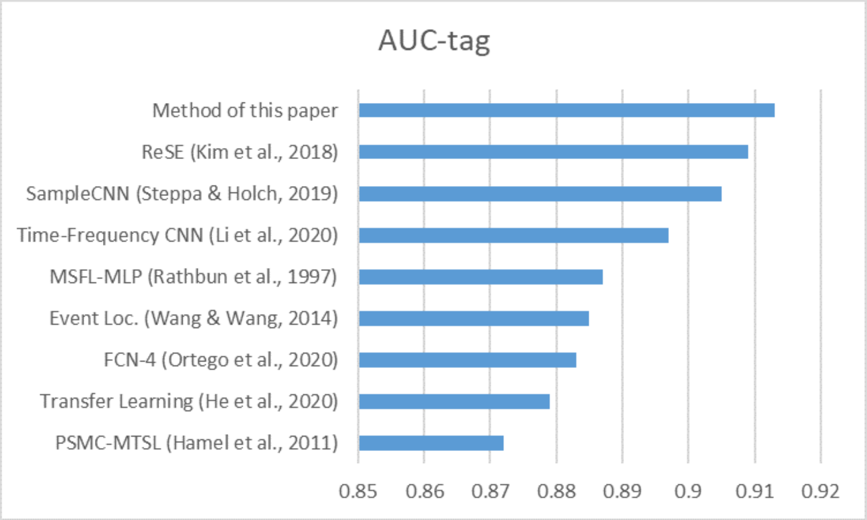
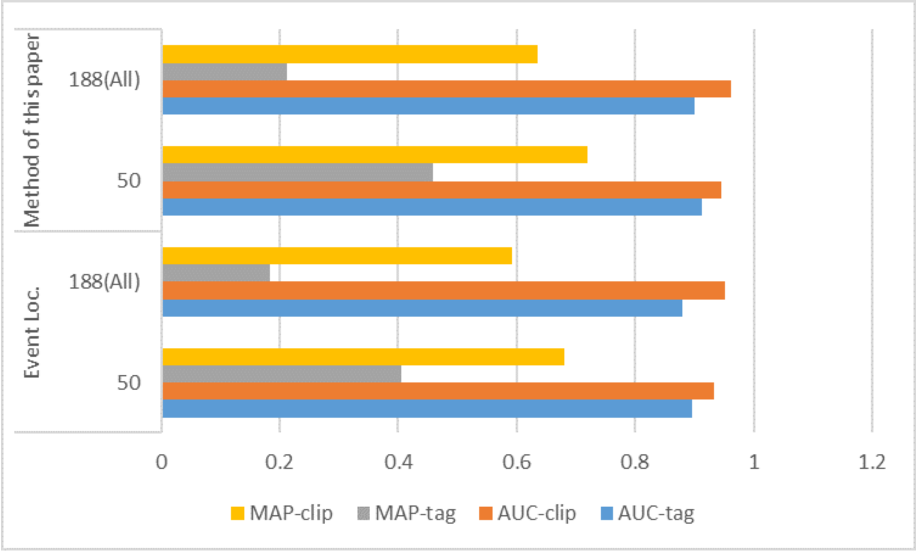


Table 5. Performance of the Proposed Method Compared With That of Event Loc in MTAT Dataset

Performance	Event Loc		Method of This Paper	
Top N tags	50	188(All)	50	188(All)
AUC-tag	0.896	0.880	0.913	0.900
AUC-clip	0.932	0.952	0.945	0.962
MAP-tag	0.406	0.183	0.458	0.212
MAP-clip	0.680	0.592	0.720	0.635

Figure 8. Comparative Analysis of Proposed Method and Event Loc Performance on MTAT Dataset



CONCLUSION

This study focuses primarily on addressing the retrieval challenges in folk music using an automatic-annotation approach. It involves constructing a deep neural network model for music annotation that integrates various musical expressions and attention mechanisms, followed by model verification. The findings indicate that the proposed method in the GLU module offers finer granularity and is better suited to the overall annotation model, leading to improved music-annotation performance across all evaluation indices. This ensures the effectiveness of hierarchical sequence modeling in music annotation. In comparison to other algorithms, the proposed method achieves the highest AUC tag score at 0.913. It demonstrates superior capability in modeling the mapping relationship between audio features of music input and text tags, as evidenced by higher scores across all evaluation indicators. Due to time and resource constraints, only the output of the final GLU module is utilized as a feature sequence in the subsequent network level within the music-representation aspect of the model. Subsequent testing is anticipated to enhance the performance of this component.

AUTHOR NOTE

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

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest.

FUNDING

This work was not supported by any funds.

CONTRIBUTIONS

The author sincerely thanks those who have contributed to this research.

PROCESS DATES

Received: 11/21/2023, Revision: 1/29/2024, Accepted: 2/6/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Lei Liu (15591165551@163.com)

REFERENCES

Aigrain, P., Zhang, H., & Petkovic, D. (1996). Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications*, *3*(3), 179–202. doi:10.1007/BF00393937

Ajoodha, R., Klein, R., & Rosman, B. (2015). Single-labelled music genre classification using content-based features. In *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)* (pp. 66–71). IEEE. doi:10.1109/RoboMech.2015.7359500

Bag, S., Kumar, S. K., & Tiwari, M. K. (2019). An efficient recommendation generation using relevant Jaccard similarity. *Information Sciences*, *483*, 53–64. doi:10.1016/j.ins.2019.01.023

Bisharad, D., & Laskar, R. H. (2019). Music genre recognition using convolutional recurrent neural network architecture. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, *36*(4), e12429. doi:10.1111/exsy.12429

Borjian, N., Kabir, E., Seyedin, S., & Masehian, E. (2018). A query-by-example music retrieval system using feature and decision fusion. *Multimedia Tools and Applications*, *77*(5), 6165–6189. doi:10.1007/s11042-017-4524-1

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696. doi:10.1109/JPROC.2008.916370

Foote, J. T. (1997). Content-based retrieval of music and audio. In C. C. J. Kuo, S. F. Chang, & V. N. Gudivada (Eds.), *Multimedia storage and archiving systems II* (Vol. 3229, pp. 138–147). SPIE., doi:10.1117/12.290336

Ghosal, D., & Kolekar, M. H. (2018). Music genre recognition using deep neural networks and transfer learning. In *Interspeech* (pp. 2087–2091). ISCA., doi:10.21437/Interspeech.2018-2045

Hamel, P., Lemieux, S., Bengio, Y., & Eck, D. (2011). Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In A. Klapuri, & C. Leider (Eds.), *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 729–734). The Printing House.

He, H., Khoshelham, K., & Fraser, C. (2020). A multiclass TrAdaBoost transfer learning algorithm for the classification of mobile lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *166*, 118–127. doi:10.1016/j.isprsjprs.2020.05.010

Hirai, T., Doi, H., & Morishima, S. (2018). Latent topic similarity for music retrieval and its application to a system that supports DJ performance. *Journal of Information Processing*, *26*(0), 276–284. doi:10.2197/ipsjjip.26.276

Kim, T., Lee, J., & Nam, J. (2018). Sample-level CNN architectures for music auto-tagging using raw waveforms. In *2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 366–370). IEEE. doi:10.1109/ICASSP.2018.8462046

Li, P., Yuan, H., Wang, Y., & Chen, X. (2020). Pumping unit fault analysis method based on wavelet transform time-frequency diagram and CNN. *International Core Journal of Engineering*, *6*(1), 182–188. doi:10.6919/ICJE.202001_6(1).0026

Li, T., & Ogihara, M. (2004). Content-based music similarity search and emotion detection. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. 5, pp. V–705). IEEE. doi:10.1109/ICASSP.2004.1327208

Manco, I., Benetos, E., Quinton, E., & Fazekas, G. (2021). MusCaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. doi:10.48550/arXiv.2104.11984

Miao, Z., & Cheng, C. (2023). Construction of multimodal music automatic annotation model based on neural network algorithm. In S. Patnaik (Ed.), *Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023)* (Vol. 12779, pp. 482–488). SPIE. doi:10.1117/12.2689482

Müller, M., Arzt, A., Balke, S., Dorfer, M., & Widmer, G. (2019). Cross-modal music retrieval and applications: An overview of key methodologies. *IEEE Signal Processing Magazine*, *36*(1), 52–62. doi:10.1109/MSP.2018.2868887

Ortego, P., Diez-Olivan, A., Del Ser, J., Veiga, F., Penalva, M., & Sierra, B. (2020). Evolutionary LSTM-FCN networks for pattern classification in industrial processes. *Swarm and Evolutionary Computation*, *54*, 100650. doi:10.1016/j.swevo.2020.100650

Raposo, F., Ribeiro, R., & Martins de Matos, D. (2016). Using generic summarization to improve music information retrieval tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(6), 1119–1128. doi:10.1109/TASLP.2016.2541299

Rathbun, T. F., Rogers, S. K., DeSimio, M. P., & Oxley, M. E. (1997). MLP iterative construction algorithm. *Neurocomputing*, *17*(3–4), 195–216. doi:10.1016/S0925-2312(97)00054-4

Sanchez, S. A., Romero, H. J., & Morales, A. D. (2020). A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. In IOP Conference Series: Materials Science and Engineering (Vol. 844, No. 1, 012024). IOP Publishing. doi:10.1088/1757-899X/844/1/012024

Seo, J. S. (2018). A local feature aggregation method for music retrieval. *IEICE Transactions on Information and Systems, E101.D*(1), 64–67. 10.1587/transinf.2017MUL0001

Shen, J., Shepherd, J., & Ngu, A. H. H. (2006). Towards effective content-based music retrieval with multiple acoustic feature combination. *IEEE Transactions on Multimedia*, *8*(6), 1179–1189. doi:10.1109/TMM.2006.884618

Shen, J., Tao, M., Qu, Q., Tao, D., & Rui, Y. (2019). Toward efficient indexing structure for scalable content-based music retrieval. *Multimedia Systems*, *25*(6), 639–653. doi:10.1007/s00530-019-00613-z

Shimizu, W., & Kanno, Y. (2018). Accelerated proximal gradient method for elastoplastic analysis with von Mises yield criterion. *Japan Journal of Industrial and Applied Mathematics*, *35*(1), 1–32. doi:10.1007/s13160-017-0280-x

Singh, P., Hrisheekesha, P. N., & Singh, V. K. (2021). CBIR-CNN: Content-based image retrieval on celebrity data using deep convolution neural network. *Recent Advances in Computer Science and Communications*, *14*(1), 257–272. doi:10.2174/2666255813666200129111928

Sotiropoulos, D. N., Lampropoulos, A. S., & Tsihrintzis, G. A. (2008). MUSIPER: A system for modeling music similarity perception based on objective feature subset selection. *User Modeling and User-Adapted Interaction*, *18*(4), 315–348. doi:10.1007/s11257-007-9035-8

Steppa, C., & Holch, T. L. (2019). HexagDLy—Processing hexagonally sampled data with CNNs in PyTorch. *SoftwareX*, *9*, 193–198. doi:10.1016/j.softx.2019.02.010

Wang, Q., Su, F., & Wang, Y. (2020). Hierarchical attentive deep neural networks for semantic music annotation through multiple music representations. *International Journal of Multimedia Information Retrieval*, *9*(1), 3–16. doi:10.1007/s13735-019-00186-7

Wang, X., & Wang, Y. (2014). Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 627–636). ACM. doi:10.1145/2647868.2654940

Xie, Z., Liu, H., Li, Z., & He, Y. (2018). A convolutional neural network based approach towards real-time hard hat detection. In *2018 IEEE International Conference on Progress in Informatics and Computing (PIC)* (pp. 430–434). IEEE. doi:10.1109/PIC.2018.8706269

Xu, Z. (2022). Construction of intelligent recognition and learning education platform of national music genre under deep learning. *Frontiers in Psychology*, *13*, 843427. doi:10.3389/fpsyg.2022.843427 PMID:35693513

Yu, Y., Tang, S., Raposo, F., & Chen, L. (2019). Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Transactions on Multimedia Computing Communications and Applications*, *15*(1), 1–16. doi:10.1145/3281746

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T., & Fang, W. (2020). Deep learning-based extraction of construction procedural constraints from construction regulations. *Advanced Engineering Informatics*, *43*, 101003. doi:10.1016/j.aei.2019.101003