Research Article

Musical Note Feature Recognition Based on Long Short-Term Memory

Yuan Yuan

Henan Institute of Science and Technology, Xinxiang, Henan 453003, China

Correspondence should be addressed to Yuan Yuan; 1208010132@stu.zuel.edu.cn

Received 6 June 2022; Revised 16 July 2022; Accepted 20 July 2022; Published 12 August 2022

Academic Editor: Liping Zhang

Copyright © 2022 Yuan Yuan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The timbre and volume of a single tone are among its fundamental characteristics. The single-tone detection technology is the key to the foundation of MNFR (musical note feature recognition), which is built on the fundamental feature extraction of single tones. A MNFR method based on LSTM (long short-term memory) is proposed because traditional methods have low accuracy in note feature classification and low accuracy in MNFR. To process the series of convolution feature maps, the feature maps are directly input into LSTM to learn hash codes. Extract the note features. Segment according to the changing trend of the physical features of the notes. Additionally, a number of feature maps are built from convolution feature maps extracted from numerous convolution layers of previously trained CNNs (convolutional neural networks), taking into account the spatial specifics and semantic features. The note start vector is produced using the enhanced peak extraction algorithm based on Gaussian kernel smoothing. The findings indicate that, when 100 samples are used, this method’s note classification accuracy differs from that of the DBN (deep belief network) and DWT (discrete wavelet transform) by 1.17 percent and 2.04 percent, respectively. The analysis in the conclusion demonstrates that the algorithm put forth in this paper is both theoretically and practically workable.

1. Introduction

MNFR (musical note feature recognition) is a general term covering many tasks, such as classification, recognition, audio stream segmentation, data retrieval, and content analysis of audio files containing music. Essentially, it is content-based audio recognition and processing, with a high degree of gender complexity. In the field of music creation, many composers use some note feature speech recognition software to automatically recognize the features of notes. Notes are used not only for learning music, but also for searching and classifying songs according to the recognition of notes, which promotes the diversified development of music [1]. If they can automatically recognize the music they play by computer and finish the music score creation automatically, it will greatly improve their creative efficiency, greatly stimulate creative inspiration, and change the inconvenience of writing music scores by hand at present.

MNFR is a branch of speech recognition. With the research of artificial intelligence becoming more and more mature, the application of intelligent technology is getting closer and closer to public life. Lai et al. proposed a framework covering the stages of feature extraction, feature selection, and classification, so that new features can be easily combined or new music styles can be tested [2]. The framework includes different classification methods, such as Bayesian classifier, nearest neighbor, and self-organizing graph. Alicia et al. think that the low-level network of deep neural network can extract features similar to speakers, while the high-level network can extract discrimination information between categories [3]. Li et al. used CNN (convective neural network) to make content-based music recommendation, realized the prediction of listeners’ listening preferences based on audio signals, and then used WMF (weighted matrix factorization) model to make score prediction [4]. Zhao et al. proposed a main melody recognition method based on improved Euclidean algorithm and dynamic programming [5]. In this method, the candidate pitch of each frame is estimated by
the improved Euclidean algorithm, and then the significance and sequential continuity of melody pitch are analyzed under the framework of dynamic programming. The traditional intelligent musical note segmentation method is based on the time discrete Fourier transform method to realize the musical note segmentation recognition function, but this method has low sensitivity to musical tone recognition and low recognition efficiency, and is prone to recognition errors.

Although musical notes can be recognized, it is by no means a simple problem. It involves multi-disciplinary support, such as music, psychology, computer, mathematics, and signal processing, which is the intersection of multi-disciplines. Traditional music information retrieval usually uses text retrieval technology to process music information. Music is manually labeled, such as song title, singer, and composer, and then searched by text retrieval technology. Compared with the traditional hash learning method, the deep hash learning method shows better results in some aspects. Therefore, this paper studies the MNFR method based on LSTM (long short-term memory). Compared with the classical algorithm, it can directly extract the acoustic features or musical features of music, get the recognition and classification results through classifier training, and improve the classification accuracy of MNFR.

Innovation of the paper:

(1) In this paper, the characteristics of note tones are used as the identification marks of notes, so as to improve the sensitivity and ability of tone identification and promote the development of musical note field.

(2) This paper presents a MNFR method based on LSTM. The vector of audio frame after STFT transformation is used as input, the drama category of the frame is used as training label to train LSTM, and the trained model is used to extract context-related features.

The content of this article is arranged as follows:

The first chapter introduces the research background and significance, and then introduces the main work of this paper. The second chapter mainly introduces the related technologies of MNFR. The third chapter puts forward the concrete methods and implementation of this research. The fourth chapter verifies the superiority and feasibility of this research model. The fifth chapter is the summary of the full text.

2. Related Work

2.1. MNFR-Related Research. Nazemi et al. comprehensively analyzed the whole process of note starting point detection, including preprocessing, signal reduction, and peak extraction [6]. The methods used in each process are analyzed in the literature, and the experimental results and evaluation are given. Generally, the research process of note starting point detection is summarized into three parts: preprocessing, signal reduction, and peak extraction. Wang et al. used AdaBoost algorithm to select a set of features from the audio feature set and then classified the music genres by aggregation algorithm [7]. Engin et al. used modulation spectrum analysis to capture time-varying information and prosody information in music signals [8]. Strisciuglio et al. used the adaptive harmony search algorithm to select the low-dimensional feature subset with the strongest correlation from the feature set, which significantly improved the accuracy of music genre classification [9].

Lee et al. used stack self-coding network for feature coding and compressed the data to a preset length with minimum reconstruction error [10]. Bibi et al. have explored various DL (deep learning) frameworks in speech emotion tasks, and their experiments have demonstrated that feedforward and RNN (recurrent neural network) structures and their variants can be used to assist speech recognition, especially emotion recognition [11]. Fornaser et al. put forward a method of identifying isolated sound events by using DBN (deep belief network). For the acoustic event classification task with 61 different categories, the classification accuracy of neural network classifier is better than that of traditional mixed model classifier [12]. Badi et al. used DWT (discrete wavelet transform) to decompose the music signal [13].

2.2. Research Status of LSTM. Different from the basic structure of traditional RNN, the hidden layer of LSTM, that is, the circulating layer, no longer simply uses activation functions to control information, but introduces the state of cells. The memory uses the “gate” structure to control the state and output at different times. Especially, the application of understanding and optimizing the core structure of LSTM-driven cell structure and cell state structure will be of great significance and value.

Geng established a new model of tomato target yield prediction based on LSTM, which proved that LSTM had high accuracy in predicting tomato target yield [14]. Yan et al. used LSTM to construct an end-to-end neural network framework for machine translation and introduced local attention mechanism into the model to improve translation quality [15]. Li et al. used the reinforcement learning actor-critic training network to evaluate the value of the output notes of LSTM network, so as to update the generation strategy of LSTM network, and the generated music has a stable structure and more style [16]. Saqib et al. conducted two experiments on two-way LSTM and one-way LSTM on speech corpus, and found that two-way LSTM was superior to one-way LSTM and conventional RNN [17].

Lu et al. established an end-to-end handwriting recognition model based on attention. Different from previous studies, this system can learn the reading order and process two-way characters without dividing the data into lines in advance [18]. Lin et al. stacked LSTM network by stacking different hidden layers with LSTM units in space to increase network capacity [19]. Hong et al. introduced deep neural network into speech recognition technology and directly replaced the Gaussian mixture distribution model with neural network [20].
3. Methodology

3.1. Feature Extraction and Analysis of Musical Note Intelligent Segmentation. Music is a fixed-pitch sound created by the regular vibration of sound-producing objects. It is the most crucial and fundamental component of tonal music. Tones make up melody and harmony in music. Tone and rhythm are complimentary organic elements. A piece of music cannot have rhythm without another piece of rhythm. Similar to this, rhythm cannot exist without rhythm. They are integral components. The processing efficiency decreases as the amount of music data stored and processed increases. The signal digitization process will result in significant waveform distortion if the data accuracy and sampling rate are too low, so 16-bit accuracy and 22,050 Hz sampling rate can be chosen.

It is found that the energy of low frequency band of speech signal is larger, while the energy of high-frequency signal is obviously smaller. The purpose of speech pre-emphasis is to enhance the high-frequency part of speech and flatten the signal spectrum, and at the same time, it can eliminate the influence of radiation from lips and improve the resolution of speech in the high-frequency part. Speech signal can be regarded as a short-term static signal. The research shows that the spectrum characteristics of speech signal can be considered as basically unchanged within 10~30 ms. Therefore, the voice signal can be divided into many short periods of equal length, each short period is called a frame, and the short-term voice signal is processed in each frame by the stationary signal investigation method. There needs to be a certain frame offset (overlap) between the previous frame and the next frame, so that the transition between the previous frame and the next frame can be smooth. In this paper, frame offset is 1/2.

After determining the frame length and the corresponding frame shifting parameters, the specific framing process is completed by windowing. The windowing process includes multiplying the signal by a window function with a finite length and a fixed shape, and then the window moves in the audio according to the frame changes. In this experiment, the audio frame length is 25 ms, the slope is 10 ms, and the Hamming window, which is commonly used in speech and audio processing, is selected as the window function. \( N \) is used to represent the window length, and the function expression is as follows:

\[
w(n) = \begin{cases} 
0.54 - 0.46 \cos \left( \frac{2\pi}{N-1} n \right), & 0 \leq n \leq (N-1), \\
0, & \text{otherwise}.
\end{cases}
\]  

(1)

The audio signal processed by framing and windowing can already be used as the simplest input, but the effect is generally not very good due to too little explicit signal extraction, so there are more advanced input processing methods below.

In our task, the dispersion coefficient obtained by the second-order dispersion transformation represents the characteristics of each frequency carried by the audio signal. Compared with the features obtained by traditional methods, these features are more invariant to local translation. At the same time, because the performance of high-pass filtering is more stable, for this task, it is necessary to extract the characteristic information contained in the high-frequency signal. The research shows that iterative wavelet modulus operation can recover high-frequency information. Therefore, the high-frequency recovery is carried out through the following operations:

\[
|W_n^f| = \left( x \ast \varphi(t), \left| x \ast \psi_{f_0}(t) \right| \right)_{x \in \mathcal{A}}.
\]  

(2)

where \( |W_n^f| \) represents the first-order scattering transformation of \( x \), and \( |W_n| \) represents the first-order transformation. For audio signals, it is generally defined that wavelet has the same filter frequency as Mel spectrum. In order to make the wavelet modulus coefficient invariant to translation, the time average unit is used. Finally, the approximate Mel spectrum coefficient is obtained by averaging the wavelet modulus coefficient with \( \varphi(t) \).

Generally, peak extraction is to set a threshold value for the detection function curve onset \((n)\), and the point beyond the threshold value is taken as the note starting point. There are two ways to set the threshold; the first is to fix the threshold \( \delta \), as follows:

\[
\text{onset}(i) = \begin{cases} 
1, & \text{onset}(i) \geq \delta, \\
-1, & \text{onset}(i) < \delta.
\end{cases}
\]  

(3)

In this way, in the climax part, the sound intensity of the note is greater than that of the flat part. At this time, if the threshold of the note starting point is still fixed, the note starting point in the flat part will be ignored. Therefore, adaptive threshold should be used for threshold setting.

In this paper, PCP (pitch class profile) feature extraction method is used. The extraction method is based on pitch and pitch theory. Pitch represents the octave of a note, and pitch represents the treble of a note. Tone formation further classifies the note characteristics according to the harmonic information represented by the note to complete the refinement. The extraction process is shown in Figure 1.

Through PCP feature extraction method, the music melody is re-converted into a tone scale spectrum, and the notes are mapped on the 12-average-law mapping tone scale. Then, the notes on each tone level are divided into frames, and the overlapping frame signals of the notes are eliminated. The specific auxiliary formula is as follows:

\[
S = \sum_{n=1}^{m} x \ast (n - m) \ast w(m),
\]  

(4)

where \( x \) represents frequency coordinates; \( n \) represents the center of the STFT (short time Fourier transform) window; \( m \) represents the length of note data frame; and \( w(m) \) refers to Hamming window.

Suppose \( y_1(C_i) = (y_1(v_1), y_1(v_2), L, y_1(v_n)) \) is the threshold vector of note class \( C_i \); \( G = \langle V, E \rangle \) is a simple directed graph with \( n \) nodes, so the square matrix \( A(G) = (a_k) \) can be called the similarity matrix of \( G \). Among them,
At this time, $A(G)$ is a symmetric matrix, which is also a musical note similarity matrix. According to the musical note similarity matrix, a note feature selection standard is given, and this standard is used as the optimization objective function of musical note feature subset.

PCP features represent music signal frames with 12-dimensional feature vectors, which can be converted into sound-level spectrum by spectrum reconstruction. In the reconstruction process, harmonics are assigned to the corresponding minority values in a many-to-one way. Therefore, the characteristics of PCP have the important characteristics of energy compression. If STFT is performed, it corresponds to the frequency number of the same PCP sound level, and the STFT amplitude is accumulated on each PCP sound level.

$$PCP(p, m) = \sum_{g(k)=p} |STFT(k, m)|.$$ (6)
The characteristics of PCP are closely related to chords in music. When the chord changes, the PCP feature vector also changes. A chord change means that a new note is created. Therefore, the start of the note can be detected by the change of PCP feature vector.

3.2. Implementation of MNFR Method

3.2.1. Introduction of LSTM Model. The basic motivation put forward by LSTM is to protect the integrity of information transmission, and the error gradient is constant in the process of reverse transmission. In order to control the influence of information in memory at different times on the flow, some irrelevant information can be selectively “shielded.” Therefore, entering the output gate, the output gate is also controlled by sigmoid function, which generates a relationship to control the output relationship among all the information of the current state of the cell. The hidden layer unit structure of LSTM model is shown in Figure 2.

The input can be divided into two parts. The first part is to determine the new information to be added to the cell state, and the second part is to determine the proportion of this new information added to the memory cell state. Its input is the hidden layer state $h_{t-1}$ at the last moment and the input $x_t$ at the current moment, and its calculation formula is

$$
\tilde{C}_t = \text{Tanh} (W_c h_{t-1} + U_c x_t + b_c),
$$

where $W, U, b$ is the weight and offset.

The gate determines how much information the internal state at the current time $t$ outputs to the external state of the neuron. The output expression of the gate is shown in

$$
O_t = \delta (U_o h_{t-1} + W_o x_t + b_o).
$$

The nonlinear function of the gate also selects sigmoid function. The principle of information screening is the same as that of input gate and forgetting 17. When the value of $O_t$ approaches 1, more information will be output from the internal state $c_t$ of $t$, to the external state $h_t$ of $t$ at the current moment.

However, LSTM can effectively solve the problem of gradient disappearance in ordinary RNN training by optimizing the internal structure of neurons. Instead of bringing the gradient to the front because there is a vector 0 in the middle, the gradient disappears. In addition, because the network is learning the threshold value all the time, through gradient descent, the network will automatically adjust when the gradient should be attenuated and kept.

3.2.2. Algorithm Realization. Generalized MNFR covers all the elements of automatic pitch labeling, including monophonic note recognition, pitch estimation, multi-syllable beat and rhythm recognition, melody and harmony extraction, multi-frequency estimation-basis polyphony, and many other topics. Pitch data localization is mainly considered in initial note separation and pitch detection. Short-duration tone signals excel in many areas where continuous tone does not. Short-term analysis is another signal processing method that works better. The biggest problem is the inability to use correlation between different channels when making decisions; this is where feature fusion comes into play. The audio file is shared by all of the article’s features, and the use of frames in the window keeps all of the features’ data synchronized.

Although the task of this chapter is MNFR, it will first preprocess the music signal and convert it into a spectrum, so it can be considered as an image recognition problem. The feature map of each convolution layer uses bilinear interpolation and similarity selection strategy to form a feature map sequence, which is then input into LSTM and hash layer, and finally recognized and classified by softmax. The MNFRDL framework proposed in this chapter is shown in Figure 3.

Each column is input into the LSTM learning feature vector in a fixed order using the spatial structure of the convolution feature map as the input. The feature maps are directly input into LSTM to learn hash codes in order to process the series of convolution feature maps. Additionally, a set of feature maps are created using convolution feature maps that are taken from pretrained CNN’s multiple convolution layers, taking into account both semantic and spatial details. Finally, a new loss function is created to maintain the semantic balance and similarity of the basic hash code while controlling the quantization error output of the hash layer.

Pool layer can be considered as a down-sampling layer, which can not only reduce the parameters reasonably, but also reduce the over-fitting problem to improve the results. On the choice of activation function, the most commonly used one in CNN is ReLUs. It is defined as follows:

$$
f(x) = \max(0, x).
$$

Due to the introduction of semilinearity, ReLUs have more efficient calculation and more effective gradient propagation, as well as life probability and sparse activation structure, while keeping enough simplicity. Therefore, ReLUs are chosen as the activation function in the experiment.
In the performance detection algorithm, the stability of signal performance is different after the iteration times of pitch part and transient part are the same, so as to identify the transient part and mark the starting point of the signal. In this algorithm, the redundant dictionary adopts real function set:

\[ g_{c}(t) = \frac{K_r}{\sqrt{s}} g(t - u) \cos(\xi t + \phi). \]  

(10)

The window function \( g \) is Gaussian window, and \( K_r \) makes \( \|g_{c}\| = 1 \). The window adjusts the time and frequency resolution of time-frequency atoms by scaling factor \( s \), and the ability to adapt to signals is enhanced. The energy distribution of the window on the time-frequency plane is an ellipse with long and short axes on the time axis and frequency axis, and the scale factor \( s \) can adjust the proportion of the long and short axes of the ellipse.

The purpose of the training is to make the correct label path sequence score higher than other competitive sequences under the decoding condition of the existing model. Therefore, in this paper, the objective function is optimized and improved by the optimization criterion based on the minimum Bayes risk framework, and the optimized expression is shown in

\[ F = \sum_{r} \sum_{s} p(s|x_r) A(s, s_r), \]  

(11)

where \( A(s, s_r) \) is a measure of the accuracy of the identified result sequence \( s \) relative to the target sequence \( s_r \).

Superparameters are initial parameters that must be set before LSTM model starts training, rather than parameters that are constantly adjusted and optimized by learning datasets. The effective selection and optimization of superparameters have great influence on the whole training process and the expected results. The generalized gradient descent method is used to minimize the above objective function, as shown in

\[ \theta_{t,j} = \theta_{t-1} - \frac{\alpha}{\sqrt{\sum_{r=1}^{t} g_{r,j}}} g_{t,j}, \]  

(12)

where \( \alpha \) is the initial learning rate of the model and \( g_r \) is the descending gradient of the \( i \) th parameter \( \theta_{t,j} \) with a step size of \( r \). According to the ratio represented by this parameter, some hidden neurons and their corresponding input and output weight parameters are randomly removed, and the propagation and gradient change of the neural network are effectively maintained on the premise of ensuring efficiency and accuracy.

4. Experiment and Results

In order to verify the comprehensive effectiveness of the proposed MNFR method based on LSTM, the following experiments are conducted: the experimental environment is Intel Core 8-460 24G memory, and the operating system is Windows 7. The methods of variable feature recognition and extraction are compared. The experimental results are shown in Table 1.

It can be seen that with the increase of the number of samples, the classification accuracy of the three methods corresponding to notes has changed. When the number of samples is 100, the difference between our method and DBN and DWT is 1.17% and 2.04%, respectively. Experimental results show that this method has the highest classification accuracy.

The pitch and starting point components make up the majority of music. When adjacent signal frames are attenuated in the pitch part, the frequency and amplitude change gradually, making the atomic ensemble and its energy sum relatively stable. As a result, the interpretive degree tends to be stable. It is important to note that the goal of this partials recognition is to recognize partials at specific frequency points rather than to identify the specific instrument (or human voice) that produces them. Let the atoms and...
frequencies of partials correspond to the auditory characteristics, which is more in line with the characteristics of music itself. The participating jitter functions are thus extracted in the matching mode.

In order to ensure the rationality and scientificity of this experiment, this paper designs DBN method and DWT recognition method as the traditional comparison methods of the experiment, and divides and recognizes the same music melody according to three methods. The experimental results of time series identification are shown in Figure 4.

It can be seen that the intelligent segmentation and note recognition time of different methods are different. When the recognition amount is 2 GB, the intelligent note segmentation recognition time of DBN recognition method is 0.021 min, that of DWT recognition method is 0.016 min, and that of this method is 0.0036 min. The recognition time of this method is much lower than that of the other two traditional methods, because it can segment notes according to the changing trend of their physical characteristics, effectively extract the features of notes, and shorten the recognition time.

Based on the original training data set of 120-hour speech data, this paper combines the rate interference technology. The speed of the experimental data speech is 1.2 times and 0.8 times that of the original speech, respectively. In this way, without adding new voice data, the amount of training data is increased by 2.5 times in this experiment, so that in practical application, if the data are missing, the model can be fully trained. Performance comparison of MNFR model based on LSTM is shown in Table 2.

In this paper, the model trained on the original data set greatly reduces the word error rate in the test set. The original data set only has a history of 120 hours, and the LSTM-based model has achieved a high accuracy, which proves the powerful function of this model. After the training data are further expanded by the rate perturbation technique, the word error rate of the model decreases again. Because the end-to-end training method requires high training data, the amount of training data currently used in this experiment is relatively small even if the rate is disturbed, so the model potential of this method has not been fully exerted.

The reliability of note detection results and the accuracy of fundamental frequency calculation are the two factors that affect how well an identification system performs. The former only involves the fundamental frequency extraction algorithm, whereas the latter is the key component, the challenge, and the means by which the algorithm can be improved. The only operations that can be obtained from the analysis window to the characteristic window are the mean and variance operations. It is the same as merging into the fused features after first calculating the mean square error in the analysis window, as opposed to the reverse. In order to avoid the direct inner product operation in high-dimensional space, the kernel function is developed to calculate in low-dimensional space and express the classification effect in that space.

When the music in the data set is divided into frames, the frame length is 20 ms, and the frame overlap rate is 1/2. The following shows the simulation results of the energy estimation method based on the coefficient vector of the music piece signal. Figure 5 shows the influence of LSTM decomposition times on the coefficient vector energy estimation algorithm.

It can be seen that the starting F value of note start detection gradually increases with the increase of attenuation times. When the attenuation times are more than 60 times, the matters needing attention in detecting F value basically remain unchanged. This is because when the number of decomposition reaches a certain value, the residual energy of the signal becomes smaller and smaller, whether it is the algorithm of expressing the degree or the algorithm of estimating the vector energy coefficient, and the residual energy of each time becomes smaller and smaller. The influence on the total energy of the whole signal may even be influenced by external noise, resulting in a slight decrease in F value.

The fundamental frequency trace of the signal can reflect two characteristics of pitch and length. For the recognition task discussed in this paper, the fundamental frequency as the feature vector of recognition can roughly meet the needs. However, when two notes with the same pitch appear continuously, only one note can be represented on the pitch frequency diagram; because of the particularity of music signal and the average process of codebook training, canceling smoothing may better meet the requirements of music signal. Figure 6 shows MNFR rates for piano, violin, and oboe.

The experimental results show that LSTM can solve the basic problems of MNFR well. Compared with the recognition rate mentioned above, the system performance has
improved a lot. This indicates that MNFR with LSTM is effective. Save the pitch, and then delete the sound from the mix. The same is true of other sound signals. In this way, the dominant fundamental frequency can be estimated. Using prior knowledge to constrain the recognition results is an inevitable way to further improve the recognition rate. In this paper, eight kinds of Chinese traditional operas are classified and studied. However, the ancient Chinese operas come down in one continuous line, and all kinds of operas learn from each other and develop with each other, forming a prosperous classical opera performance. Figure 7 shows the classification accuracy of any two types of operas on each feature.

In classifying eight operas, it can be seen that the time context function works best, though there are some instances where the octave spectrum contrast, normalized spectrum envelope, and pitch family function work better. The relative distribution of spectral peaks and troughs in each sub-band is most clearly highlighted in the octave spectrum, which also has the best contrast effect. As a result, it is easier to see how the peaks and valleys of the opera audio signals are distributed. It is generally not appropriate to use accuracy or root mean square error as a measurement standard in labeled data sets because many labeled dimensions may be 0. The performance of hashing methods, which have two main benefits, is therefore evaluated in this experiment using three commonly used indicators. The first one is fairly resilient to unbalanced data sets, and the second one uses a set of straightforward numbers as its description indicator. MAP (mean average precision): The similarity score of a pair of images can be calculated by calculating the Hamming distance between the test sample and the training sample using the learned binary semantic features. Precision@$k$ is the proportion of accurate results in the image with the closest match to the test image among the first $k$ images. The percentage of accurate results when the Hamming distance between the test sample and the training sample is less than 2 is known as HAM2 (Hamming distance less than 2). The recognition results of feature maps of different sizes are shown in Figure 8.

It can be seen that when the size of the feature map is $6 \times 6$, a better recognition result can be obtained. This is because the intelligent note segmentation recognition method studied in this paper preprocesses the note data of the notes to be segmented before cutting and removes the redundant tones in the melody without breakpoint conditions. The intelligent note segmentation recognition method based on audio feature technology uses layered segmentation and filtering to complete the calculation of a melody, which reduces the task of note segmentation and improves the efficiency of note segmentation recognition.

The recognition process includes estimating various states experienced according to the feature stream and obtaining the best state sequence, thus obtaining the note sequence. Considering multiple candidates, the search process should return candidates with different lengths, corresponding to different note detection results, and finally get the recognition results through multi-candidate decision. It can be said that selecting reasonable features can get twice the result with half the effort, and a simple classifier can also
achieve good results; if an invalid or even confusing feature is selected, the result will obviously be poor. Functional integrity ensures functional usability; that is, functions can not only distinguish audio, but also classify audio into the correct category. In short, the reliability of features guarantees the accuracy of features.

5. Conclusions

After years of development, MNFR technology development system has become more and more perfect. On the basis of single-tone extraction and recognition, the extraction and recognition techniques of single-tone melody and multi-part polyphony are developed, including timbre, rhythm, speed, volume, and harmony. Aiming at the shortcomings of the traditional MNFR method, a MNFR method based on LSTM is proposed. According to the recognition results of fused features under different classifiers, Gaussian kernel smoothing algorithm is applied to smooth the curve of the detection function, then the moving window of the detection function is normalized, and the threshold and the starting point detection function are set. Experiments show that the depth network mentioned above is more suitable for music recognition technology. The application of DL in MNFR technology needs to combine some domain knowledge in computer vision, bringing new research directions and key research points for the application of DL in MNFR technology.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] D. Wang, “Recognition and error correction techniques for piano playing music based on convolutional cyclic hashing method,” Wireless Communications and Mobile Computing, vol. 2022, pp. 1–11, 2022.
[2] C. F. Lai, W. C. Chien, L. T. Yang, and W. Qiang, “LSTM and edge computing for big data feature recognition of industrial electrical equipment,” IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2469–2477, 2019.
[3] R. Zazo and D. T. Toledano, “An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition,” PLoS One, vol. 12, no. 8, Article ID e0182580, 2017.
[4] S. Li, H. Zhang, Y. Shi, and J. Yang, “Novel local coding algorithm for finger multimodal feature description and recognition,” Sensors, vol. 19, no. 9, p. 2213, 2019.
[5] S. Zhao and B. Zhang, “Joint deep convolutional feature representation for hyperspectral palmprint recognition,” Information Sciences, vol. 489, pp. 167–181, 2019.
[6] A. Nazemi, Z. Azimifar, M. J. Shafiee, and W. Alexander, “Real-time vehicle make and model recognition using unsupervised feature learning,” in Proceedings of the IEEE Transactions on Intelligent Transportation Systems, vol. 99, pp. 1–11, Paris, July 2019.
[7] J. T. Wang, G. L. Yan, H. Y. Wang, and J. Hua, “Pedestrian recognition in multi-camera networks based on deep transfer learning and feature visualization,” Neurocomputing, vol. 316, pp. 166–177, 2018.
[8] M. A. Engin, S. Aras, and A. Gangal, “Extraction of low-dimensional features for single-channel common lung sound classification,” Medical & Biological Engineering & Computing, vol. 60, no. 6, pp. 1555–1568, 2022.
[9] N. Strisciuglio, M. Vento, and N. Petkov, “Learning representations of sound using trainable COPE feature extractors,” Pattern Recognition, vol. 92, pp. 25–36, 2019.
[10] S. Lee, H. Eun, S. Oh, and C. Kim, “Landmark-free clothes recognition with a two-branch feature selective network,” Electronics Letters, vol. 55, no. 13, pp. 745–747, 2019.
[11] S. Bibi, N. Anjum, and M. Sher, “Automated Multi-feature human interaction recognition in complex environment,” Computers in Industry, vol. 99, pp. 282–293, 2018.
[12] A. Fornaser, T. Mizumoto, H. Suwa, and M. D. YasumotoCecco, “The influence of measurements and feature types in automatic micro-behavior recognition in meal preparation,” in Proceedings of the IEEE Instrumentation and Measurement Magazine, vol. 21, no. 6, pp. 10–14, London, December 2018.

[13] H. Badi, S. H. Hussein, and S. A. Kareem, “Retraction Note to: feature extraction and ML techniques for static gesture recognition,” Neural Computing & Applications, vol. 28, no. 4, p. 851, 2017.

[14] B. Geng, “Text segmentation for patent claim simplification via bidirectional long-short term memory and conditional random field,” Computational Intelligence, vol. 38, no. 1, pp. 205–215, 2022.

[15] W. Yan, J. Wang, J. Cheng, and K. WanXingGao, “Long short-term memory networks and bayesian optimization for predicting the time-weighted average pressure of shield supporting cycles,” Geofluids, vol. 2021, no. 4, Article ID 8895844, 2021.

[16] G. Li, X. Wang, R. Hu, and S. ZhangKe, “Intelligibility enhancement via normal-to-lombard speech conversion with long short-term memory network and bayesian Gaussian mixture model,” in Proceedings of the IEEE Transactions on Multimedia, vol. 23, pp. 3035–3047, New York, March 2021.

[17] M. Saqib, E. Şentürk, S. A. Sahu, and M. A. Adil, “Comparisons of autoregressive integrated moving average (ARIMA) and long short term memory (LSTM) network models for ionospheric anomalies detection: a study on Haiti (Mw = 7.0) earthquake,” Acta Geodaetica et Geophysica, vol. 57, no. 1, pp. 195–213, 2022.

[18] X. Lu, Y. H. Sha, Z. Li, and J. C. HuangChenChenShenChenFung, “Development and application of a hybrid long-short term memory - three dimensional variational technique for the improvement of PM2.5 forecasting,” Science of the Total Environment, vol. 770, no. 574, Article ID 144221, 2021.

[19] L. Lin, S. Gong, S. Peeta, and X. Wu, “Long short-term memory-based human-driven vehicle longitudinal trajectory prediction in a connected and autonomous vehicle environment,” Transportation Research Record: Journal of the Transportation Research Board, vol. 2675, no. 6, pp. 380–390, 2021.

[20] H. Hong, H. Jeon, C. Youn, and H. Kim, “Incorporation of shipping activity data in recurrent neural networks and long short-term memory models to improve air quality predictions around busan port,” Atmosphere, vol. 12, no. 9, p. 1172, 2021.