Research on the Construction of Music Performance Robot Based on Beat Recognition

Zhang Ting

Abstract

This paper addresses the issue of music performance robots frequently misjudging complex rhythmic patterns, leading to abnormal performance movements. Focusing on rhythm recognition and gesture recognition, we propose an adaptive rhythm identification model based on bidirectional recurrent neural networks and gesture recognition technology for a performing robot system. Experimental results demonstrate the model's superior performance in robot dance performances with complex musical rhythms. Specifically, the bidirectional recurrent neural network adaptive model enhances the accuracy of rhythm recognition to over 90%, while gesture recognition technology matches the rhythm identification results to dance movements through complete information transformation, which are then precisely presented by the robot. Ultimately, experiments confirm that the recognition system meets the design requirements for music performance robots.

Keywords Rhythm Recognition; Bidirectional Recurrent Neural Networks; Gesture Recognition

1. Introduction

At present, the recognition of music by intelligent performance robots can be divided into two aspects: music style and music beat, but the robots have more misjudgment problems in the recognition of complex music beats, and the misjudgment is mainly due to the recognition system's fuzzy classification of beat feature extraction, a single recognition model, and incomplete information transfer. To address the above problems, Wang Li et al. proposed a combined data preprocessing scheme of short-time Fourier transform combined with Mel inverted spectrum through the analysis of audio features to further clarify the results of beat data feature extraction and classification [1]; Meng Zhen et al. proposed a Mel spectrum and power spectrum fusion algorithm of image deep learning by analyzing the frequency domain and spectral characteristics of beats, which improved the correct rate of beat recognition by 6% [2]. Li Xin et al. proposed a new independent recurrent neural network model after analyzing the decision tree, logistic regression, neural network and other models to address the problem of a single model for music beat recognition, which can be classified and trained for different beat features through the adaptive selection mode of multiple models, and obtain targeted and more accurate recognition results [3]; Zeng Shengqiang et al. proposed an algorithm to address the problem of misjudgment caused by the lack of information mining in beat recognition methods, and proposed an image deep learning algorithm to improve the correct rate of beat recognition by 6% [3]; Zeng Shengqiang et al. proposed an algorithm to address the problem of misjudgment caused by the lack of information mining in beat recognition methods. Zeng Shengqiang et al. proposed a deep learning-based gesture recognition technology for the problem of misjudgment caused by insufficient information mining in beat recognition methods, which improves the completeness of information capture and transmission when 2D information is transformed into 3D information, and significantly improves the accuracy of beat recognition and performance action matching [4].

In summary, the study tries to use a combination of short-time Fourier transform combined with Meier spectral computation to preprocess the beat data and clarify the feature extraction scheme; then the independent recurrent neural network model is

improved to enrich the model adaptive selection scheme; finally, the drawbacks of the incomplete information conversion are improved by using the gesture recognition technology to realize the construction of the performance robot based on beat recognition.

2. Performance robot framework design based on beat recognition

The framework design of the performing robot is shown in Figure 1. Among them, the beat recognition module and the dance performance part are constructed based on the results of music style recognition. The main reference feature for music beat data preprocessing is the result of style recognition. The gesture estimation is to extract the movements in the dance video, summarize and construct the dance movement library, and finally, according to the music beats recognized by the robot, select the appropriate movements from the movement library at the corresponding beat time to complete the dance performance [5].

3. Music beat recognition

3.1 Idea of music beat recognition

Beat recognition is the basic core of the robot music performance, and the basic beats of music are divided into two beats, three beats, and four beats. To accurately distinguish music styles and recognize beats, the robot recognition system needs to evenly distribute data in different recognition dimensions such as music rhythm, harmony, timbre, etc. Therefore, the system needs to add two modules of style recognition and model selection before beat recognition, so that the system can efficiently and accurately recognize and learn music beats [6]. The basic structure of beat recognition is shown in Figure 2.

3.2 Music Style Recognition

The music style recognition process first used the scattering transform method to extract features from the audio data, and then to carry out style recognition and classification using the IndRNN network model.IndRNN network is a kind of independent recurrent neural network model, which is mainly trained on the audio data using the acceleration model of the BN layer in the model, and adjusts the distribution of the audio tempo data, to realize the recognition of the music style. The specific structure of the IndRNN network model is shown in Figure 3.

In Figure 3, Weight represents the input processing of audio, BN represents the model learning and training layer, Recurrent+RELU represents the cyclic processing, and RE-LU represents the activation function of the model. The scattering transform formula of the style recognition module is shown in equation (1).

where λ denotes the audio wavelength, $x * \psi\lambda$ denotes the audio signal, $\varphi(t)$ denotes the high-frequency signal to be rejected, and m denotes the scattering order.

3.3 Beat Recognition

3.3.1 Beat Signal Preprocessing and Feature Extraction

The data preprocessing for beat recognition is to convert the audio signal data into frequency domain by three short-time Fourier transform algorithms with different window lengths, and combine the results with its first-order median positive difference

stack as the extracted features.

The data preprocessing first utilizes three windows to segment the captured audio x(n) into several subframes, with segmentation window lengths of 21.5ms, 42.6ms, and 91.6ms, respectively, corresponding to the number of samples of 1024, 2048, and 4096. To avoid the signal of the stacked part of the overlapped frames being weakened due to the addition of the window, the time shift of the overlapped frame samples is set to 10ms, and the overlap length corresponds to the window segmentation length one by one, which is 23.2ms, 46.4ms, and 92.8ms, respectively [7]. The specific steps and formulas for data preprocessing are as follows.

The formula for calculating the short-time Fourier transform of complex audio is shown in equation (2).

where X(n) denotes the input signal of the nth frame, W(l) denotes the Hamming window function, k represents the frequency index, and h denotes the length of the sample time shift between two neighboring segmentation frames. After the signal is converted from audio to spectrum, to omit the phase in the spectrum, the phase information needs to be removed by the power spectrum calculation, which is simplified to the signal power spectrogram S(n, k), and the transformation formula is shown in equation (3).

The transformed audio spectrogram has a large dimension, and the Mel scale filter bank is generally used to convert and adjust the audio Hertz size to obtain the appropriate acoustic frequency signal characteristics. After conversion by the Mel scale filter bank, the Mel scale to the sound Hertz perception of the amount of the human ear to achieve consistency, Mel frequency scale and ordinary scale conversion relationship expression shown in equation (4), the conversion spectral curve shown in Figure 3 [8].

According to the curve information in Fig. 4, it can be seen that the conversion curve shows a log relationship, and the conversion speed becomes slower with the increase of spectrum samples. To address this problem, when the signal is converted, 20 triangular filter banks are set equidistant on the Mel scale, and 1 is added before the calculation of the logarithm is taken to ensure that the amplitude value of the conversion result is positive. The improved conversion formula is shown in equation (5).

Where M(n, m) represents the Mel spectrogram, S(n, k) represents the signal power spectrogram, and F(m, k)T represents the Mel cepstrum, the Mel spectrogram and the power spectrogram cepstrum conversion of the signal at that point at the moment T. Sound spectrum converted by short-time Fourier transform, sub-frame and Mel scale filter conversion that is the completion of the preprocessing, will be converted to obtain the spectrogram as a music beat extraction features to time-frequency, for example, to obtain audio time-frequency feature image shown in Figure 5.

3.3.2 Beat Recognition Model Construction

Studies have shown that independent music recognition classification recurrent neural networks have the problem of insufficient information support when targeting multi-performance task scenarios, and it is often necessary to use subsequent information to assist in judging music beats. To address this problem, a recognition model based on a bidirectional recurrent neural network (bidirectional LSTM) is proposed, which attempts to model the before and after times of the input data to detect music beat timing signals [9]. The schematic diagram of the recognition network structure is shown in Fig. 6.

Considering the network computing power and the reaction time of the dancing robot, a bidirectional LSTM model containing five hidden layers is finally adopted. In the model training, it is found that the original audio signal is complicated by the features, which leads to the decline of the network generalization ability of the recognition model. It is difficult to recognize diverse music styles by a single recurrent neural network model. To address this problem, music style features are added to the model training to construct an adaptive multi-model selection beat recognition system, and the system structure is shown in Figure 7. The construction idea of the model selection module is that, before the beat data preprocessing, the music style data subset is divided according to the music style classification standard, and all of them are thrown into the network to carry out adaptive learning to establish the recognition model with different features [10].

4. Action library construction for gesture estimation

According to the structure of Fig. 7, it can be seen that the performing robot first estimates the dance gestures based on the recognized music beats and styles, and then matches the dance movements from the dance movement library. The gesture estimation, as the core foundation of dance performance, is mainly for the estimation of the key points of action performance, such as the position of the head, hands, feet, and other parts. Intelligent robot gesture estimation is mainly categorized into single gesture and multi-person gesture according to the gesture mode. Considering the independence of the robot, the single-person posture estimation algorithm is finally chosen to construct a dance action library, the specific steps are: firstly, detect the 2D key point data of human action in the frame rate; secondly, convert the 2D key point information into 3D; finally, match different actions according to the 3D information [11]. The specific construction process is shown in Figure 8.

According to the structure of Fig. 8, it can be seen that the three main algorithmic techniques of the posture estimation module are YOLOv3 portrait detection technology, HRnet2D key point information acquisition algorithm, and TCN high-precision 3D posture reconstruction technology, and the specific calculation formula and process of the algorithms and techniques are as follows.

4.1 YOLOv3 Portrait Detection

YOLOv3 is an end-to-end target detection algorithm, the third generation of the YOLOv algorithm has stronger stability and flexibility and can balance the relationship between speed and accuracy by adjusting the size of the model structure. Using different scales of extracted features for target detection, the resolution of the feature map is reduced, and its medium-small feature map connecting its dimension-matching feature map is taken as the detection target in the subsequent stage [11]. The specific formula of the YOLOv3 function is shown in equation (6).

where λcoord denotes the weight of prediction frames containing objects, λnoobj denotes the weight of prediction frames without objects, Ibj denotes prediction frames containing objects, and Ioobj denotes prediction frames without objects, wi, hi, xi, and yi denote the sizes and coordinate positions of labeled frames, and ‖ i, i, i, and i denote the size and position; ci denotes whether the object exists or not, i denotes whether the network predicts the target exists or not, 1 denotes existence and 0 denotes non-existence, pi(c) denotes the probability that the detected region is of category c, and i(c) denotes the probability that the network-predicted region belongs to category c.

4.2 HRnet 2D pose acquisition

2D pose estimation is the basis of 3D pose reconstruction, and the conventional 2D pose estimation method is the process of restoring the image resolution to a certain high resolution in order from low to high. Considering the multiple neural network structure of the performing robot, the conventional 2D pose estimation method cannot meet the robot's pose acquisition requirements, therefore, this study adopts the new HRnet architecture for robot 2D pose acquisition, the specific structure is shown in Fig. 9. The advantage of the HR-net2D pose acquisition technique is that the high resolution of the image can be maintained throughout the entire pose acquisition process. The high-resolution subnet is used as the first stage of information acquisition during the posture acquisition, and then the subnets are added one by one in the order of resolution from high to low, and then the subnets are connected sequentially to acquire multiple stages. In the process of adding subnets, since the resolution is different, the information needs to be exchanged and fused repeatedly between the subnets to obtain richer information, which can improve the accuracy of the attitude key point information acquisition [12].

4.3 TCN3D Attitude Reconstruction

Robot performance motion requires high-precision 3D posture to provide joint information, and the full convolution model in the null time domain is mainly used when 2D posture key point information is converted to 3D. The emerging TCN sequence data prediction network in the model can accurately predict the 2D joints' pose in 3D, and the network structure is shown in Fig. 10. The innovation of the TCN3D sequence data prediction model lies in the following two points, firstly, using the parameter design of the one-dimensional fully convolutional network, which can keep the network input and output data in the same dimension; and secondly, the parameter design of the network makes the convolutional network layers have a causal relationship with one another, which ensures that the network has no cause and effect relationship before. causal relationship, to ensure that the network before the transmission of information and future prediction data has a more complete inheritance relationship between the network, to reduce the information data in the transmission of the phenomenon of missing and miscommunication. Based on these two characteristics, TCN in the field of music modeling information transfer and pose reconstruction results are more accurate, 2D conversion 3D pose schematic shown in Figure 11 [13].

5. Experimental Validation

5.1 Validation of music beat recognition algorithm

5.1.1 Evaluation Metrics

Music beat recognition metrics are selected as follows.

(1) F-measure. F-measure is a composite index of rhythm check rate (positive) and beat check rate (Negative), and the F-value increases with the increase of check rate and check rate.

(2) P-score and Gemgil: P-score represents the beat tracking accuracy, if the measured value is less than 1/5 of the annotated beat interval, the beat tracking is accurate; Gemgil is the summation of beat scores by Gaussian function, and evaluates the beat tracking accuracy based on the summation value. Compare the normalized P-score value and Gemgil value, and take the larger one as the evaluation value.

(3) CMLc&CMLt. CMLc denotes the longest recognized segment and CMLt denotes the correctly tracked beats, and the system is considered to be correctly recognized if the measured values of both are less than 17.5% of the beat and phase tolerances in the case of the correct metric.

(4) D&Dg. D denotes information gain and Dg denotes global information gain, comparing the results of both annotation and detection. Statistics and calculation of the error dispersion of the two detections, the lower the dispersion value, the more accurate the recognition effect [14].

5.1.2 Experimental results

Based on the beat recognition process, the accuracy of the system for recognizing music styles is first tested. Two kinds of music with different styles are randomly selected on the network and recognized separately, and the obtained beat recognition results are shown in Figure 12.

In Figure 12, the beat of Figure (a) is more soothing, the actual audio is hummed by the singer, there is no fixed beat, and the overall music is gentle, the system recognizes it as Blues style music beat, and it can be determined that the recognition results are more accurate. Figure (b) has a faster beat, the system recognizes it as Rock style music beat, and the actual music is a rock song. The experimental results show that the system has high accuracy in recognizing the music style.

Based on the above experimental results, the experiment continues to test the system's beat recognition effect. The experiment for the evaluation index of 4.1.1, combined with the dataset of Big Data Ballroom [15], selected the first 1000 frames of the training set of the model and carried out beat recognition detection, and the detection results are shown in Figure 13.

In the figure, the red part is the actual beat moment of the music dataset, and the blue part represents the beat recognition model training results. According to the results in Fig. 12, it can be seen that the best moments in the dataset can be recognized on the training set and output a larger probability value from the network model. The image shows that although the system may have occasional misjudgment of beat recognition at some moments, the overall recognition effect is more accurate and can be applied to the

beat recognition stage of performing robots. To further verify the accuracy of the recognition system on music style and beat recognition, the following randomly selected a piece of music from Ballroom big data, for the model evaluation index, were tested, the test results are shown in Table 1.

In Table 1, Kerbs is a music style, Humantapper denotes the vocal harmony recognition model, and BeatTracker denotes the beat tracking model. According to the experimental results, it can be seen that the F-value of beat recognition is 0.90, P-score and Gemgil take the larger value of 0.872, and the three indicators are at a high level; the indicator CMLt is 0.854, and the difference in compatibility is 15%, which is lower than the standard value of 17.4%, and it can be determined that the recognition is correct; the Dg-value is 2.392, which is a low value, with a low degree of dispersion, and can be considered to be accurate in the recognition of beats.

5.2 Performance robot action test based on beat recognition

5.2.1 NAO robot

Based on the algorithm described above, the NAO robot is selected as the carrier for testing, which is the most widely used robot in various academic fields, with a height of 58cm and a Linux operating system computer at its core. The action commands are mainly for the robot head, arms, hands, legs, feet, and other parts of the robot, and the commands are executed by the drive motors during the performance to complete the coordinate joint rotation.

5.2.2 Analysis of the results of the realization of the performing robot based on beat recognition

A piece of music is randomly selected for beat recognition, and the results are shown in Table 2, where the serial number indicates the beat point of the music, and the time indicates the time point at which the beat point appears in the music.

Based on the performing robot in 5.2.1, the experimental results are shown in Fig. 14, combining the previous research content and experimental collection of pictures. The experimental process of action acquisition is more, like taking a certain beat action as an example to recognize the conversion, the specific results are shown in Figure 14.

The 2D key point human skeleton map and the 2D transformation 3D key point skeleton map are shown in Figure 14. According to the transformation results, the final dance movement of the robot is shown in Figure 15.

The above experimental results can be verified, based on the bidirectional LSTM network model of the beat recognition system applied to the performance of the robot has better performance, the accuracy of the system to identify beats can be maintained at about 90%; on this basis, the robot uses the gesture recognition technology to recognize the results of the robot through the robot action performance to achieve the performance of the performance robot based on the beat recognition of the action performance, the experiments have fully proved that the robot based on the beat recognition is a good performance. The experiment fully proves that the beat recognition system based on the bidirectional recurrent neural network can meet the design requirements of the performing robot.

6. Conclusion

Through the study of beat recognition and robot action training and other related contents, a music-performing robot based on beat recognition is constructed, and the feasibility of robot action performance is verified through a series of experiments. The following conclusions were obtained in the research and experiments.

(1) Through algorithmic analysis, the Melt sign and power spectrum features are identified for beat data preprocessing.

(2) An adaptive selection based on a bidirectional recurrent neural network is proposed to classify and train music styles and beats by combining the results of music style recognition to improve the accuracy of music beat recognition.

(3) Build a dance movement library based on gesture recognition technology. Enrich the action library samples through gesture recognition technology to provide more dance action matching options for the robot.

(4) Verify the feasibility of the beat recognition system applied to the performing robot through a series of experiments, and successfully realize the music-performing robot based on beat recognition through the design system.

References

1. Li Wang, Xin Wang, Lingyun Xie. A review on the characterization of music signal processing. Journal of Communication University of China (Natural Science Edition), 2021, 28(6):59-72.

2. MENG Zhen, WANG Hao, YU Wei, et al. Research on vocal music classification based on feature fusion. Data Analysis and Knowledge Discovery, 2021, 5(5):59-70.

3. Xin Li, Hongjuan Mi, Xuejun Wu. Comparison of multiple machine learning models for classification of music genres. Journal of Yibin College, 2020, 20(12):42-47.

4. Shengqiang Zeng, Lin Li. A 2D/3D skeleton action recognition method based on pose correction and pose fusion. Computer Application Research, 2022, 39(3):900-905.

5. Yi Ru. Deep learning-based real-time pose recognition algorithm for generating character 2D animation. Journal of Taiyuan College (Natural Science Edition), 2022, 40(1):69-74.

6. Ren R. A music beat recognition system based on audio technology. Microcomputer Applications, 2022, 38(3):58-61+69.

7. YANG Wenwen, SHI Mengluo. Deep learning-based music feature extraction and genre classification. Yangtze River Information and Communication, 2021, 34(5):9-11.

8. Fan Yongguan. Electronic music labeling algorithm based on Fourier transform and cepstrum coefficient. Modern Electronic Technology, 2020, 43(13):155-158.

9. Zheng Qingjie, Long Hua, Shao Yubin, et al. A speech music classification model based on beat spectrum. Communication Technology, 2020, 53(11):2675-2679.

10. Tian Jialu, Zhang Yan. A method for automatic music recognition and real-time visualization. Computer and Information Technology, 2020, 28(4):9-12.

11. Wan Y. Design research on multimodal yoga movement posture detection. Sports Research and Education, 2021, 36(4):90-96.

12. Ren Guoyin , Lv Xiaoqi , Li Yuhao . Real-time action recognition based on multi-feature fusion of 2D to 3D skeleton. Advances in Lasers and Optoelectronics, 2021, 58(24):241-249.

13. YANG Jinlong, SHI Minghui, CHAO Fei, et al. A dance robot based on deep learning for motion imitation. Journal of Xiamen University (Natural Science Edition), 2019, 58(5):759-766.

14. H.L. Zhang. Research on automatic generation of robot dance movements based on artificial intelligence. Automation Technology and Application, 2022, 41(4):82-85+165.

15. Chi Shanjiao. Research on the Promotion and Innovation of Line Dance Intelligent Robot. Journal of Tonghua Normal College, 2020, 41(6):91-95.

Author Biography

Zhang Ting (1989-) female, Shangluo, Shaanxi, master's degree, lecturer. 726000