Scholar Publishing Group
International Journal of Art Innovation and Development
https://doi.org/10.38007/IJAID.2023.040108
ISSN 2789-7192 Vol. 4, Issue 1: 102-109

The Design of a Music Scene Recognition Method and System Based on Artificial Intelligence

Yang Nan[1, 2*]

[1] *Jinzhong University, Jinzhong, China*

[2] *Philippine Christian University, Manila, Philippine*

sxtyjzxy@163.com

[*] *corresponding author*

Keywords: Artificial Intelligence, Music Scene, Scene Recognition, System Design

Abstract: The main melodic information of music can be used for content-based audio retrieval, management as features of audio files and to meet the business needs of specific departments and scenarios. The purpose of this paper is to study music scene recognition methods and system design based on artificial intelligence. The overall system architecture design is integrated by analysing the current stage of music scene recognition, proposing a modelling approach for music scene recognition system for music scene recognition needs and a CNN-based music scene recognition approach. A music recognition module was implemented based on the music scene recognition framework, and the generic scene text recognition system was tested. The experimental results show that the average time consumed for model computational efficiency is 60ms and the cpu resource consumption is maintained at about 28%.

1. Introduction

In recent years, with the rapid development of computer technology and the Internet, the information and data in the network is exploding, among which audio information occupies a large part of the data volume in the Internet and still shows a rising trend year by year, the most proportion of audio information is also the most important information is music information [1-2]. The rapid growth of music information provides convenience to people's life, and how to quickly and accurately retrieve the required information from the huge amount of song resources also has an important practical value [3-4].

Scene recognition is one of the fundamental problems in computer vision research and has a wide range of applications in robotics [5].Veronica Naosekpam introduces a new adaptive scene recognition method that exploits self-supervised translation between modalities. In fact, learning from RGB to depth and vice versa is an unsupervised process that can be trained jointly on data from multiple cameras and helps to bridge the gap between the extracted feature distributions. Experimental results confirm the effectiveness of the proposed approach [6]. Khosro Rezaee proposed a new end-to-end scene recognition framework, called the Recursive Memory Attention Network (RMAN) model, which performs object-based scene classification by recursively locating and memorising objects in images. Based on the proposed framework, they introduce a multi-tasking mechanism that continuously focuses on different underlying objects in the scene images and recursively memorises and fuses the object features focused by the attention model to improve scene recognition accuracy. Experimental results show that the RMAN model achieves better classification performance on the constructed dataset and two common scene datasets, outperforming state-of-the-art image scene recognition methods [7]. Leena Mary Francis proposed a cross-modal music emotion recognition method that correlates music samples with emotions in a common space by considering their general and specific features. Since human subjective perception leads to uncertainty in the association of music samples with emotions, we compute compound loss-based embeddings to maximise two statistical features, a correlation between music samples and emotions based on canonical correlation analysis, and a probabilistic similarity between music samples and emotions with KL divergence. Experiments on two benchmark datasets demonstrate the superiority of our approach over a one-way baseline [8].

This paper describes the main components of an audio scene recognition system and presents the corresponding key techniques. An overview of convolutional neural networks used in audio scene recognition tasks is given, and two CNN classifier network structures are designed, inspired by the VGGNet network structure and incorporating differences in the number of input channels present in audio signals.

2. Research on Artificial Intelligence Based Music Scene Recognition Method and System Design

2.1. Requirements Analysis

The music scene recognition system in this study is designed as a scalable and deployable platform, not only for general users but also for developers using the system services, with the aim of providing a flexible platform capability to meet the usage needs of different users. The needs of the two roles of use are expressed in three ways:

(1) Generic needs, hope that the system provides the recognition service, for common scenarios are guaranteed recognition effect;

(2) For the general user, it is hoped that the system can be used directly through an interactive and friendly interface operation, and that the system can provide common scene recognition functions and their layout analysis

(3) For developers who use the system package service, they hope that the service package is high and convenient for function expansion.

2.2. CNN-based Music Scene Recognition Method

2.2.1. Feature Extraction

The distribution of the central frequency of each sub-band of CQT shows an exponential pattern. Moreover, the bandwidth of each sub-band varies, but the ratio of the centre frequency of each sub-band to its bandwidth is a constant, which is called Q. This is also the origin of the name of CQT.

In order to obtain data from the CQT spectrogram that can be processed by DNN, the CQT spectrogram is further processed using two steps, segmentation and pooling. The main purpose of this is to reduce the dimensionality of the data while at the same time obtaining features that can be valid for future training.

Firstly, suppose there are L segments of scene audio in the training set, and the previously extracted CQT spectrogram has a total of T frames and the number of bands is P. Then the CQT spectrogram of the Ith audio in the L samples is divided into M segments without overlap, and the length of each segment is Q=T/M. Using Sml to denote the mth segment of the Ith audio sample, which contains a total of Q frames, the whole spectrogram of this segment can be represented as $S^l = \left| S_0^l, S_1^l, ..., S_{M-1}^l \right|$, and then for each Sml is averaged on the time axis and Sml is transformed into a column vector and represented by vml, then the feature parameters of the lth short segment of the scene audio can be represented as $V^l = \left| v_0^l, v_1^l, ..., v_{M-1}^l \right|$, and finally the feature parameters of the whole training set are $V = \left| v^1, v^l, ..., v^L \right|$, and $V \in \mathrm{P}\Re_+^{R\times N}$, where N=ML.

2.2.2. Convolutional Neural Network Structure

There are two key steps in updating the parameters of a neural network using the BP algorithm, the first step is the forward propagation of the network and the second step is the back propagation of the network. Forward propagation is the use of the network parameters to map the network input to an output. The purpose of back propagation is to optimise the network parameters based on the error between the input obtained from forward propagation and the actual value, and after a period of training, when the error reaches the training criteria, which generally requires an error of less than 0.01, stop training.

The derivation process of the computational formula for updating the parameters of the CNN's back propagation is vastly complex, and given that the focus of this paper is not on this, the formula for updating the parameters is briefly described below. In the case that the current layer is a convolutional layer and the sensitivity of that layer is known, the formula for calculating the sensitivity of the previous pooling layer is:

$$\delta_j^l = \beta_j^{l+1}(f'(u_j^l) \otimes up(\delta_j^{l+1})) \tag{1}$$

Where up (.) is an up-sampling operation. The gradient of b can also be obtained by summing the nodes of sensitivity of the current layer l by equation (2):

$$\frac{\partial E}{\partial b_j} = \sum_{u,v}(\delta_j^l)_{uv} \tag{2}$$

Then the gradient of the weights of the convolution kernel can be found:

$$\frac{\partial E}{\partial b_j}=\sum_{u,v}(\delta_j^l)_{uv}\;(p_i^{l-1})_{uv} \tag{3}$$

Where $(p_i^{l-1})_{uv}$ is the local of the feature map that acts with the convolution kernel when performing the convolution calculation.

2.2.3. Overfitting

The BN mechanism is essentially the process of data normalisation. The data is normalised to make the data distribution consistent. The goal of a neural network is to learn this distribution of the input data. However, if the data distribution keeps changing while the network is being trained, it will slow down the learning speed of the network and, more seriously, affect the learning ability of the network. During network training, each layer in the network learns the distribution of the input data for that layer, but since the network parameters are constantly updated, the data distribution will change after the data passes through one of the layers after a parameter update is completed, and this change will be continuously amplified as it passes through multiple layers of the network, which results in a small change in the data distribution in the front, but a large change in the data distribution in the back of the network This results in significant changes in the distribution of data in the later layers of the network, which affects the network's ability to generalise.

3. Investigation and Research on Artificial Intelligence Based Music Scene Recognition Method and System Design

3.1. Modelling of Music Scene Recognition System

This paper designs a convolutional neural network based audio scene recognition system. After data processing, four different types of input signals will be generated, including two-channel signals including left and right channels, central/side channels and HPSS, and the single-channel signal obtained by the background difference method. For the different number of input signal channels, two classifier networks are proposed in this paper to solve this problem, namely single channel classifier network (1-conv) and two channel classifier network (2-conv). The prediction results of the system are finally obtained by integrating the prediction results of several groups of individual classifier networks for learning. The system somewhat overcomes the limitation caused by the insufficient amount of data and improves the overall learning ability and generalization capability.

3.2. Overall Architecture Design

The system adopts a layered architecture design scheme, with a clear hierarchy and interfaces for communication between layers, which enables developers to only care about the business logic of each layer, coupling functional implementation and basic services, ensuring the specification of atomic function-based development, a more logical system architecture and easier reuse of functions. The platform is divided into three layers, from top to bottom, the system representation layer, the basic service layer and the model service layer, as shown in Figure 1.

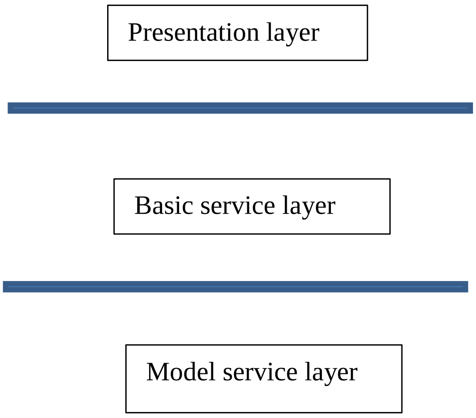

Figure 1. Platform structure

3.3. Experimental Platform

In this paper, in order to verify the classification performance of the proposed music scene classification system, the audio scene recognition system was built on the dcase_util toolbox, using TensorFlow (v1.4.0) as the backend. In particular, the dcase_util toolbox is designed for sound scene and event detection classification (DCASE). The utilities in the toolbox are mainly used to process metadata and various other forms of structured data, and to provide a standardised API for audio datasets from different sources.

4. Analysis and Research of Artificial Intelligence Based Music Scene Recognition Method and System Design

4.1. Music Recognition Module

Each piece of music has a unique musical emotional character in different sections, some are upbeat, some are sad and some are rousing. The music recognition module identifies the different types of music and the emotional tone of each piece of music in different sections. Music beats and genres are extracted through a series of variations and analysis of the music files. The music recognition module mainly includes the implementation of functions such as music file analysis, transformation, music beat extraction, music emotion feature analysis and music and beat association. The flowchart is shown in Figure 2.

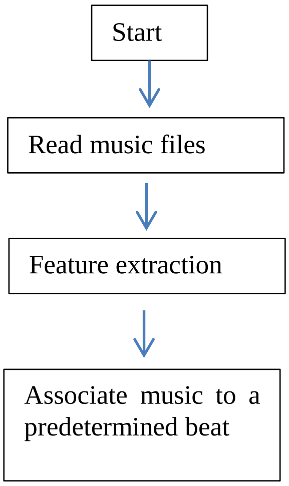

Figure 2. Work flow diagram of music recognition module

4.2. System Non-functional Testing

4.2.1. Recognition Time Test

The data used for the test was one episode from each of the three TV series selected as a music test sample. By automating the data filling recognition, the music recognition results and model inference time were outputted by way of system logs and then human statistics were performed. The final test results of each category are shown in Table 1 and Figure 3.

Table 1. Time test table

Test items	Test result 1	Test result 2	Test result 3
Identification time	56ms	64ms	60ms

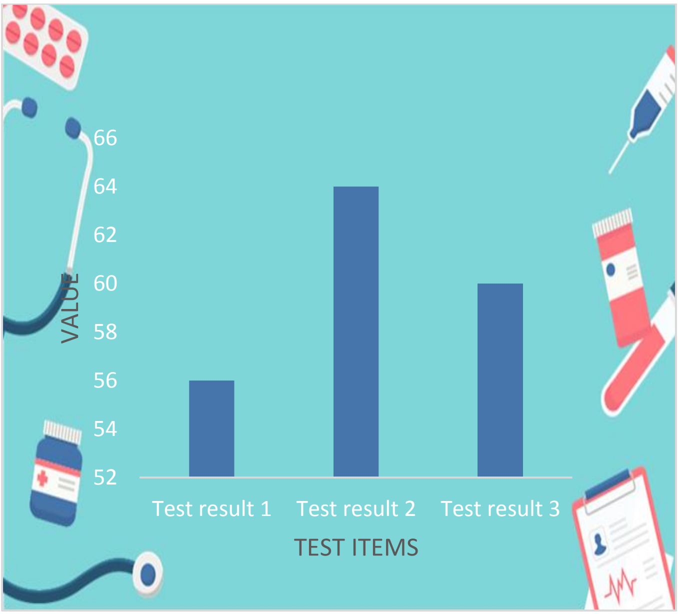

Figure 3. Test result

4.2.2. Memory Resource Usage Test

The user experience is very important during the running of software. If several large pieces of software are running in the background, then running the application again will cause the user to wait and have a poor user experience. To ensure the smoothness of the system, we reduce the CPU usage and memory usage when the application is running, as shown in Figure 4.

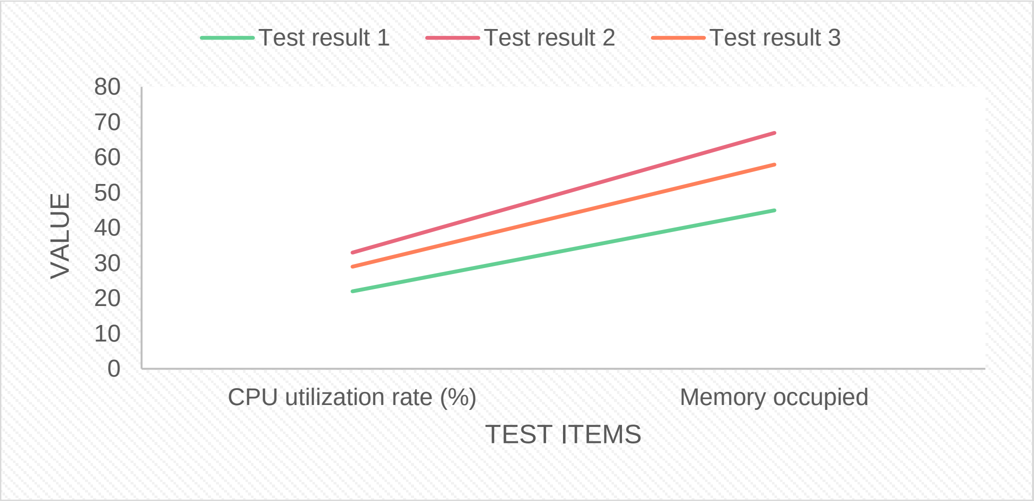

Figure 4. Resource usage test

5. Conclusion

Usage scenes is an emerging music metadata recently proposed in the field of music information retrieval, while the current research works related to music usage scenes are relatively few and all in the early exploration stage. In this paper, we build a CNN-based audio scene recognition system. CNN is chosen because it is the most mainstream network structure in deep learning, and has good performance in both image recognition and speech recognition. The use of the BN mechanism in the network can effectively prevent the occurrence of overfitting problems. Although a lot of work has been done in this paper during the research of audio scene recognition methods, deep learning and audio processing techniques have entered into in-depth research, and the recognition ability of the proposed audio scene recognition system has been improved to a certain extent, there are many factors affecting the effect of audio scene recognition, and the content of deep learning is very complex, so there is still a lot of research space.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Gwenaelle Cunha Sergio, Minho Lee. Scene2Wav: a deep convolutional sequence-to-conditional SampleRNN for emotional scene musicalization. Multim. Tools Appl. (2021) 80(2): 1793-1812.

[2] Bela Usabaev, Anna Eschenbacher, Angela Brennecke. The Virtual Theremin: Designing an Interactive Digital Music Instrument for Film Scene Scoring. I-com. (2022) 21(1): 109-121.

[3] Neeraj Gupta, Anand Singh Jalal. Traditional to transfer learning progression on scene text detection and recognition: a survey. Artif. Intell. Rev. (2022) 55(4): 3457-3502.

[4] Recep Sinan Tümen, T. Metin Sezgin. Segmentation and Recognition of Offline Sketch Scenes Using Dynamic Programming. IEEE Computer Graphics and Applications. (2022) 42(1): 56-72.

[5] T. Mithila, R. Arunprakash, A. Ramachandran. CNN and Fuzzy Rules Based Text Detection and Recognition from Natural Scenes. Comput. Syst. Sci. Eng. (2022) 42(3): 1165-1179.

[6] Veronica Naosekpam, Nilkanta Sahu. Text detection, recognition, and script identification in natural scene images: a Review. Int. J. Multim. Inf. Retr. (2022) 11(3): 291-314.

[7] Khosro Rezaee, Mohammad Reza Khosravi, Maryam Saberi Anari. Deep-Transfer-Learning-Based Abnormal Behavior Recognition Using Internet of Drones for Crowded Scenes. IEEE Internet Things Mag. (2022) 5(2): 41-44.

[8] Leena Mary Francis, N. Sreenath. Robust scene text recognition: Using manifold regularized Twin-Support Vector Machine. J. King Saud Univ. Comput. Inf. Sci. (2022) 34(3): 589-604.

[9] Riadh Harizi, Rim Walha, Fadoua Drira, Mourad Zaied. Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. Multim. Tools Appl. (2022) 81(3): 3091-3106.

[10] Riadh Harizi, Rim Walha, Fadoua Drira, Mourad Zaied. Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. Multim. Tools Appl. (2022) 81(3): 3091-3106.

[11] Fatemeh Naiemi, Vahid Ghods, Hassan Khalesi. Scene text detection and recognition: a survey. Multim. Tools Appl. (2022) 81(14): 20255-20290.

[12] Andreea Glavan, Estefanía Talavera. InstaIndoor and multi-modal deep learning for indoor scene recognition. Neural Comput. Appl. (2022) 34(9): 6861-6877.

[13] Nitika Nigam, Tanima Dutta, Hari Prabhat Gupta. FactorNet: Holistic Actor, Object, and Scene Factorization for Action Recognition in Videos. IEEE Trans. Circuits Syst. Video Technol. (2022) 32(3): 976-991.

[14] Cinu C. Kiliroor, S. Shrija, R. Ajay. Augmented Scene Text Recognition Using Crosswise Feature Extraction. Wirel. Pers. Commun. (2022) 123(1): 421-436.

[15] Payel Sengupta, Ayatullah Faruk Mollah. Journey of scene text components recognition: Progress and open issues. Multim. Tools Appl. (2021) 80(4): 6079-6104.

[16] L. T. Akin Sherly, T. Jaya. An efficient indoor scene character recognition using Bayesian interactive search algorithm-based adaboost-CNN classifier. Neural Comput. Appl. (2021) 33(22): 15345-15356.

[17] S. Chandrakala, S. L. Jayalakshmi. Generative Model Driven Representation Learning in a Hybrid Framework for Environmental Audio Scene and Sound Event Recognition. IEEE Trans. Multim. (2020) 22(1): 3-14.

[18] Anna Zhu, Seiichi Uchida. Scene word recognition from pieces to whole. Frontiers Comput. Sci. (2019) 13(2): 292-301.