Hindawi
Scientific Programming
Volume 2021, Article ID 2388697, 9 pages
https://doi.org/10.1155/2021/2388697

Research Article

Digital Audio Scene Recognition Method Based on Machine Learning Technology

Sihua Sun

Anhui Art College, Hefei, Anhui Province 230011, China

Correspondence should be addressed to Sihua Sun; 114039@ahua.edu.cn

Received 19 October 2021; Revised 2 November 2021; Accepted 10 November 2021; Published 26 November 2021

Academic Editor: Bai Yuan Ding

Audio scene recognition is a task that enables devices to understand their environment through digital audio analysis. It belongs to a branch of the field of computer auditory scene. At present, this technology has been widely used in intelligent wearable devices, robot sensing services, and other application scenarios. In order to explore the applicability of machine learning technology in the field of digital audio scene recognition, an audio scene recognition method based on optimized audio processing and convolutional neural network is proposed. Firstly, different from the traditional audio feature extraction method using mel-frequency cepstrum coefficient, the proposed method uses binaural representation and harmonic percussive source separation method to optimize the original audio and extract the corresponding features, so that the system can make use of the spatial features of the scene and then improve the recognition accuracy. Then, an audio scene recognition system with two-layer convolution module is designed and implemented. In terms of network structure, we try to learn from the VGGNet structure in the field of image recognition to increase the network depth and improve the system flexibility. Experimental data analysis shows that compared with traditional machine learning methods, the proposed method can greatly improve the recognition accuracy of each scene and achieve better generalization effect on different data.

1. Introduction

As an information carrier, sound is an important way for us to perceive the external environment. With the development of signal processing technology and computer science, the audio processing task of extracting information from sound assisted by machine has attracted more and more researchers' attention [1–6]. Compared with images with multimedia information, the acquisition of audio files is not limited by light environment and visual obstacles. In addition, audio occupies less capacity and faster processing speed compared with images [7–9]. Audio processing tasks include speech recognition, audio fingerprint, music mark, audio scene recognition, etc.

Audio scene recognition belongs to the subfield of computational auditory scene analysis. Its main goal is to enable devices to understand and distinguish their environment by analyzing sound. The implementation principle is that the equipment extracts different audio features through audio scene recognition technology to obtain the corresponding features and then models the audio scene according to these features, that is, constructs a classifier. After learning enough samples, the classifier will judge the audio scene category according to the extracted audio features [10–13]. The applications of audio scene recognition include context-aware services, intelligent wearable devices, robot sensing, and robot hearing. In addition, audio scene recognition also complements the research in several related fields. Among them, the detection and classification of audio events are often associated with audio scene recognition because the audio scene can be regarded as the product of the superposition of several audio events. On the other hand, audio scene recognition can improve the performance of sound event detection by providing a priori information about the probability of some events.

Previously, the implementation of audio scene recognition often applied general classifiers (such as Gaussian mixture model [14–16], support vector machine, and hidden

Markov model) to manually extracted features, such as mel-frequency cepstrum coefficient. In recent years, thanks to the improvement of computer speed and the rapid development of deep learning, people gradually realize that the characteristics of automatic feature extraction of deep learning can replace the inefficient manual extraction in the past. At the same time, more and more recording devices such as smart phones have made great contributions to the expansion of audio datasets [17]. With a large amount of audio data, it is possible to realize the deep learning method which is difficult to realize in the past.

2. Literature Review

Influenced by psychoacoustic/psychological technology, most of them emphasize the local and global features of audio scene recognition. In addition, a few researchers focus on the time-domain features of audio. Mitsukura [18] used mel-frequency cepstrum (MFCC) to describe the local spectral envelope of audio signal and Gaussian mixture model (GMM) to describe its statistical distribution. Zhao et al. [19] proposed to train the hidden Markov model by using the discriminant algorithm of the knowledge of training signal types to explain the time-domain evolution of GMM. Abrol and Sharma [20]. further improved the recognition performance by considering more features and adding a feature transformation step in the classification algorithm and obtained an average accuracy of 58% in 18 different sound fields [21].

Convolutional neural network (CNN) is a deep learning network model inspired by animal visual system. Its network composition imitates the principles of various cells in the visual system to construct the network model [22–24]. CNN was originally designed for feature extraction of two-dimensional data. It can directly establish the mapping relationship from low-level features to high-level semantic features and has achieved remarkable results in the field of two-dimensional image classification. Zhao et al. [25] proposed driver fatigue state recognition based on CNN. Cai et al. [26] proposed a CNN-based video classification method, which uses convolution filter and global average pool layer to obtain more detailed features. Tan et al. [27] discussed whether CNN in deep learning can be effectively applied to audio scene recognition.

Although applying a variety of deep learning methods to audio scene classification speeds up the research process in this field, there are still two problems worth discussing:

(1) The mainstream audio feature extraction methods are based on mel spectrum. Is there any special processing skill for mel spectrum to make it more in line with specific application scenarios, so as to improve the accuracy of audio classification?

(2) Since most of the results after audio feature extraction also belong to images, can the excellent volume of CNN architecture in the field of image recognition be applied to audio scene classification to improve the system performance?

Aiming at the above two problems, this paper proposes an audio scene recognition method based on optimized audio processing and CNN. The basic system is improved from two aspects: audio processing and network structure. In audio processing, binaural representation and harmonic impact source separation are used to process the original audio and extract the corresponding features, so that the system can make use of the spatial features of the scene, and then the classification accuracy has been significantly improved. In terms of network structure, we try to learn from the VGGNet structure in the field of image recognition, improve the system flexibility while increasing the network depth, and finally achieve better generalization effect on different data. Finally, the effectiveness of the proposed method is verified by experimental data analysis.

3. Audio Processing Optimization Method

3.1. Binaural Representation. Traditional audio processing methods always use MFCC feature in feature extraction. Although MFCC can describe features concisely with more than a dozen coefficients, MFCC is not competent for audio scene classification in scenes with obvious spatial features such as libraries.

Although it is common to use stereo equipment for recording, the signal is usually averaged to make it a mono before processing. Although using mono is easy for processing and feature extraction, it will lose a lot of spatial information. If important audio information is captured well in only one of the channels, problems may occur. Because averaging two channels to one channel will reduce the signal-to-noise ratio, the difference between the two channels is not easy to be reflected, so it is very easy to cause confusion in the classification process. Analyzing the two channels separately can alleviate this problem. In view of the excellent results obtained in previous challenges using binaural representation, this paper decides to use LR (left right) representation and MS (mid-side) representation in feature processing.

LR representation represents the left and right channels in conventional stereo recording. For example, when a car passes in front of a microphone, the sound moves from L channel to R channel or from R channel to L channel, which is only reflected in amplitude change in mono. By introducing LR representation, the movement of sound source in space can be reflected. In this section, only the left and right channels of the source audio file are separated. The MS representation emphasizes the time difference between sounds reaching each side of the stereo microphone. MS representation obtains the final result by summing and subtracting the waveforms of two stereo input channels, respectively. Mel spectrum will be extracted for the above four representations, as shown in Figure 1.

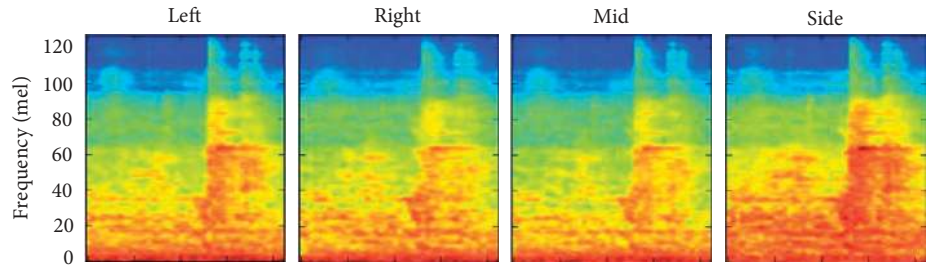

Figure 1: MFCC spectrum of binaural representation.

For LR spectrum and MS spectrum, they are input into the subsequent convolution neural network as one of the classification features.

3.2. Harmonic Percussive Source Separation Method. Sound can generally be divided into two types: harmonic and impact sound. In the traditional research work, harmonic impulse source separation (HPSS) algorithm is proposed under the background of music signal processing. The goal is to decompose the input audio signal into all harmonics and signals composed of all impulse sources. In order to solve the problem of poor generalization and dependence on learning data in audio scene classification, this section draws lessons from music signal processing to try to improve the classification performance of the system. The steps of HPSS algorithm are given below.

Assume that the input discrete input audio signal is $\lambda \in \mathbb{R}$. The harmonic component signal x_h and the shock source component signal x_p shall be calculated so that $x = x_h + x_p$.

First, the short-time Fourier transform of x can be expressed as

$$X(t,k) = \sum_{n=0}^{N-1} x(n+tH)w(n)\exp\left(\frac{-2\pi ikn}{N}\right), \quad (1)$$

where T is the number of frames, N is the frame size of Fourier transform, $w(n)$ is the window function, and H is the frame offset.

The input power spectrum Y can be calculated by the following formula:

$$Y(t,k) = |X(t,k)|^2. \quad (2)$$

Next, the harmonic enhancement spectrum $\widetilde{Y}_h$ and the shock source enhancement spectrum $\widetilde{Y}_p$ are calculated by median filtering Y. Assuming that A is a set composed of a column of real numbers and N is the number of real numbers in the set, the median filtering of A is defined as

$$\text{median}(A) = \begin{cases} a_{N-1/2}, & N \text{ is an odd number,} \\ \dfrac{a_{N/2} + a_{N-1/2}}{2}, & N \text{ is an even number.} \end{cases} \quad (3)$$

Then, according to the definition of median filter, the harmonic enhancement spectrum $\widetilde{Y}_h$ and impact source enhancement spectrum $\widetilde{Y}_p$ can be obtained by performing one horizontal and one vertical median filter on Y, respectively.

$$\begin{aligned} \widetilde{Y}_h(t,k) &= \text{median}\left(Y(t-l_n,k),\ldots,Y(t+l_n,k)\right), \\ \widetilde{Y}_p(t,k) &= \text{median}\left(Y\left(t,k-l_p\right),\ldots,Y(t,k+l_n)\right), \end{aligned} \quad (4)$$

where l_n and l_p are filter lengths.

Then, a variable β is introduced, which is called the separation factor. The original input signal $X(t,k)$ can be intuitively judged as harmonic or impact source component. Through this rule, binary masks M_h and M_p can be defined.

$$\begin{aligned} X_h(t,k) &= X(t,k)\cdot M_h(t,k), \\ X_p(t,k) &= X(t,k)\cdot M_p(t,k). \end{aligned} \quad (5)$$

Finally, the required signals x_h and x_p can be calculated by transforming these spectra into time domain by using inverse short-time Fourier transform.

In the separation process, with the help of the decompose.hpss method in the Librosa library, the separation factor for the experiment is 1.05. Convert stereo audio to mono before separation. As shown in Figure 2, when the HPSS algorithm is applied to the input signal, the harmonic tends to form a horizontal structure (in the time direction) and the impact source tends to form a vertical structure (in the frequency direction) on the mel spectrum.

4. Audio Scene Recognition Method Based on Convolutional Neural Network

Convolution layer is the core module of convolutional neural network, which can complete most of the heavy computing work. Its function is to extract the local features of a region. The core operation performed by the convolution layer is called convolution. Convolution is a common operation method in analytical mathematics. It is a mathematical operator that generates the third function through two functions. In the application of machine learning, convolution is usually embodied in sliding a filter on an image or some feature and using this operation to obtain a set of new features.

In this study, an audio scene classification system based on convolutional neural network will be designed and compared with traditional machine learning methods. The classification system mainly includes training and testing, and its process design is shown in Figure 3.

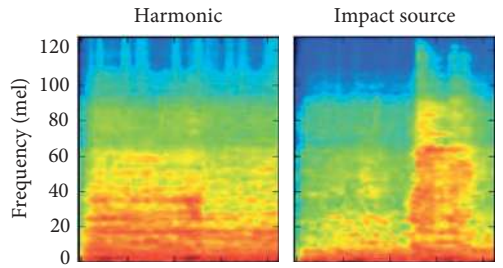

Figure 2: MFCC spectrum of harmonic percussive source separation method.

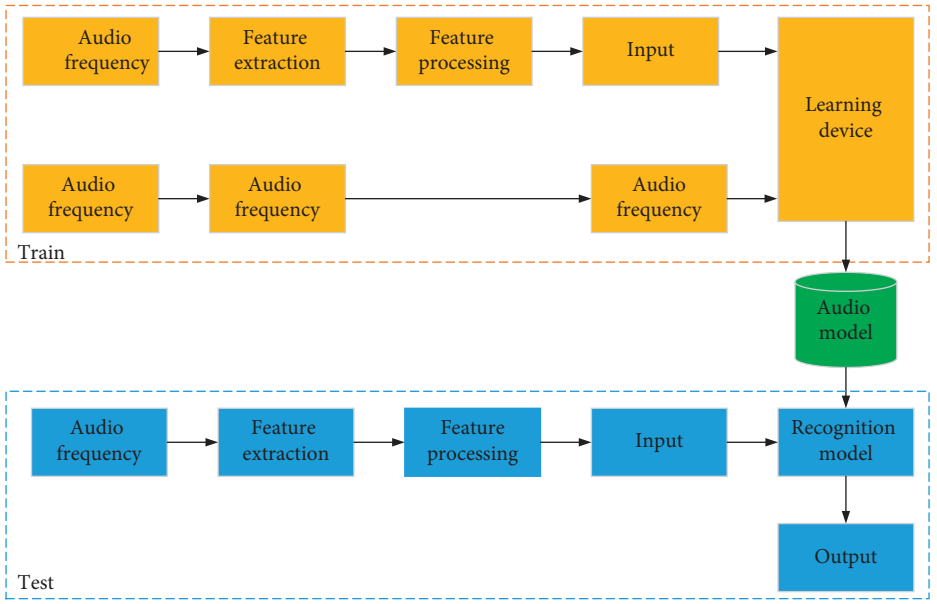

Figure 3: Audio scene recognition process based on convolutional neural network.

4.1. Convolutional Neural Network Architecture. Convolutional neural network structure also has the problem of relying too much on training parameters in the training process [28]. At present, there are many convolutional neural network frameworks with good performance that can be widely used. Whether its structure can be slightly modified and used in audio scene classification to improve system performance is also the focus of this section. The convolutional neural network module of the system core is shown in Figure 4.

The first layer performs convolution on the input spectrum. Since the size of convolution filter is reflected in the size of local block diagram and the size of local block diagram with different sizes determines the speed of feature extraction [29], filter sizes with different sizes will be arranged for comparison in the experiment. In addition, the number of filters will also affect the analysis angle of features, so different filter numbers will be arranged for comparison during the experiment. More filters will increase the angle of feature analysis but also increase the amount of calculation, and too many filters may lead to parameter redundancy. Generally, the number of filters is 2^n. After the convolution process is completed, the maximum pooling layer is used to subsample the obtained feature map.

The second convolution layer is basically the same as the first convolution layer, except that the second layer uses more cores (twice the first layer) to represent features at a higher level. Then, the second and last subsampling is performed for the "destruction" of the timeline. Therefore, the maximum pooling layer is still used, which operates over the entire sequence length. The activation function for the kernel in the convolution layer is ReLU.

Finally, because the classification involves 15 different classes, the last layer is the softmax layer composed of 15 fully connected neurons, which normalizes the output results of the network and makes the system output the classification results. Assuming that y_i is the output of the upper layer neuron i, y_i can be defined as

$$y_i = \text{softmax}\left(x_j\right) = \frac{\exp\left(x_j\right)}{\sum_{j=1}^{N} \exp\left(x_j\right)}, \tag{6}$$

where N is the total number of categories, x_i is the nonlinear input, and y_i is the prediction score of the input sequence belonging to class i.

4.2. Batch Standardization and Dropout Mechanism. In order to solve the problem of decreasing the network

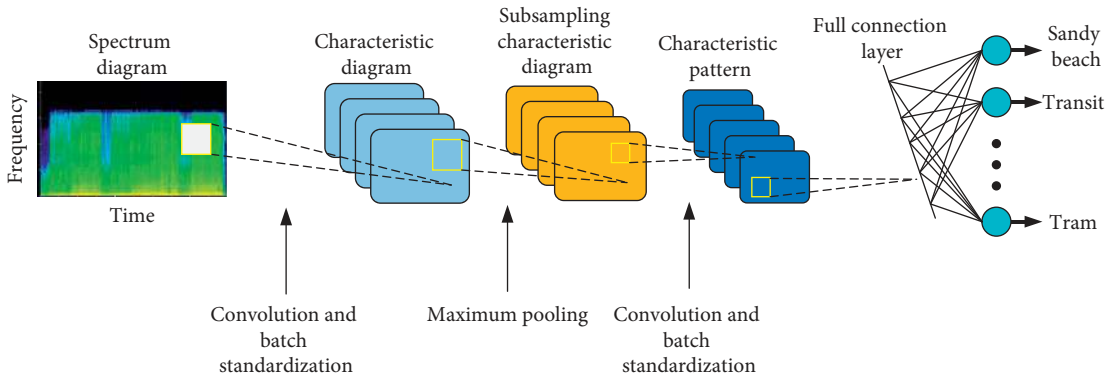

Figure 4: Composition of convolutional neural network module.

learning speed and convergence speed caused by internal covariate transformation, batch standardization is introduced to change the distribution of input data. The batch standardization layer applies linear transformation $BN_{\beta,\gamma}$ to its input x, as shown in the following formula:

$$BN_{\beta,\gamma} = \frac{\gamma}{\sqrt{\mathrm{Var}(x) + \varepsilon}} \cdot x + \left(\beta - \frac{\gamma \cdot E[x]}{\sqrt{\mathrm{Var}(x) + \varepsilon}}\right), \quad (7)$$

where $E[x]$ is the average value of batch standardization layer input, $\mathrm{Var}(x)$ is the variance of batch standardization layer input, and β and γ represent the transformation parameters to be learned during training.

By using batch standardization, the input feature distributions have the same mean and variance, and the correlation between features is removed. Although it increases the complexity of the model, it slows down the transformation process of internal covariates, greatly reduces the training convergence time, and speeds up the learning progress.

In the convolutional neural network model, overfitting often occurs if there are too many parameters and too few training samples. Overfitting is embodied as follows: the loss function of the model is small in the training stage and the prediction accuracy is high, but the loss function is large and the accuracy is low in the testing stage.

In order to solve the over fitting phenomenon, the dropout mechanism is adopted in this paper. Suppose the output of the neural network is x and the output is y. After dropout is introduced, half of the hidden neurons in the network will be hidden randomly, as shown in Figure 5, while the input and output neurons remain unchanged. Then, the input x is still propagated forward through the modified network, and the loss result is propagated back through the network. After a batch of training samples, the parameters were updated by the gradient descent method on the neurons that were not deleted. Finally, continue to repeat the process.

Because half of the hidden neurons are deleted randomly, the network structure changes. The whole dropout process is equivalent to averaging multiple different neural networks. By introducing the average effect of dropout mechanism, some opposite fitting in the network is offset, so as to achieve the effect of similar model integration and reduce overfitting.

4.3. Improved Network Structure Design. The traditional convolutional neural network contains two convolution modules and uses a single mel spectrum as the input. Due to the simple network structure, the system performance can only be improved by adjusting parameters one sidedly. However, too many parameters will make the model dependent on data, resulting in weak generalization. Therefore, changing the network structure, such as increasing the depth, to enhance the classification ability of the system for different datasets is an important means to enhance the system performance.

In recent years, deep convolution neural network has been widely used in the field of computer vision. One of the benefits brought by the increase of network depth is that the flexibility of the system is greatly enhanced. In the current widely used framework, VGGNet has been widely used because of its simple architecture and strong expansibility. VGGNet was jointly developed by the computational vision group of Oxford University and Google DeepMind. A major feature of VGGNet is that the entire neural network uses a convolution kernel size of $3 * 3$ and a maximum pool size of $2 * 2$. In addition, although VGGNet has deeper layers and more parameters than conventional neural networks, VGGNet can converge with only a few iterations because the depth of the network and the small-size filter play an implicit normalization role. On the other hand, it initializes the parameters using the data obtained by pretraining in a specific layer.

Inspired by VGGNet, the improved convolution neural network structure decides to also use the convolution kernel size of $3 * 3$ and more convolution layers to improve the classification performance. The overall architecture of the designed system is shown in Figure 6.

The whole network organization includes three layers: convolution block, mono model, and dual channel model. Among them, the convolution block is responsible for the convolution operation of the system core and includes the steps of zero filling, batch standardization, activation function, and so on. The size of each convolution kernel in the convolution block is $3 * 3$, and the zero filling size is $1 * 1$. The mono channel model consists of convolution blocks, maximum pooling, and overall average pooling steps, which is responsible for processing one of the input

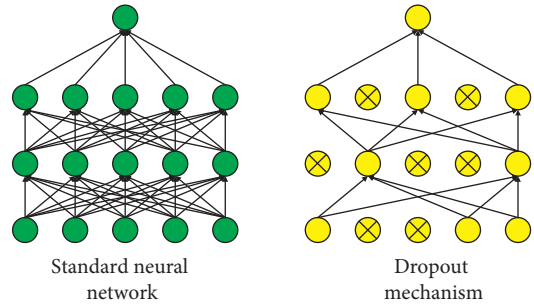

Figure 5: Dropout mechanism.

channels, and each channel is provided with four layers of convolution blocks. Similar to the design of VGGNet, the number of filters in each convolution block is doubled to 32, 64, 128, and 256.

5. Experimental Results and Analysis

5.1. Experimental Environment and Dataset. This experiment is implemented under the Ubuntu system with Linux as the kernel. The processor used in the system is an 8-core Intel E3-1270@3.8 GHz CPU, with 32 GB memory. The experiment uses Python language and introduces the external library Librosa for feature extraction. In the training part, sklearn library is introduced for unsupervised learning using the GMM model. The audio feature part includes 60-dimensional MFCC feature vector, including 20 MFCC static coefficients. For each audio file, log mel energy is extracted in 40 frequency bands using 40 ms analysis frame and 50% frame shift window. The improved convolutional neural network structure was introduced in Section 4.3. Its input feature is a single channel learner that only extracts mel spectrum. In addition, there is a dropout layer behind each CNN layer, and the dropout ratio is 30%. The parameters of convolutional neural network are shown in Table 1.

The TUT Acoustic Scenes 2017 dataset was used in this experiment, and the team responsible for collection is the audio research group of Tampere University of Technology. The dataset consists of 15 audio scenes with different labels: beach, bus, cafe/restaurant, car, city center, forest path, grocery store, family, library, subway station, office, park, community, train, and tram. All audio files are cut into segments with a length of 30 seconds, and the audio file format is wav. The dataset used in this paper is divided into development set and verification set. Among them, the development set contains 4680 audio files, and the number of files in each type of scene is 312. About 70% of the data are used to train the audio scene classification model, and the remaining 30% are used for testing. In the verification set, there are 1620 audio files, including 108 audio files of each type. The length of each audio segment is 10 seconds, and the audio of each scene is 18 minutes in total.

5.2. Identification of Performance Evaluation Indicators. The evaluation standard adopted is accuracy (ACC), that is, the ratio between the predicted correct number of samples and the total number of samples, which is calculated as follows:

$$ACC = \frac{N_{\text{true}}}{N_{\text{total}}}, \tag{8}$$

where N_{true} represents the number of correctly predicted samples and N_{total} represents the total number of samples.

5.3. Result Analysis. The classification accuracy of each scene in the development set and verification set of the proposed method is shown in Table 2.

By observing Table 2, we can find that although the average classification accuracy of the verification set is significantly lower than that of the development set, it still achieves a good result of 81.5%. It can be seen that the generalization ability of the improved system has been significantly improved, especially in the scenes of subway station, forest path, car, and home, and the classification accuracy is more than 90%, but there is still room for improvement in parks, libraries, and other scenes.

In addition, the proposed method is compared with GMM and traditional CNN in the development and verification of two datasets in order to analyze the specific performance of the improved method. The comparison between GMM and the proposed method is shown in Figure 7.

As can be seen from Figure 7, the classification accuracy of the proposed method is significantly ahead of GMM in most scenes, especially in beach, bus, library, park, and other scenes, both in the development set and verification set. Analyzing the possible reasons, on the one hand, the convolutional neural network itself has stronger learning ability for data; on the other hand, the sound field space of the above scene is large and has fixed bottom noise, which just fits the role of audio processing on the audio environment of fixed scene.

The comparison between CNN and the proposed method is shown in Figure 8.

It can be seen from the analysis of Figure 8 that although the proposed method does not achieve overall advantages in the development set, the overall classification accuracy of the proposed method in the verification set is about 19% higher than that of traditional CNN. This shows that the generalization performance of the network structure has been significantly enhanced after the introduction of VGGNet. It also shows that the flexibility performance of the system has been greatly improved by increasing the network depth and simplifying the network parameters.

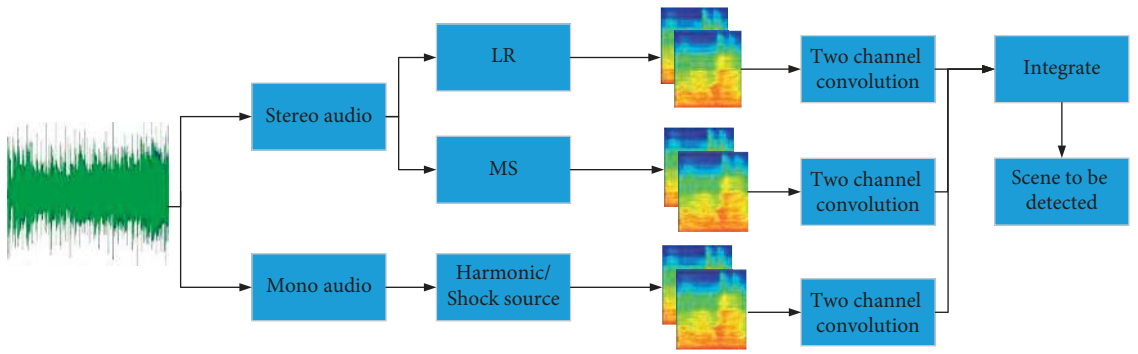

Figure 6: Improved convolution network structure design.

Table 1: Parameters of convolutional neural network.

Layer (type)	Output size	Parameter
Conv_1	(40, 500, 128)	6400
Batch standardization_1	(40, 500, 128)	160
Activation function_1	(40, 500, 128)	0
Maximum pooling_1	(8, 100, 128)	0
Dropout_1	(8, 100, 128)	0
Conv_2	(8, 100, 256)	1605888
Batch standardization_2	(8, 100, 256)	32
Activation function_ 1	(8, 100, 256)	0
Maximum pooling_1	(2, 1, 256)	0
Dropout_1	(2, 1, 256)	0
Dense (full connection layer)_1	100	51300
Dropout_3		0
Dense (full connection layer)_2	100	1515

Table 2: Scene classification results of the proposed method (%).

Audio scene	Development set	Validation set
Sandy beach	89.6	76.3
Transit	98.2	71.9
Coffee/restaurant	88.3	81.2
Automobile	99.0	92.4
Center	89.7	88.7
Forest giants	99.8	95.5
Grocery store	93.6	75.8
Home	84.1	93.6
Library	88.5	62.1
Metro station	99.2	98.1
Office	98.9	83.4
Park	80.3	66.8
Neighborhood	86.8	74.3
Train	91.4	90.6
Tram	92.7	71.0
Average	92.0	81.4

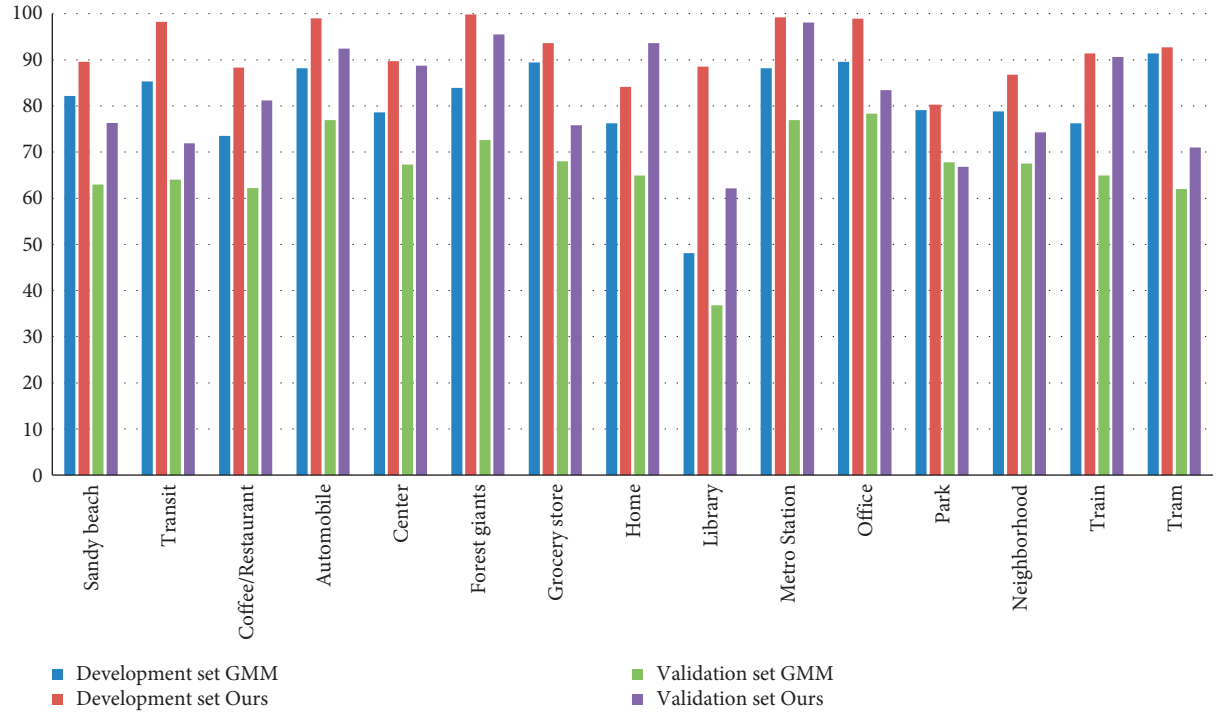

Figure 7: Comparison between GMM and the proposed method.

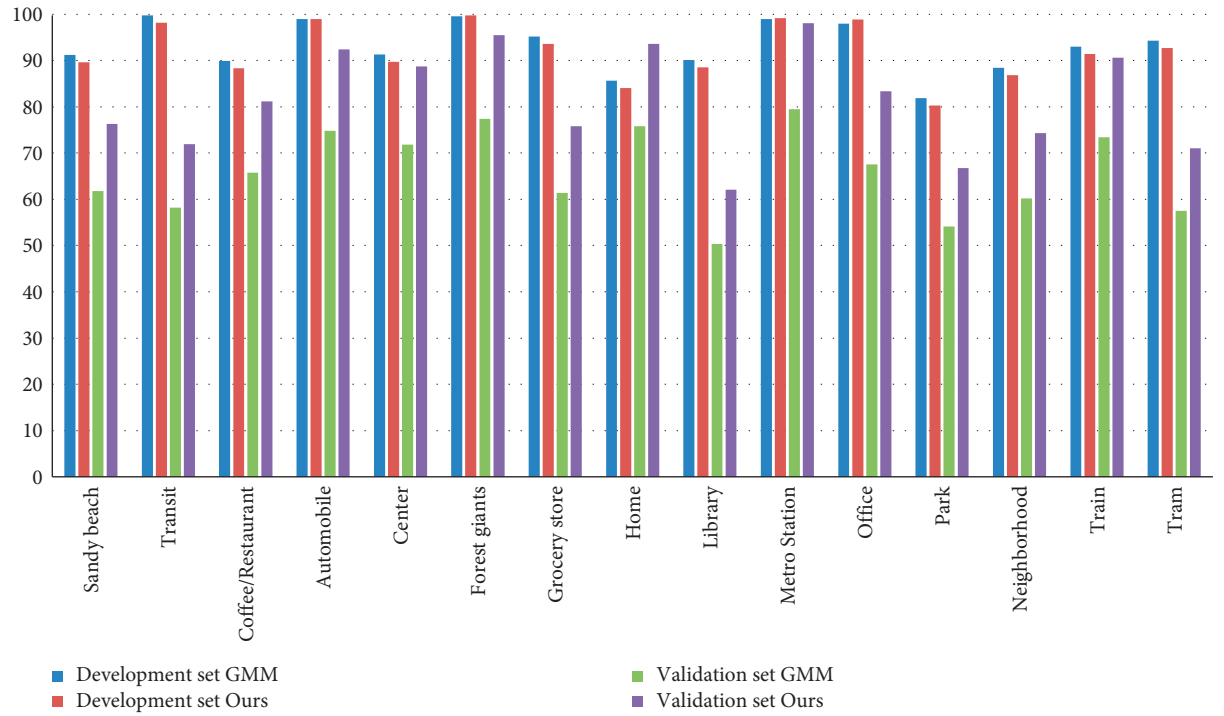

Figure 8: Comparison between CNN and the proposed method.

6. Conclusions

This paper presents an audio scene recognition method based on optimized audio processing and convolutional neural network. In audio processing, binaural representation and harmonic impact source separation are used. In terms of network structure, VGGNet-like structure is introduced to improve the flexibility of network structure. Experimental results show that compared with other existing methods, the proposed method can have better generalization performance while maintaining high recognition accuracy.

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] K. Nogueira, O. Penatti, and J. Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2016.

[2] X. Qi, C.-G. Li, G. Zhao, X. Hong, and M. Pietikäinen, "Dynamic texture and scene classification by transferring deep image features," *Neurocomputing*, vol. 171, no. 6, pp. 1230–1241, 2016.

[3] C. Gong, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geoscience & Remote Sensing Letters*, vol. 14, no. 10, pp. 1735–1739, 2017.

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[5] D. Liang and H. Ling, "Dynamic scene classification using redundant spatial scenelets[J]," *IEEE Transactions on Cybernetics*, vol. 46, no. 9, pp. 2156–2165, 2016.

[6] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Information Sciences*, vol. 348, pp. 209–226, 2016.

[7] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor scene classification," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4829–4841, 2016.

[8] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins, "Improved audio scene classification based on label-tree embeddings and convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1278–1290, 2017.

[9] A. Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1253–1265, 2017.

[10] M. Raissi and G. E. Karniadakis, "Hidden physics models: machine learning of nonlinear partial differential equations," *Journal of Computational Physics*, vol. 357, pp. 125–141, 2018.

[11] C. Voyant, G. Notton, S. Kalogirou et al., "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, no. 5, pp. 569–582, 2017.

[12] M. P. Pound, J. A. Atkinson, A. J. Townsend et al., "Deep machine learning provides state-of-the-art performance in image-based plant phenotyping," *Gigascience*, vol. 6, no. 10, pp. 1–10, 2018.

[13] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, no. C, pp. 104–116, 2017.

[14] W. Yang and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1315–1321, 2017.

[15] M. A. Alamir, "A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers," *Applied Acoustics*, vol. 172, no. 3, pp. 112–122, 2020.

[16] X. Lin, X. Wang, and L. Li, "Intelligent detection of edge inconsistency for mechanical workpiece by machine vision with deep learning and variable geometry model," *Applied Intelligence*, vol. 50, no. 7, pp. 2105–2119, 2020.

[17] T. Kim, I. Y. Jung, and Y. C. Hu, "Automatic, location-privacy preserving dashcam video sharing using blockchain and deep learning," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–23, 2020.

[18] Y. Mitsukura, "EEG signal processing for real applications," *Journal of Signal Processing*, vol. 20, no. 1, pp. 1–7, 2016.

[19] F. a. Zhao, X. Zhang, X. Mu, Z. Yi, and Z. Yang, "Learning multi-modality features for scene classification of high-resolution remote sensing images," *Journal of Computer and Communications*, vol. 06, no. 11, pp. 185–193, 2018.

[20] V. Abrol and P. Sharma, "Learning hierarchy aware embedding from raw audio for acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 157, no. C, pp. 552–559, 2020.

[21] S. Waldekar and G. Saha, "Two-level fusion-based acoustic scene classification," *Applied Acoustics*, vol. 170, no. 5, Article ID 107502, 2020.

[22] B. S. Chandra, C. S. Sastry, and S. Jana, "Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 3, pp. 710–717, 2019.

[23] S. Ravikumar and D. Kavitha, "CNN-oppositional-based Henry gas solubility optimization model for autonomous vehicle control system," *Journal of Field Robotics*, vol. 38, no. 4, pp. 1–30, 2021.

[24] R. Yang, X. Zha, K. Liu, and S. Xu, "A CNN model embedded with local feature knowledge and its application to time-varying signal classification," *Neural Networks*, vol. 142, no. 1, pp. 564–572, 2021.

[25] L. Zhao, Z. Wang, G. Zhang, and H. Gao, "Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector," *Multimedia Tools and Applications*, vol. 11, no. 2, pp. 1–19, 2020.

[26] J. Cai, J. Hu, S. Li, J. Lin, and J. Wang, "Combination of temporal-channels correlation information and bilinear feature for action recognition," *IET Computer Vision*, vol. 14, no. 8, pp. 634–641, 2020.

[27] D. Tan, H. Nguyen, D. T. Ngo, L. Pham, and H. H. Kha, "Acoustic scene classification using a deeper training method for convolution neural network," in *Proceedings of the 2019 International Symposium on Electrical and Electronics Engineering (ISEE)*, pp. 131–136, Ho Chi Minh City, Vietnam, October 2019.

[28] S. Bao, S. Ma, and C. Yang, "Multi-scale retinex-based contrast enhancement method for preserving the naturalness of color image," *Optical Review*, vol. 27, no. 6, pp. 475–485, 2020.

[29] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognition*, vol. 72, pp. 15–26, 2017.