Acoustic Scene Classification Using Aggregation of Two-Scale Deep Embeddings

Ho Ka Chon
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
hokachon@hotmail.com

Yanxiong Li*
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
eeyxli@scut.edu.cn

Wenchang Cao
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
wenchangcao98@163.com

Qisheng Huang
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
839508665@qq.com

Wei Xie
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
chester.w.xie@gmail.com

Wenfeng Pang
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
wenfengpang@gmail.com

Jiyue Wang
School of Electronic and Information Engineering, South China University of Technology,
Guangzhou, China
jiyuewang@outlook.com

2021 IEEE 21st International Conference on Communication Technology (ICCT) | 978-1-6654-3206-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCT52962.2021.9658086

***Abstract*—Acoustic scene classification (ASC) is a topic related to the field of machine listening whose important role is to recognize and categorize audio data in a predefined label which describes a scene location. In most of the state-of-the-art works for ASC, hand-crafted features and single-scale deep embeddings were adopted as the input of back-end classifiers. Inspired by the success of multi-scale deep embeddings in the field of computer vision, we propose an ASC method by aggregating two-scale deep embeddings that are independently learned by two convolutional neural networks (CNNs). We perform ASC experiments on two official datasets of the challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), i.e., DCASE-2019 and DCASE-2017. Experimental results show that the proposed method using the aggregation of two-scale deep embeddings improves the performance of the ASC system. The proposed method obtains the improvement of classification accuracies by 0.11 and 0.09 on DCASE-2019 and DCASE-2017 respectively compared to the baseline system. Code is available: *https://github.com/hokachon/Two-scale-Agg*.**

Keywords—acoustic scene classification, two-scale deep embedding, convolutional neural network

I. INTRODUCTION

ASC is the task of classifying the type of environments from the sounds they produce [1], [2]. Humans rely on many different environmental sounds to detect danger and improve understanding [3]. In a general way, audio is an advantageous modality that compliments visual data such as images and videos, and has the advantage of being easily collected and stored. There are many applications and technologies that can hear and understand different environmental sounds, such as context-aware devices [4], [5], audio surveillance [6], and audio-based multimedia analysis [7], [8].

Although many efforts were recently made on ASC, it is still a tough issue far from being solved [2], [9]. In our opinions, the difficulty mainly originates from two aspects. First, acoustic scenes have quite complex and diverse characteristics. The foreground and background acoustic events contained in each audio sample of the same class of acoustic scenes are not always identical, whereas audio samples of different class of acoustic scenes include some identical acoustic events in most time. As a result, the intra-class divergences of one acoustic scene are large, while the inter-class distances among different class of acoustic scenes become small. Second, time-frequency scales of various acoustic scenes are not consistent. The variations in duration and spectrum distribution are relatively large. Hence, extracting and transforming features under one fixed scale might be not effective for ASC.

To overcome the difficulties above, we propose an ASC method using aggregation of two-scale deep embeddings. In the proposed method, two CNNs with different architectures are used to learn two-scale deep embeddings whose concatenation is then fed into the full-connection and Softmax layers for obtaining classification result. The motivation for using the aggregation of two-scale deep embeddings for ASC is that two-scale deep embeddings are expected to be effective for representing characteristic differences among different acoustic scenes and thus obtain a better result compared to single-scale deep embedding and hand-crafted features. Using two CNNs as the extractor of deep embeddings helps to get more feature maps with different resolutions and thus helps to acquire more discriminative information for ASC. Mel frequency cepstral coefficient (MFCC) is used as the input of CNNs for extracting two-scale deep embeddings due to its excellent performance in most of audio processing tasks. CNN is used to produce two-scale deep embeddings because it has powerful ability to learn discriminative information by using some effective operations such as convolution, pooling.

* Corresponding author: Yanxiong Li (eeyxli@scut.edu.cn).

This work was partly supported by national natural science foundation of China (61771200, 62111530145), international scientific research collaboration project of Guangdong Province, China (2021A0505030003), and Guangdong basic and applied basic research foundation, China (2021A1515011454). Ho Ka Chon is an international master student who is sponsored by the Chinese government.

The rest of the paper is organized as follows. In Section II, we briefly introduce related works. We describe the proposed method and experiments in Sections III and IV, respectively. Finally, we conclude this paper in Section V.

II. Related Works

The development of ASC techniques is promoted by some evaluation campaigns, such as the Classification of Events, Activities and Relationships (CLEAR) [10], [11], and the DCASE challenge [12]-[14]. Feature extraction (learning) is one critical step for an ASC system with higher performance [9]. Researchers focused on their efforts on designing discriminative features for improving the performance of their ASC systems. Related works are summarized as follows.

Many hand-crafted features were adopted in previous works, which were proved to be effective for ASC in specific cases. These hand-crafted features mainly include: logarithm Mel-band energy, linear prediction cepstral coefficients, spectral centroid magnitude cepstral coefficients, MFCCs, Mel-frequency discrete wavelet coefficients, spectral flux, spectrogram, Gabor filterbank, cochleogram, I-vector, higher-order ambisonic features, histogram of gradients of time-frequency representations, hash features, and local binary patterns [14]-[25]. However, these hand-crafted features are difficult to be generalized for other situations and thus lack a flexibility in practice. In addition, most of hand-crafted features need to be combined with other features for obtaining satisfactory results, leading to a large feature vector [22].

To overcome the drawbacks of hand-crafted features, transformed features were proposed using non-negative matrix factorization (NMF) [26], [27] and deep neural network (DNN) [28]-[34]. The ASC methods using NMF-based features represent the state-of-the-art performance in the field of matrix factorization [26], [27]. On the other hand, the ASC methods using DNN-based features were proposed and outperformed the methods using both hand-crafted features and NMF-based features under the condition of the availability of sufficient training data [13].

In addition, CNN was adopted as a dominant feature extractor to transform input feature vectors at various convolutional layers. The transformed vector outputted from one layer (e.g., the final convolutional layer) was used as the deep embedding [9], [29]-[34]. Here, the transformed feature outputted from one layer of a CNN is called single-scale deep embedding. Recent studies demonstrated that it was still of insufficiency to adopt only single-scale deep embedding for representing the complex time-frequency characteristics of various acoustic scenes [35].

III. Method

The flowchart of the proposed method is shown in Fig. 1. The proposed method consists of two parts: extractors of two-scale deep embeddings and classifier.

A. Extractors of Deep Embeddings

As shown in Fig. 1, MFCC is extracted from each audio sample and then fed into CNN 1 and CNN 2. The deep embeddings of scale 1 and scale 2 are extracted by these two CNNs with different structures which model acoustic scene at different time-frequency scales. Hence, they are expected to own complementary and discriminative information with various resolutions.

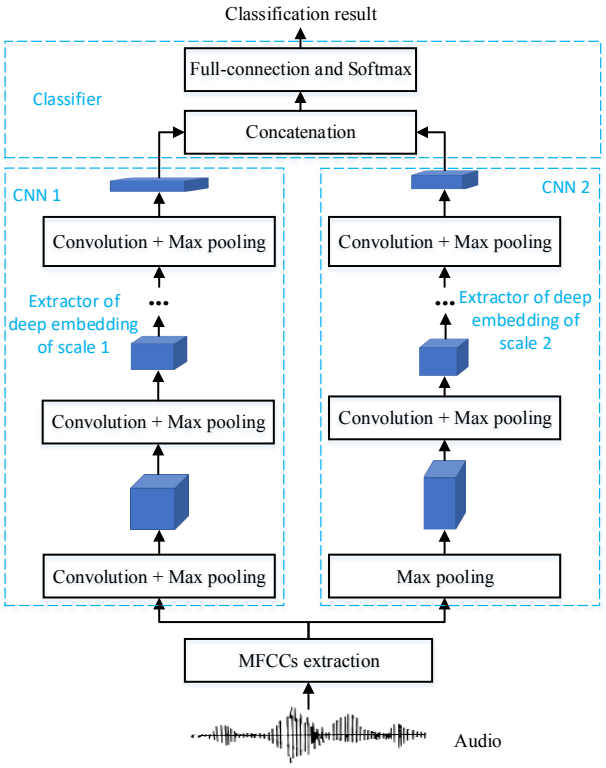

Fig. 1. Flowchart of the proposed method.

The common hyperparameters of these two CNNs are learning rate, batch size, optimizer, loss, training epoch and activation. For determining learning rate, it is tuned many times from 0.001 to 0.0001, and the best accuracy is produced by using 0.001. Batch size of 64, optimizer Adam and loss function of categorical cross-entropy are used since they perform well for audio classification. The number of training epochs is from 50 to 100.

The CNN 1 consists of 10 layers where the filter depth ranges from 64 to 1024, and the kernel size in each filter depth is 3×3 to maintain the weight sharing and to avoid heavy computational load. After each convolutional layer, max pooling (MP) layer is applied to down-sample the feature maps by calculating the maximum value in each section of the feature maps. Dropout (DP) layer with 0.3 is used to achieve acceptable variance and to avoid the possibility of overfitting. Batch normalization (BN) and Rectified Linear Unit (ReLU) are also used in the network and finally flatten function is applied to convert the feature map into one-dimension. The structure of CNN 2 is different from that of CNN 1. CNN 2 consists of 11 layers with filter depth ranging from 128 to 2048, where a max pooling layer and batch normalization are used as the first layer of CNN 2. That is, it starts to down-sample the feature map before feeding to the convolutional layers. The kernel size of CNN 2 is the same as CNN 1. After a convolutional layer, max pooling layer and dropout are used with the same values as CNN 1. Finally, a flatten function is used to convert the feature map

into one-dimension. The parameters of these two CNNs are presented in Fig. 2.

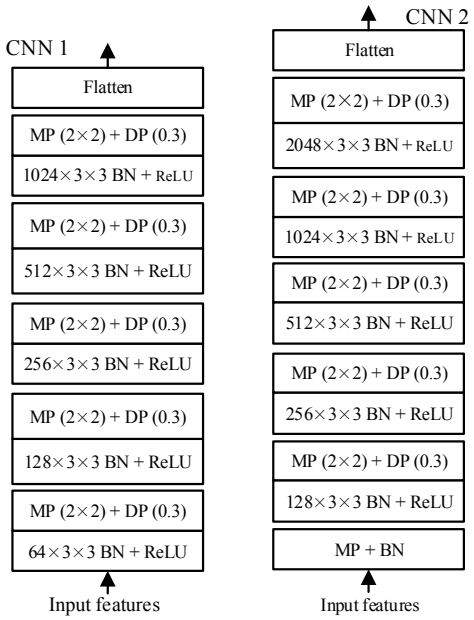

Fig. 2. The parameters of CNN 1 and CNN 2.

It should be noted that we apply the flatten function in both CNN 1 and CNN 2, because they are separate networks and thus have different outputs. After getting the outputs of these two networks, the concatenation is applied to combine their outputs where both batch normalization and dropout of 0.3 are adopted.

B. Classifier

The classifier consists of two full-connection layers and one Softmax layer. In the classifier, the concatenation of two-scale deep embeddings is fed into the full-connection layers for further transformation, and then used as the input of the Softmax layer for producing the final classification result. The activation function of the output layer is Softmax whose neurons are 10 and 15 for DCASE-2019 and DCASE-2017 datasets, respectively.

IV. Experiments

In this section, we describes experimental datasets, setup and results.

A. Experimental Datasets

Experimental datasets are DCASE-2019 and DCASE-2017 datasets. DCASE-2019 dataset has 10 classes of acoustic scenes, including *Airport*, *Shopping mall*, *Metro station*, *Pedestrian street*, *Public square*, *Street traffic*, *Tram*, *Metro*, *Urban park*, and *Bus*. Its audio samples were recorded in 10 different cities, and saved in WAV format with sampling frequency rate of 48 kHz. Each class has a total of 1440 samples with a duration of 10 seconds per file which makes the total length of the DCASE-2019 dataset of 40 hours. DCASE-2017 dataset has 15 classes of acoustic scenes which include indoor and outdoor environment sounds, including *Bus*, *Car*, *City center*, *Café/restaurant*, *Forest path*, *Grocery store*, *Home*, *Beach*, *Library*, *Metro station*, *Office*, *Residential area*, *Train*, *Tram*, and *Urban park*. All recordings were saved in WAV format with the sampling frequency rate of 44.1 kHz. Each class has 420 samples with a duration approximately 10 seconds per sample, which makes the total length of the DCASE-2017 dataset of 17.5 hours.

B. Experimental Setup

All the programming is done using the Pycharm with Python version 3.6. Some of the libraries are used in the algorithm are: Pandas, Numpy and Matplotlib for data processing. SoundFile library is used to read and write the audio files and the feature extraction of MFCC is done by Python_speech_features library. The neural network design is done by using Keras library where the confusion matrix calculation is done by using Scikit-learn. Other libraries are also used, such as Tqdm for showing the progression bar whenever the data is processed.

The datasets are provided by the organizer of the DCASE-2019 challenge where only the training dataset are publicly available. For DCASE-2017 dataset, 80% training data and 20% testing data are publicly available. From the training data, 3744 samples are used as training subset and 936 samples are used as validation subset, respectively. For the testing data, 1620 samples are used as testing subset in the experiments. DCASE-2019 dataset is manually divided into two parts: training data and testing data. The training and testing data account for 75% and 25%, respectively. In the training data of DCASE-2019 dataset, 7200 samples are used as training subset, while 2880 samples are used as validation subset. The testing data of DCASE-2019 dataset consists of 4320 samples.

The performance metric for ASC is classification accuracy (CA) which is defined by the number of correctly predicted audio samples among the total number of audio samples. Each audio sample is regarded as an independent test sample.

C. Experimental Results

A baseline system is designed using a CNN with 4 layers where MFCC is used as input feature [36]. The parameters of the baseline system are optimally tuned on experimental datasets. To know the contribution of each scale of deep embedding in the proposed method to the ASC performance, ablation experiments are also done by using CNN 1 or CNN 2 only for learning deep embedding (as shown in Fig. 1), i.e., the CNN 1 based method and CNN 2 based method.

Tables I and II show the CAs obtained by the baseline system, the CNN 1 based method, the CNN 2 based method and the proposed method for each class of acoustic scenes on DCASE-2019 and DCASE-2017 datasets, respectively. The CAs obtained by the baseline system are 0.60 and 0.64 on DCASE-2019 and DCASE-2017 datasets, respectively. The CNN 1 based method obtains the CAs of 0.63 and 0.66 on DCASE-2019 and DCASE-2017 datasets, respectively. The CNN 2 based method obtains the CAs of 0.65 and 0.69 on DCASE-2019 and DCASE-2017 datasets, respectively. The proposed method obtains a CA of 0.71 on DCASE-2019 dataset and a CA of 0.73 on DCASE-2017 dataset. Compared to the baseline system, the proposed method obtains improvements of CA by 0.11 and 0.09 on DCASE-2019 and DCASE-2017 datasets, respectively. Similarly, in terms of the metric of CA, the proposed method also outperforms both the CNN 1 based method and the CNN 2 based method, in which single-scale deep embedding is used.

In conclusion, the results above indicate that using the aggregation of two-scale deep embeddings is better than using single-scale deep embedding. The reason is probably that the two-scale deep embeddings can learn more discriminative information from feature maps during training.

TABLE I. CAS OBTAINED BY VARIOUS METHODS ON DCASE-2019 DATASET

Class	Baseline	CNN 1	CNN 2	Proposed
Airport	0.80	0.60	0.63	0.57
Bus	0.65	0.70	0.75	0.70
Metro	0.73	0.72	0.76	0.78
Metro Station	0.47	0.60	0.59	0.75
Park	0.94	0.80	0.91	0.88
Public Square	0.28	0.50	0.42	0.56
Shopping Mall	0.40	0.40	0.62	0.86
Street Pedestrian	0.44	0.60	0.32	0.39
Street Traffic	0.87	0.95	0.95	0.93
Tram	0.34	0.50	0.58	0.66
Average	**0.60**	**0.63**	**0.65**	**0.71**

TABLE II. CAS OBTAINED BY VARIOUS METHODS ON DCASE-2017 DATASET

Class	Baseline	CNN 1	CNN 2	Proposed
Beach	0.08	0.25	0.40	0.86
Bus	0.49	0.50	0.55	0.62
Café/Restaurant	0.54	0.61	0.70	0.50
Car	0.75	0.78	0.80	0.78
City Center	0.82	0.79	0.89	0.83
Forest Path	0.76	0.81	0.75	0.76
Grocery Store	0.53	0.59	0.65	0.69
Home	0.81	0.89	0.88	0.80
Library	0.66	0.64	0.40	0.63
Metro Station	0.99	0.97	0.98	0.91
Office	0.70	0.69	0.55	0.62
Park	0.25	0.33	0.45	0.63
Residential Area	0.86	0.86	0.75	0.75
Train	0.78	0.73	0.70	0.70
Tram	0.59	0.59	0.65	0.79
Average	**0.64**	**0.66**	**0.69**	**0.73**

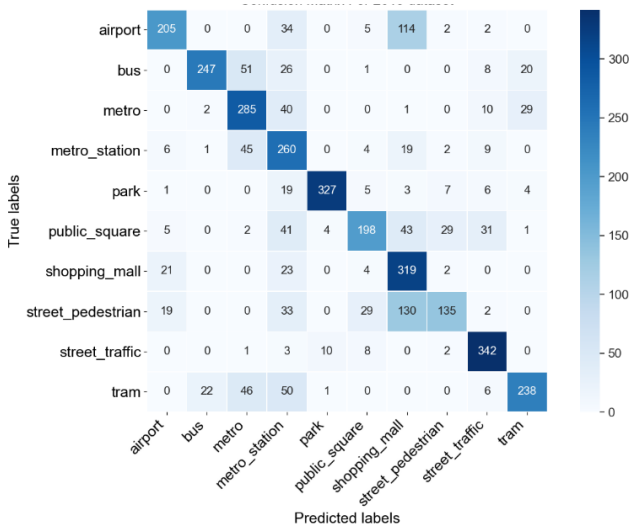

Fig. 3. Confusion matrix for our method on DCASE-2019 dataset.

The confusion matrix is able to show the numbers of audio samples that are correctly and incorrectly predicted. The confusion matrix for our method on DCASE-2019 dataset is shown in Fig. 3. We can see that *Metro*, *Park*, *Shopping mall* and *Street traffic* are the classes which have high accuracy in classification result. Other classes which have low accuracy are *Airport*, *Public square* and *Street pedestrian*. *Airport* is mis-predicted as *Shopping mall* 114 times (true negative) and 21 times (false negative). This is the problem of audio samples having crowds and people noise which could lead the network to predict them incorrectly.

The confusion matrix obtained by the proposed method on DCASE-2017 dataset is depicted in Fig. 4. It can be seen that *Beach*, *Car*, *City center*, *Home* and *Metro station* classes have high accuracy in classification results. Some other classes which are somehow complex to be correctly predicted are *Bus*, *Café/Restaurant*, *Library* and *Office*. The main reason for this kind of problem to occur may be due to the noise in the audio samples of these classes of acoustic scenes. For example, in Fig. 4, the *Car* class and *Train* class were mis-predicted 23 times and 14 times respectively, because they are all automotive vehicles which would have similar time-frequency properties.

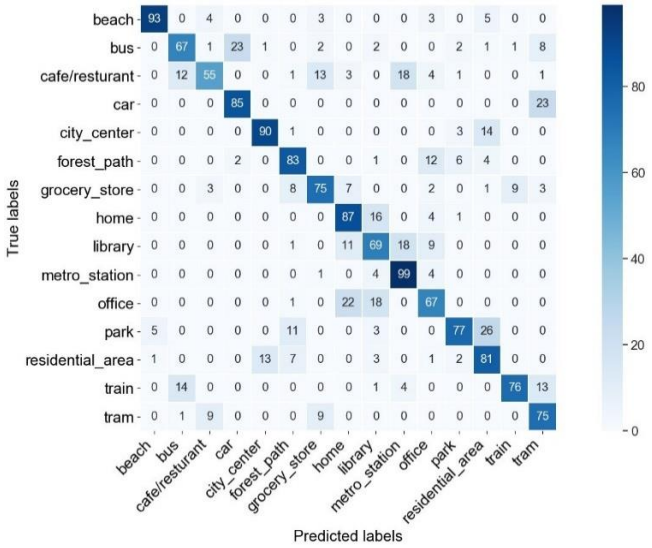

Fig. 4. Confusion matrix for our method on DCASE-2017 dataset.

V. CONCLUSIONS

In this work, we propose a method for ASC using the aggregation of two-scale deep embeddings. Two CNNs are used in this work with different structures, and then the outputs of these two CNNs are concatenated and used as the input of back-end classifier. The results show that the proposed method exceeds the baseline system. The shortcoming of the proposed method is that its parameter size is larger than that of the baseline system.

It should be noted that the work in this paper is a preliminary study for the ASC task using multi-scale deep embeddings because the deep embeddings of only two scales are adopted in our method. In next work, we will further investigate the ASC problem using aggregation of deep embeddings of more than two scales. In addition, we will explore other framework of

neural networks to learn multi-scale deep embeddings for decreasing its parameter size with higher performance for ASC.

ACKNOWLEDGMENT

We'd like to thank Dr. Emmanouil Benetos from Queen Mary University of London for his constructive comments and suggestions on this paper.

REFERENCES

[1] S. Mun, and S. Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *Proc. of ICASSP*, 2019, pp. 845-849.

[2] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, and X. Feng, "Acoustic scene classification using deep audio feature and BLSTM network," in *Proc. of ICALIP*, Jul. 2018, pp. 371-374.

[3] J. Abeber. "A review of deep learning-based methods for acoustic scene classification." *Applied Sciences*, vol. 10, pp. 16-21, 2020.

[4] V. Abrol, and P. Sharma, "Learning hierarchy aware embedding from raw audio for acoustic scene classification," *IEEE/ACM TASLP*, vol. 28, pp. 1964-1973, 2020.

[5] Y. Zeng, Y. Li, Z. Zhou, R. Wang, and D. Lu, "Domestic activities classification from audio recordings using multi-scale dilated depthwise separable convolutional network," in *Proc. of IEEE MMSP*, pp. 1-5, 2021.

[6] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58043-58055, Oct. 2018.

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE TMM*, vol. 17, no. 10, pp.1733-1746, Oct. 2015.

[8] S. Chandrakala, and S.L. Jayalakshmi, "Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition," *IEEE TMM*, vol. 22, no. 1, pp. 3-14, Jan. 2020.

[9] Y. Li, M. Liu, W. Wang, Y. Zhang, and Q. He, "Acoustic scene clustering using joint optimization of deep embedding learning and clustering iteration," *IEEE TMM*, vol. 22, no. 6, pp. 1385-1394, Jun. 2020.

[10] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," *Lecture Notes in Computer Science*, vol. 4122, pp. 311-322, 2007.

[11] R. Stiefelhagen, K. Bernardin, R. Bowers, R.T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 Evaluation," *Lecture Notes in Computer Science*, vol. 4625, pp. 3-34, 2008.

[12] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 challenge: generalization across devices and low complexity solutions," in *Proc. of DCASE Workshop*, 2020, pp. 1-5. Online: https://arxiv.org/abs/2005.14623

[13] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M.D. Plumbley, "Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge," *IEEE/ACM TASLP*, vol. 26, no. 2, pp. 379-393, Feb. 2018.

[14] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," in *Proc. of DCASE Workshop*, 2017, pp. 98-102.

[15] C. Paseddula, and S.V. Gangashetty, "Late fusion framework for acoustic scene classification using LPCC, SCMC, and log-Mel band energies with deep neural networks," *Applied Acoustics*, vol. 172, pp. 1-12, 2021.

[16] A. Rakotomamonjy, and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 142-153, Jan. 2015.

[17] A. Jiménez, B. Elizalde, and B. Raj, "Acoustic scene classification using discrete random hashing for Laplacian kernel machines," in *Proc. of ICASSP*, 2018, pp. 146-150.

[18] W. Yang, and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM TASLP*, vol. 25, no. 6, pp.1315-1324, Jun. 2017.

[19] S. Abidin, R. Togneri, and F. Sohel, "Spectrotemporal analysis using local binary pattern variants for acoustic scene classification," *IEEE/ACM TASLP*, vol. 26, no. 11, pp. 2112-2121, Nov. 2018.

[20] M.D. McDonnell, and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *Proc. of ICASSP*, 2020, pp. 141-145.

[21] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *Proc. of IEEE ICASSP*, 2020, pp. 286-290.

[22] S. Waldekar, and G. Saha, "Two-level fusion-based acoustic scene classification," *Applied Acoustics*, vol. 170, pp. 1-11, 2020.

[23] S. Waldekar, and G. Saha, "Analysis and classification of acoustic scenes with wavelet transform-based Mel-scaled features," *Multimed. Tools Appl.*, vol. 79, pp. 7911-7926, 2020.

[24] J. Xie, and M. Zhu, "Investigation of acoustic and visual features for acoustic scene classification," *Expert Systems with Applications*, vol. 126, pp. 20-29, 2019.

[25] Z. Lin, Y. Li, Z. Huang, W. Zhang, Y. Tan, Y. Chen, and Q. He, "Domestic activities clustering from audio recordings using convolutional capsule autoencoder network," in *Proc. of IEEE ICASSP*, 2021, pp. 835-839.

[26] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM TASLP*, vol. 25, no. 6, pp. 1216-1228. June 2017

[27] A. Rakotomamonjy, "Supervised representation learning for audio scene classification," *IEEE/ACM TASLP*, vol. 25, no. 6, pp. 1253-1265, Jun. 2017

[28] Y. Li, X. Zhang, H. Jin, X. Li, Q. Wang, Q. He, and Q. Huang, "Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic events detection," *Multimed. Tools Appl.*, vol. 77, no. 1, pp. 897-916, Jan. 2018.

[29] Y. Wu, and T. Lee, "Time-frequency feature decomposition based on sound duration for acoustic scene classification," in *Proc. of ICASSP*, 2020, pp. 716-720.

[30] X. Bai, J. Du, J. Pan, H. Zhou, Y. Tu, and C. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *Proc. of ICASSP*, 2020, pp. 656-660.

[31] L. Pham, I. McLoughlin, H. Phan, R. Palaniappan, and A. Mertins, "Deep feature embedding and hierarchical classification for audio scene classification," in *Proc. of IJCNN*, 2020, pp. 1-7.

[32] S.S.R. Phaye, E. Benetos, and Y. Wang, "SubSpectralNet - Using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *Proc. of ICASSP*, 2019, pp. 825-829.

[33] Y. Wu, and T. Lee, "Enhancing sound texture in CNN-based acoustic scene classification," in *Proc. of ICASSP*, 2019, pp. 815-819.

[34] Z. Ren, Q. Kong, J. Han, M.D. Plumbley, and B.W. Schuller, "Attention-based atrous convolutional neural networks: visualisation and understanding perspectives of acoustic scenes," in *Proc. of ICASSP*, 2019, pp. 56-60.

[35] L. Yang, L. Tao, X. Chen, and X. Gu, "Multi-scale semantic feature fusion and data augmentation for acoustic scene classification," *Applied Acoustics*, vol. 163, pp. 1-10, 2020.

[36] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of DCASE 2017 Workshop*, 2017, pp. 85-92.