# 摘要

音频场景分类（Acoustic Scene Classification）是一项通过音频分析使设备能够理解其所处环境的任务，属于计算机听觉场景领域的一个分支。目前该技术已经广泛使用于智能可穿戴设备、机器人导航系统、上下文感知服务等应用场景。得益于近年来深度学习热潮的掀起，将深度学习中的方法应用到音频场景分类中的例子越来越多。通过引入诸如卷积神经网络等手段，使场景分类准确率有了可观的提升，甚至能使机器超过人类水平。为了探究卷积神经网络在音频场景分类领域的适用性并探究系统性能的提升方法，本文中设计了几组系统并进行实验以验证，主要工作及贡献如下：

（1）针对普通单声道模型不能有效利用音频中空间信息导致分类准确率不佳的问题，文中使用了双耳表示法及谐波-冲击源分离法对原始音频进行处理，使系统在具有明显空间特征的场景中分类准确率得到了明显的提升。（2）针对一般卷积神经网络应对不同数据泛化能力不佳的问题，本文尝试借鉴图像识别领域中的VGGNet结构，在增加卷积神经网络深度的同时增加了系统的灵活性，最终在不同的数据上取得了更好的泛化效果。（3）本文还尝试使用集成学习领域中流行的Stacking方法将多个基于不同特征的独立模型融合，融合后的系统相比其中的子系统分类性能又有了进一步的提升。（4）本文还在卷积神经网络训练时参数调整阶段将训练时间考虑到系统性能评估的要素中去，使得系统在性能逐渐提升的同时训练时间也处于可控范围内。

本文先从设计基于MFCC和GMM的基线系统开始，用传统机器学习的方法构造了一个典型的基线系统作为之后系统的对照组。接着研究一般的基于卷积神经网络的音频场景分类方法与原理，并实现了一个具有两层卷积结构的基本系统。训练系统时通过调整参数以发挥其分类潜力，并分析系统在各类别上的分类准确率与混淆矩阵，借此评价其优点与不足。最后，根据之前系统体现出的问题，从特征处理和网络结构两个方面进行改进，得到了最终的改进系统。在改进系统中分类的准确率相对前两个系统有了明显的提升。

最终得出的结论是，在音频场景分类这一领域中，卷积神经网络相比于一般机器学习方法所需要的数据更少、学习能力更强。在设计卷积神经网络时应注意网络的灵活性，避免因参数过多而造成系统的泛化能力不佳。此外，通过引入集成学习的方法将多组模型进行融合通常可以显著的提升性能，但集成时应注意模型间的独立性。最后，在音频特征提取阶段如果能利用到立体声信息，可以提升系统对空间的感知能力，进而提升分类准确率。

关键词：音频场景分类；卷积神经网络；梅尔频率倒谱系数；集成学习

# Abstract

Acoustic Scene Classification (ASC) is a task that enables devices to make sence of their environment through audio analysis，which belongs to a branch of the computational auditory scene analysis (CASA). At present, ASC has been widely used in intelligent wearable devices, robotics navigation systems, context-aware services and so on. Thanks to the upsurge of deep learning in recent years, there are more and more examples the learning methods applied to audio scene classification. By introducing means such as Convolutional Neural Networks(CNN), the accuracy of scene classification has been greatly improved, and even the machine can be made to exceed the human level. In order to explore the applicability of CNN in ASC field and the ways to improve the performance of the system, several sets of systems are designed and tested to verify it in this paper. The main work and contributions are as follows:

(1) The problem that the ordinary mono model can not effectively utilize the spatial information in audio leads to poor classification accuracy. In this paper, the binaural representation and the harmonic-percussive source separation method are used to process the original audio. The classification accuracy of the system in the scene with obvious spatial characteristics has been significantly improved (2) In view of the problem that the general CNN is not good at generalizing different data, this paper attempts to draw from the VGGNet structure in the field of image recognition, which increases the flexibility of the system, While increasing the depth of CNN. Finally, a better generalization effect was achieved on different data. (3) This paper also attempts to use the popular Stacking method in the field of ensemble learning to fuse multiple independent models based on different features. The merged system performance is further improved compared with that of subsystem classification. (4) In the parameter adjustment phase of convolutional neural network training, the training time is taken into account in system performance evaluation, so that the training time is also in the controllable range while the performance is gradually improved.

This paper begins with the design of a baseline system based on MFCC and GMM, using traditional machine learning method to construct a typical baseline system as a comparison for further systems. Then, the general method and principle of ASC based on CNN are studied, and a basic system with two-layer convolution structure is implemented. When the system trained, the parameters are adjusted to give full play to its classification potential, and the classification accuracy and confusion matrix of the system in each category are analyzed to evaluate its advantages and disadvantages. Finally, according to the problems reflected in the previous system, the two aspects of feature processing and network structure are improved, and the final improved system is obtained. The classification accuracy in improved systems has been significantly improved over the first two systems.

The final conclusion is that in the field of ASC, convolutional neural networks require less data and are more capable of learning than general machine learning methods. The network flexibility should be paid attention to while the CNN designed to avoid the poor generalization ability of the system caused by too many parameters. In addition, the integration of multiple sets of models by introducing ensemble learning methods can generally results in significant performance improvements. However, we should pay attention to the independence between models when integrating. Finally, if the stereo information can be utilized in the audio feature extraction stage, the system's perception of space can be improved, thereby the classification accuracy will be improved as well.

Keywords: Acoustic Scene Classification; Convolutional Neural Network; Mel frequency cepstral coefficient; ensemble learning

# 参考文献

Eronen A J, Peltonen V T, Tuomi J T, et al. Audio-based context recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, 14(1):321-329.

Malkin R G, Waibel A. Classifying user environment for mobile applications using linear autoencoding of ambient audio[C]// ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing. 2005.

Krijnders J, t Holt G. Tone-fit and MFCC scene classification compared to human recognition[J]. Energy [dB], 2013, 400(450): 500.

Sawhney N, Maes P. Situational awareness from environmental sounds[J]. Tech-nical Report, Massachusetts Institute of Technology, 1997.

Clarkson B, Sawhney N, Pentland A. Auditory context awareness via wearable computing[J]. Energy, 1998, 400(600): 20.

Patil K, Elhilali M. Multiresolution auditory representations for scene classification[J]. cortex, 2002, 87(1): 516-527.

Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

Piczak K J. Environmental sound classification with convolutional neural networks[C]//2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015: 1-6.

Ballas J A. 1993. Common factors in the identification of an assortment of brief everyday sounds[J]. Journal of experimental psychology: human perception and performance, 19(2): 250.

Peltonen V T K, Eronen A J, Parviainen M P, et al. 2001. Recognition of everyday auditory scenes: potentials, latencies and cues[J]. PREPRINTS-AUDIO ENGINEERING SOCIETY.

Dubois D, Guastavino C, Raimbault M. 2006. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories[J]. Acta acustica united with acustica, 92(6): 865-874.

Tardieu J, Susini P, Poisson F, et al. 2008. Perceptual study of soundscapes in train stations[J]. Applied Acoustics, 69(12): 1224-1239.

Eronen A, Tuomi J, Klapuri A, et al. 2003. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]//Acoustics, Speech, and Signal Processing, 2003.

Lin M, Chen Q, Yan S. 2013. Network in network[J]. arXiv preprint arXiv:1312.4400.

Davis S，Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEe transactions on acoustics, speech，and signal processing, 1980, 28 (4): 357-366.

Lidy T, Schindler A. 2016. CQT-based convolutional neural networks for audio scene classification[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). DCASE2016 Challenge, 90: 1032-1048.

Aytar Y, Vondrick C, Torralba A. 2016. Soundnet: Learning sound representations from unlabeled video[C]//Advances in Neural Information Processing Systems. 892-900.

Santoso A, Wang C Y, Wang J C. Acoustic scene classification using network-in-network based convolutional neural network[R]. DCASE2016 Challenge, Tech. Rep, 2016.

Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE transactions on speech and audio processing, 1995, 3(1): 72-83.

Dempster A . Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 1977, 39.

Nuttall A H . Some Integrals Involving the Q-Function[J]. IEEE Transactions on Information Theory, 1972, 21(1):95-96.

Rijsbergen C J V. Information Retrieval[M]. 1979.

Wiesel T N , Hubel D H . EXTENT OF RECOVERY FROM THE EFFECTS OF VISUAL DEPRIVATION IN KITTENS[J]. Journal of Neurophysiology, 1965, 28(6):1060-1072.

Fukushima K. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

Hinton G E, Osindero S, Teh Y W, et al. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

Salakhutdinov R, Mnih A, Hinton G E, et al. Restricted Boltzmann machines for collaborative filtering[C]. international conference on machine learning, 2007: 791-798.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function[J]. Journal of statistical planning and inference, 2000, 90(2): 227-244.

|  |
| --- |
|  |

Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.

周志华. 机器学习[M]. 清华大学出版社, 2016.

Eghbal-Zadeh H, Lehner B, Dorfer M, et al. CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks[J]. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

Ono N, Miyamoto K, Le Roux J, et al. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram[C]//2008 16th European Signal Processing Conference. IEEE, 2008: 1-4.

Park J, Shin J, Lee K. Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(5): 1061-1074.

Moore A W, Jorgenson J W. Median filtering for removal of low-frequency background drift[J]. Analytical chemistry, 1993, 65(2): 188-191.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

Wolpert D H. Stacked generalization[J]. Neural networks, 1992, 5(2): 241-259.

Van der Laan M J, Polley E C, Hubbard A E. Super learner[J]. Statistical applications in genetics and molecular biology, 2007, 6(1).

# 致 谢

本