DOI:10.19651/j.cnki.emt.2108486

基于频谱位移模块的环境声音识别方法*

李传坤[1,3] 郭锦铭[1,3] 李 剑[1,2] 孙袖山[1,3]
(1.中北大学信息与通信工程学院 太原 030051;2.中北大学省部共建动态测试技术国家重点实验室 太原 030051;
3.中北大学无损检测技术中心 太原 030051)

摘 要:针对卷积操作只能提取局部频谱信息,不能有效地挖掘频谱之间相关信息的问题,提出了一种基于频谱位移模块的神经网络。该网络采用密集卷积神经网络的架构,并在支路上使用频谱位移模块实现频谱信息之间的交互。利用这种频谱移位取代了频谱间的下采样操作,实现了频谱的全局化特征提取,同时避免了下采样过程中信息的丢失,进一步地提高了频谱特征图质量。并在公开的数据集ESC10和ESC50上验证频谱位移密集模块,在两种数据集的分类准确度分别达到了96.00%和88.75%,与原有的网络相比准确度分别提升了2.1%和2.25%。实验结果表明,和现有的其他卷积神经网络方法相比,所提出的网络能够更好有效地挖掘全局时频信息,具有更高的识别准确率。

关键词:环境声音识别;卷积神经网络;频谱位移;密集神经网络;深度学习
中图分类号:TP391 **文献标识码**:A **国家标准学科分类代码**:510.4040

Environmental sound recognition method based on spectrum shift module

Li Chuankun[1,3] Guo Jinming[1,3] Li Jian[1,2] Sun Xiushan[1,3]
(1. School of Information and Communication Engineering, North University of China, Taiyuan 030051, China;
2. State Key Laboratory of Dynamic Testing Technology, North University of China, Taiyuan 030051, China;
3. NDT Technology Center, North University of China, Taiyuan 030051, China)

Abstract: Convolutional operation only extracted local time-frequency information, and cannot effectively mine the relevant information between spectra. In order to solve this problem, a spectrum shift densenet was proposed. The module adopted structure of dense convolutional module, and the spectrum shift module was used to realize the information interaction between the spectra. It replaced the down-sampling operation between spectra and extracted the global feature from the spectrum. Meanwhile, it avoided the loss of information in the down-sampling process and further improved the quality of the spectrum feature maps. The proposed method was verified on two widely used dataset ESC10 and ESC50 respectively. The classification accuracy of ESC10 and ESC50 datasets is 96.00% and 88.75% respectively. Compared with the existing networks, the accuracy is improved by 2.1% and 2.25%. Compared with convolutional neural networks based other methods, the proposed module can effectively mine more time-frequency information and has higher accuracy.

Keywords: environmental sound recognition;convolutional neural network;spectrum shift;DenseNet;deep learning

0 引 言

环境声音识别是人机交互的一个重要研究领域,也是人工智能交互时代极具应用前景的技术之一,并在人机交互、音频监控[1]和智能房间监控[2]等场景得到了初步应用。例如智慧家居系统能全方位监测独居老人和孩子的生活状态,可以识别老人、婴儿摔倒等异常声音,发出警报并紧急联系其家人。

传统的环境声音识别方法使用手工特征,比如音频能量、过零率和梅尔谱倒频谱系数(MFCC)特征[3]等,并使用高斯混合模型[4-5]、隐马尔可夫模型和支持向量机(SVM)[6]作为分类器。但手工特征方法的泛化能力比较弱,性能依赖数据集。近年来,以数据为驱动的深度学习方法得到了快速的发展,在图像融合[7]、目标检测[8]方面得到广泛的应用。深度学习通过非线性映射,实现了复杂函数逼近,并展现了从少数样本集中提取数据集本质特征的强大能力。当

收稿日期:2021-12-01
*基金项目:国家自然科学基金(62101512)、山西省青年科学基金(20210302124031)项目资助

前,深度学习在机器视觉和自然语言处理等方面展现了优异的性能,在具有挑战性的环境声音识别任务中也取得了显著的成果。利用深度学习的环境声音识别方法已经超越了传统手工特征的方法,成为主流研究方向。根据神经网络的类型可以将声音识别的方法分为如下两个方面:基于一维卷积(1D CNN)神经网络的环境声音识别和基于二维卷积神经网络(2D CNN)的环境声音识别。

基于一维卷积神经网络声音识别的方法,不需要对音频进行预处理,直接将音频信号输入到一维卷积神经网络,提取深度特征用于声音识别。Tokozume 等[9]设计了一个EnvNet 网络,用2个一维卷积层、3个池化层和2个全连接层提取特征,通过卷积核的长度提取短时间信息。为了获取不同长度的时间信息,Zhu 等[10]利用不同的一维卷积核提取多尺度时间信息用于声音识别。Abdoli 等[11]使用Gammatone filterbank 初始化一维卷积核来提高声音识别准确率。但一般音频序列比较长,采样点数比较多,需使用滑窗的方式选取音频序列作为一维卷积的输入,复杂度比较高,而且容易受到噪声影响。

为了提高对噪声的鲁棒性,越来越多的学者对音频序列进行预处理,把音频信号编码成时频谱图,如对数梅尔谱图(Log-Mel)、梅尔谱倒频谱系数图等,并利用二维卷积神经网络提取深度的时频信息。Piczak[12]构建了一个两层的二维卷积网络用于提取梅尔谱图的特征。但使用的卷积神经网络比较浅,无法提取有效的深度特征,为了解决这个问题,不少学者使用更深的卷积神经网络[13-15]去挖掘时频信息。同时有些学者为了提取局部显著性特征,引入注意力机制[16]来提高声音识别率。

但上述这些方法仅仅用卷积提取固定相邻频谱间的局部信息,不能有效地提取更多频谱之间的相关特征。为了更好利用卷积神经网络从频谱图挖掘有效的时频信息,本文提出了频谱位移密集神经网络探索频谱之间的相关特征,解决了卷积只能提取固定相邻频谱信息的问题,不需要频谱间下采样也能提取不同频谱之间的全局化信息,从而挖掘出更有效的时频特征,提高环境声音识别的准确率。

1 算法基本框架

该算法的整体框架如图1所示,包括音频输入、特征提取(对数梅尔谱图)、频谱位移密集神经网络和声音识别4个部分。文中重点论述特征提取和频谱位移密集神经网络这2个部分内容。

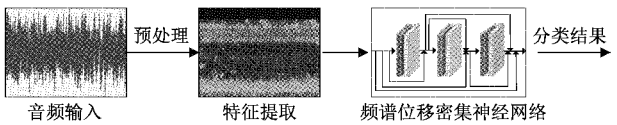

图1 整体算法框架

1.1 特征提取

由于输入的音频数据存在大量噪音,使用原始的一维音频数据作为特征很难凸显出有用的相关信息。为了解决此问题,目前比较流行的方式是将音频数据映射到梅尔谱或MFCC谱,即将输入声音信息通过梅尔谱特征提取的方式进行滤波,提高环境声音识别精度。本文采用对数梅尔图谱作为识别网络的输入,其原理是模仿人耳听觉特性设计滤波器,将原始音频数据经滤波器处理后再进行识别。其具体生成步骤如图2所示。

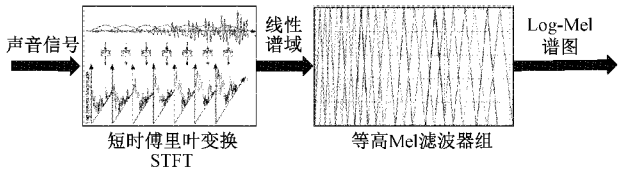

图2 Log-Mel 谱提取

首先对输入音频信号进行短时傅里叶变换(STFT),其具体步骤如下。

1)分帧:将输入的音频信号 $x(n)$ 分成多个短时帧,每帧长度为43 ms,帧移为21 ms。

2)加窗:将每一个短时帧乘以汉明窗,增加帧左右两端的连续性。

3)FFT:对各帧信号进行一个1 024点的快速傅里叶变换,将信号转换为频域上的能量分布。

4)取功率谱:将声音信号的频谱取模的平方,获得信号的谱线能量。

其次将所得功率谱通过等高Mel滤波器组:Mel滤波器组由 M 个三角滤波器组成(本文 $M=40$),滤波器在低频区域分布集中,高频区域分布稀疏。三角滤波器的频率响应为:

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)<k\leqslant f(m+1)\\ 0, & k>f(m+1)\end{cases} \tag{1}$$

其中,$0\leqslant m\leqslant M$,$f(m)$ 是中心频率。对于经过式(1)处理后得到的功率谱,分别与 M 个三角滤波器进行频率相乘累加,即可得到该帧数据对应频段的 M 个能量值。

最后将得到的梅尔谱图的幅度值(颜色)取对数转化为分贝,得到近似于同态变换的结果,如图3所示。

1.2 频谱频移密集神经网络

密集神经网络(DenseNet)已经被广泛地应用到计算机视觉的各个任务中,与传统的卷积神经网络相比,密集神经网络通过密集连接的方式实现各个卷积层的信息流通,可以有效地缓解深度网络中梯度消失的问题,同时密集连接的方式可以有效地利用不同层的信息进一步挖掘新的有用特征。本文采用密集连接的方式构建频谱位移密集神经

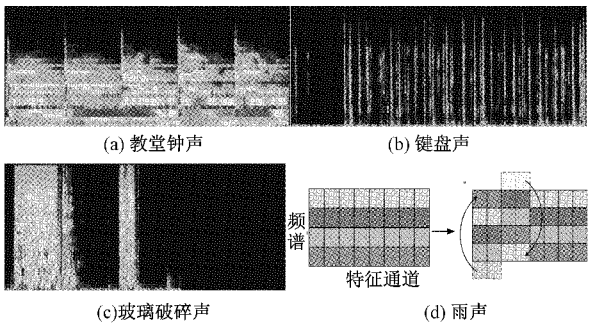

图 3　对数梅尔谱示意图

网络，如图 4 所示，该网络有 3 个卷积层、4 个频谱位移密集模块和 1 个全局池化层(global pooling)组成，每个卷积层包含 1 个二维卷积操作，1 个批量正则化层(BN 层)和激活函数(ReLU)。本文所提出的网络框架采用端到端的学习方式，实现从对数梅尔谱图到对应环境声音识别的直接映射。首先将对数梅尔谱图送入到一个卷积层提取浅层特征图，卷积核的时间维度步长设为 2，并将获得的特征图输入到 4 个频谱位移密集模块，然后把输出的特征图输入到两个卷积层，最后通过全局池化层输出声音分类的概率值。

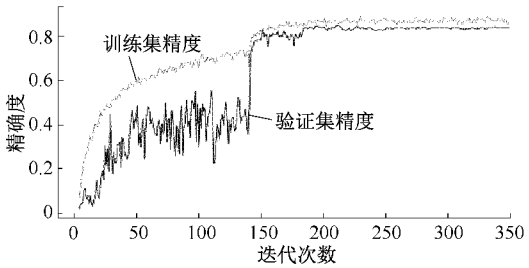

图 4　频谱位移密集神经网络

对于传统的卷积神经网络，卷积层通过大量的卷积核，在局部感受野下提取有效的时频特征，然后通过非线性激活函数和下采样实现全局特征的挖掘，但下采样会造成有用信息的丢失。为了减少信息的丢失，频谱图中频谱宽度会比较小，在下采样过程中，只在时间轴上进行下采样，保留频率轴上的全部信息。但利用卷积核的感受野只能提取相邻频谱的局部信息，为了解决此问题，本文提出了频谱位移密集模块，如图 5 所示。浅层特征图输入到两路，上面主路是 2 个卷积层，提取相邻频谱之间的局部特征，且第 1 层卷积的时间步长设为 2，进行时间上的下采样；下面的支路是 2 个卷积层和 1 个频谱位移模块，先经过第 1 个卷积层(时间步长设为 2)，然后通过频谱位移模块，实现频谱之间的位置变化，再通过卷积提取其他相邻频谱之间的信息，最后上下两路的信息通过密集连接的方式，实现不同频谱信息融合。

为了挖掘不同频谱之间的有用信息，本文设计了一个频谱位移模块，如图 6 所示，简单起见，只显示了频谱和特征通道，不同的频谱特征在每一行用不同的颜色表示，传统

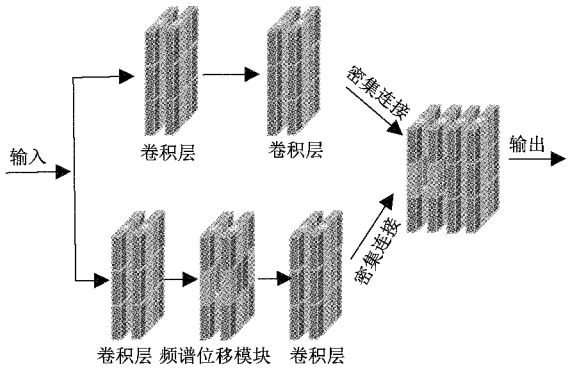

图 5　频谱位移密集模块

的二维卷积操作在不同的信道间进行操作，即沿着每一行单独进行操作，这样只能提取相邻频谱之间的信息。为了能够挖掘更多频谱之间的信息，本文将 1/4 的特征通道向下移动一位，最下面的特征通道移动到最上面；同时将 1/4 的特征通道向上移动一位，最上面的特征移动到最下面；剩下的特征通道位置保持不动。这样对每一行卷积操作时，能够提取不同频谱间的信息，生成更有效的时频特征。

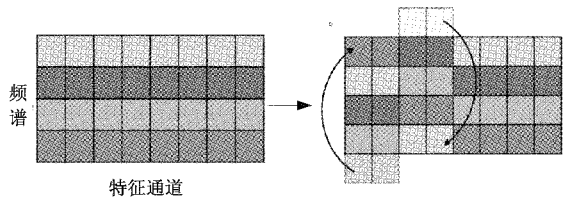

图 6　频谱位移模块

与 DenseNet 类似，在频谱位移密集模块中，下支路采用固定数量的卷积核，即第 1 层的卷积核设为 3×3×96，第 2 层的卷积核设为 3×3×24。对于上面主路，不同的频谱位移密集模块采用不同的卷积核数量，4 个频谱位移密集模块的卷积核依次为：3×3×24、3×3×48、3×3×96 和 3×3×192。整体的网络参数如表 1 所示。

表 1　整体网络参数

操作	卷积核	输出尺寸
卷积层 1	3×3×24	40×256×24
频谱位移密模块 1	3×3×24	40×128×48
频谱位移密模块 2	3×3×48	40×64×72
频谱位移密模块 3	3×3×96	40×32×120
频谱位移密集块 4	3×3×192	40×16×216
卷积层 2	3×3×384	40×16×384
卷积层 3	3×3×类别数	40×16×类别数
全局池化层	—	类别数

2　实验结果及数据对比

为了验证本文提出网络的有效性，在公认的 2 个数据集 ESC10 和 ESC50[17] 上进行验证。首先将两种音频数据

集中的训练集音频文件输入到本文所提出的网络中进行网络模型训练，然后将验证集音频文件输入到训练后的网络模型中，得到的分类结果并与真实结果相对比，获得相应的分类准确度。同时为了更好地体现出本文网络的优势，在对比本文网络与现有网络的准确度的同时，提出消融实验，进一步验证本文提出的方法对于声音识别准确度的提升，具体细节描述如下。

2.1 数据集

ESC10数据集总共10个声音类别分别为打喷嚏、狗吠、时钟滴答声、婴儿哭啼声、公鸡啼叫、雨声、海浪声、火声、直升机声和电锯声。每个声音平均约秒左右，数据采样频率44.1 kHz。该数据集总共400个样本，每种音类别40个样本。按照文献[17]的验证方法，采用5折交叉验证法进行样本的划分，即把数据划分成5份，4份数据作为训练数据，剩下的1份数据作为测试数据。

数据集ESC50比ESC10更难，更具有挑战性，该数据集有50种声音类别，主要分为如下几类：动物的声音、自然界的声音、人类的声音、室内声音和城市声音。总共2 000个样本，每种音类别40个样本，每个样本约5 s，采样频率44.1 kHz。该ESC50数据集同样采用5折交叉验证法。

2.2 实验细节

本次实验以keras作为环境，tensorflow作为后端，在Quardro RTX 6000 GPU上完成数据集的训练验证，Batch_size在ESC50和ESC10数据集上分别设为128和64，SGD作为优化函数，初始学习率设为0.1，采用数据样本混合的方式进行数据增强。在训练过程中，根据验证集的结果保存最优模型，且patience设为60，即迭代60轮，验证集的结果没有继续增加，则降低学习率，然后继续训练，直到学习率降到0.000 1时，停止训练。

2.3 数据集实验结果

ESC50训练过程中，训练集和验证集的精确度如图7所示，在150次迭代附近系统降低了学习率，导致在之后精确度有了一个陡峭的跳变，同时在之后的迭代过程中精确度的震荡有所减小。从图7可以看出，本网络模型有效地避免了过拟合问题。根据5折交叉验证法，数据集ESC10和ESC50上验证结果如表2所示，从表2可以看出，本文算法在各折的准确率，最后5折均值作为数据集的识别率。

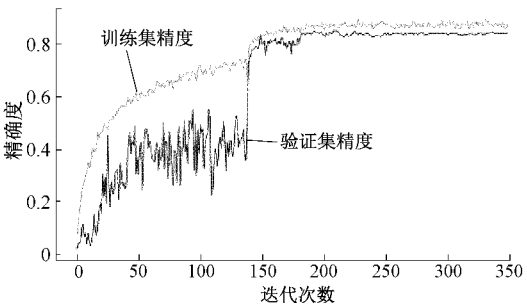

图7 ESC50数据集下精确度曲线

表2 5折交叉验证结果

交叉验证	ESC10	ESC50
1折	96.25	90.00
2折	97.50	87.50
3折	95.00	88.75
4折	97.50	90.25
5折	93.75	87.25
平均	96.00	88.75

同时为了验证频谱位移模块的有效性，本文把每个频谱位移密集模块中的频谱位移模块去除，其他的网络结构和参数保持不变，构建一个基础网络，与本文提出的网络进行了对比，结果如表3所示，通过实验验证，本文的提出网络结果高于基础网络，从而说明频谱位移模块能够更好地挖掘时频信息。

表3 频谱位移模块有效性验证

网络	ESC10	ESC50
基础网络	93.90	86.50
本文提出的网络	96.00	88.75

2.4 与现有方法对比

在数据集ESC10上，将本文算法与现有的算法进行对比，如表4所示，其中比较有代表性的方法是，Tokozume等[18]提出的一维卷积网络直接从音频信息提取深度特征；Li等[16]利用多路卷积网络(1DCNN+2DCNN)，提取不同的特征，然后通过特征融合提高声音的识别精度；Guzhov等[19]利用注意力机制来提高识别精度。本文网络的环境声音识别精度优于其他网络模型。同时从图8的混合矩阵可以看出，对于每个声音类别，本文提出的频谱位移密集神经网络都能很好地识别。

表4 数据集ESC10的算法对比

方法	结果
Piczak等[12]	95.70
Tokozume等[18]	91.30
Piczak等[20]	80.50
Boddapati等[13]	91.00
Salamon等[21]	91.70
Li等[16]	94.20
Guzhov等[19]	94.25
本文	96.00

由表5可知，数据集ESC50与ESC10相比，更具有挑战性，环境声音类别数更多，很多声音难以区分识别。在该数据集下，本文算法仍优于其它算法。由图9可知，对于ESC50的大部分环境声音，本文的方法都能很好地区分

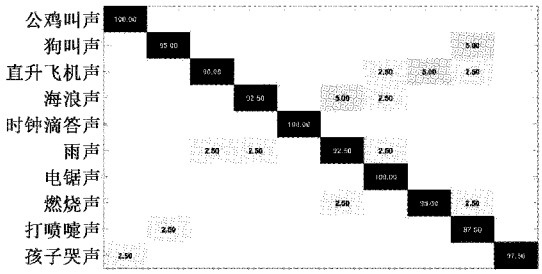

图 8　ESC10 数据集上的测试混淆矩阵

开，但对于一些声音表现的一般，比如风声和直升机声音。主要因为这类声音的音频低频占比过高，与日常环境噪声类似，当其他声音数据信噪比较低时，生成的频谱图与它们相似度很高，导致网络难以区分。

表 5　数据集 ESC50 的算法对比

方法	结果
Piczak 等[12]	81.30
Tokozume 等[18]	84.70
Zhu 等[10]	75.10
Sailor 等[22]	86.50
Piczak 等[20]	64.50
Boddapati 等[13]	73.00
Tak 等[23]	84.15
Salamon 等[21]	83.90
Li 等[16]	84.00
Guzhov 等[19]	83.15
Park 等[24]	88.10
本文	88.75

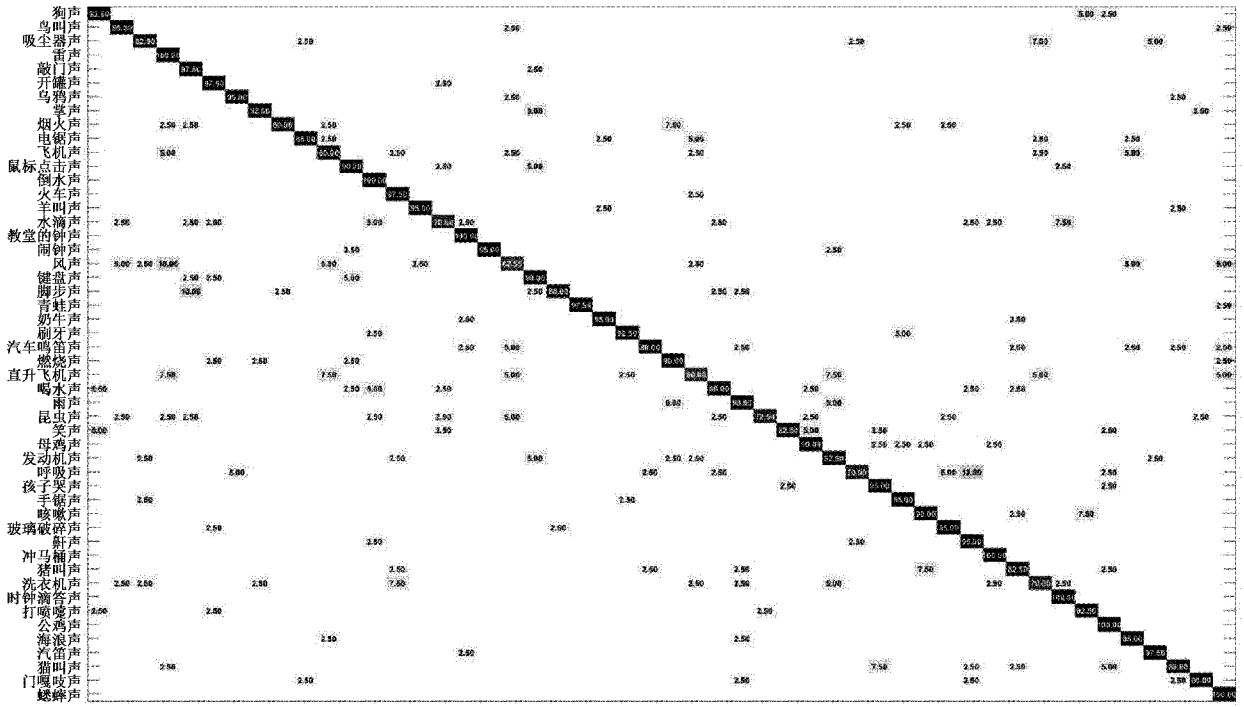

图 9　ESC50 数据集上的测试混淆矩阵

同时由于本网络模型大小仅有 21.9 MB，可以将此网络部署至一些便携式设备端进行应用，在树莓派端部署网络，嵌入家庭保健监测系统中，对家庭中的声音事件进行识别，利用树莓派的外设麦克风收集周围的声音信号，输入至 Linux 系统中，再将此声音信号的 wav 文件读取进入本文的网络模型中，对家庭中孩子哭泣声、警报声、电话铃声、笑声、尖叫声、咳嗽声等声音进行实时识别，经过实验验证，本网络模型可以高效准确地识别家庭中的声音，有效地提升家庭服务的智能化，实现全方面的家庭智能监测。

3　结　　论

本文针对卷积网络感受野的局限性，无法从频谱图中提取有效特征的问题，设计了一种端到端的频谱位移密集神经网络，将频谱位移嵌入密集神经网络中，提取不同频谱之间的相关信息，改善因下采样信息大量丢失的情况，提高了频谱特征图的质量，从而提高了环境声音识别精度。并在 ESC10 和 ESC50 两个公开常用的数据集上验证了该本文提出的方法，并取得了优异的结果。但仍有许多方面需要进一步研究改进：进一步优化频谱图，减少低频噪声对环境声音的影响，进一步提高声音识别精度等。

参考文献

[1]　RADHAKRISHNAN R, DIVAKARAN A, SMARAGDIS A. Audio analysis for surveillance applications[C]. IEEE Workshop on Applications of

Signal Process-ing to Audio and Acoustics, New Paltz,USA, 2005: 158-161.

[2] VACHER M, SERIGNAT J F, CHAILLOL S. Sound clssification in a smart room environment: An approach using GMM and HMM methods [J]. Computer Science, 2007:1-11.

[3] LI D, SETHI I K, DIMITROVA N, et al. Classification of general audio data for content-based retrieval [J]. Pattern Recognit. Lett., 2001(22): 533-544.

[4] VUEGEN L, BROECK B, KARSMAKERS P, et al. An MFCC-GMM approach for event detection and classific-ation [C]. Signal Process. Audio Acoust. New Paltz, UAS, 2013:1-3.

[5] MESAROS A, HEITTOLA T, ERONEN A, et al. Event detection in real life recordings[C]. Eur. Signal Process. Conf. Budapest, Hungary, 2010: 1267-1271.

[6] UZKENT B, BARKANA B D, CEVIKALP H. Non-speech environmental sound classification using SVMs with a new set of features [J]. Int. J. Innovative Comput., Inform. Control, 2012(8): 3511-3524.

[7] 刘小利.基于深度学习算法的图像融合[J].国外电子测量技术,2020,39(7):38-42.

[8] 包本刚.融合多特征的目标检测与跟踪方法[J].电子测量与仪器学报,2019,225(9):93-99.

[9] TOKOZUME Y, HARADA T. Learning environmental sounds with end-to-end convolutional neural net-work[C]. Speech and Signal Processing (ICASSP). New Orleans, Louisiana, 2017: 2721-2725.

[10] ZHU B, XU K, WANG D, et al. Environmental sound classification based on multitemporal resolution convolutional neural network combining with multilevel features[C]. Hefei, China, 2018: 528-537.

[11] ABDOLI S, CARDINAL P, KOERICH A L. End-to-end environmental sound classification using a 1d convolutional neural network[J]. Expert Systems with Applications, 2019(136): 252-263.

[12] PICZAK K J. Environmental sound classification with convolutional neural networks [C]. Boston, Massachusetts, USA, 2015: 1-6.

[13] BODDAPATI V, PETEF A, RASMUSSON J, et al. Classifying environmental sounds using image recognition networks[J]. Procedia Computer Science, 2017(112): 2048-2056.

[14] ZHANG Z, XU S, ZHANG S, et al. Learning attentive representations for environmental sound classification[J]. IEEE Access, 2019(130): 327-339.

[15] 孔子迁,邓蕾,汤宝平,等.基于时频融合和注意力机制的深度学习行星齿轮箱故障诊断方法[J].仪器仪表学报,2019,40(6):221-227.

[16] LI X, CHEBIYYAM V, KIRCHHOFF K. Multistream network with temporal attention for environmental sound classification [C]. Interspeech, Graz, Austria, 2019:3604-3608.

[17] PICZAK K J. ESC: Dataset for environmental sound classification[C]. Brisbane Australia, 2015: 1015-1018.

[18] TOKOZUME Y, USHIKU Y, HARADA T. Learning from between class examples for deep sound recognition[C]. ArXiv Preprint, 2017,ArXiv:1711.10282.

[19] GUZHOV A, FEDERICO RAUE, JÖRN HEES, et al. ESR-esNet: Environmental sound classification based on visual domain models [C]. Milan, Italy, 2021: 4933-4940.

[20] PICZAK K J. ESC: Dataset for environmental sound classification[C]. Br-isbane Australia, 2015: 1015-1018.

[21] SALAMON J, BELLO J P. Deep convolutional neural networks and data augmentation for environmental sound classification [J]. IEEE Signal Process-ing Letters, 2017(24): 279-283.

[22] SAILOR B H, AGRAWAL D M, PATIL H A. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification [C]. Stockholm, Sweden, 2017: 3107-3111.

[23] TAK R N, AGRAWAL D M, PATIL H A. Novel phase encoded mel filterbank energies for environmental sound classification [C]. Kolkata, India, 2017: 317-325.

[24] PARK H, YOO C D. CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification [J]. IEEE Signal Processing Letters, 2020:411-415.

作者简介

李传坤,工学博士,讲师,主要研究方向为深度学习、模式识别。

E-mail:chuankun@nuc.edu.cn

郭锦铭(通信作者),硕士研究生,主要研究方向为声音识别、深度学习。

E-mail:guojinming0927@163.com