基于双输入卷积神经网络的环境声事件识别

李芳足[1] 罗丽燕[1] 王 玫[1,2*]

[1](桂林电子科技大学认知无线电与信息处理教育部重点实验室 广西 桂林 541004)

[2](桂林理工大学信息科学与工程学院 广西 桂林 541007)

摘 要 针对前融合的特征融合方式不利于卷积神经网络提取高阶特征的问题,提出一种基于双输入卷积神经网络的特征融合框架。该特征融合框架将两种声学特征分别经过不同的卷积和池化策略进行高阶特征提取,将高阶特征进行拼接并送入输出层输出分类结果。这种方式不仅为不同的特征匹配不同的卷积和池化策略,还避免了单位或尺度不同的特征拼接在一起干扰卷积核的特征提取。经公开数据集的评估结果显示,该多特征融合框架相比单一特征和现有的融合方式性能更优。此外,将此框架应用于实际场景下的汽车鸣笛声的识别,结果显示,查全率达到 87.7%,查准率达到 84.7%,F1 度量达到 86.2%,优于其他方法,验证了该方法在实际应用中的可行性。

关键词 环境声事件识别 特征融合 卷积神经网络

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.07.025

ENVIRONMENTAL SOUND EVENT RECOGNITION BASED ON DOUBLE-INPUT CONVOLUTIONAL NEURAL NETWORK

Li Fangzu[1] Luo Liyan[1] Wang Mei[1,2*]

[1](*Provincial Ministry of Education Key Laboratory of Cognitive Radio and Signal Processing*, *Guilin University of Electronic Technology*, *Guilin 541004*, *Guangxi*, *China*)

[2](*School of Information Science and Engineering*, *Guilin University of Technology*, *Guilin 541007*, *Guangxi*, *China*)

Abstract Aiming at the problem that the feature fusion method of pre-fusion is not conducive to the extraction of high-order features for convolutional neural network, we propose a feature fusion framework based on double-input convolutional neural network. The proposed framework extracted two acoustic features through different convolution and pooling strategies for high-order feature extraction. It concatenated the high-order features and sent them to the output layer in order to output the classification results. This method chose suitable convolution and pooling strategies for different features, and avoided that fusion-feature affected the feature extraction of convolutional kernel when features units or scales were different. The evaluation results of the public data set show that this multi-feature fusion framework performs better than the single feature and existing feature fusion method. The framework was applied to the recognition of car horn sounds in actual scenarios. The results show that recall rate reaches 87.7%, precision rate reaches 84.7%, and F1 measure reaches 86.2%, which are better than other methods. The feasibility of proposed method was verified in actual scenarios.

Keywords Environmental sound event recognition Features fusion Convolutional neural network

---

收稿日期:2020-03-09。国家自然科学基金项目(61771151);广西重点研发计划项目(2017AB08072);广西自然科学基金项目(2016GXNSFBA38014);中国博士后科学基金项目(2016M602921XB);广西研究生科研创新项目(YCSW2019139);桂林电子科技大学创新项目(2019YCXS038)。**李芳足**,硕士生,主研领域:环境声信号处理,深度学习。**罗丽燕**,讲师。**王玫**,教授。

0 引 言

视频监控在公共安全管理中发挥着重要作用,为保护人民生命财产安全提供了有力支撑。但由于室外环境下视频数据的采集过程易受环境因素的干扰,且视频采集设备通常布点固定,所以会出现“监控盲区”的问题。单纯地以增加视频采集设备为代价解决“监控盲区”问题,无疑会较大地增加设备成本与存储成本。因此,如何在低成本的条件下实现监控无死角覆盖成为了急需解决的问题。而声传播的全向性、声接收设备成本较低等优点使得基于声的监控手段得到了广泛关注,例如针对道路交通环境下的异常声事件监测[1]、针对动物声识别的动物习性和生活区域监测[2]、针对地铁环境的异常声事件监测[3]等。

环境声事件识别是指对采集的环境声数据进行分析进而识别出其中包含的声学事件的技术。经过近年来对该技术的研究,研究人员借鉴语音识别框架总结出一套环境声事件识别框架。该框架包含两个重要部分:声学特征提取和分类器识别[4]。早期的环境声事件识别的研究中,由于识别任务较为简单加之计算机的算力不足,常使用 K 近邻算法(K-Nearest Neighbor, KNN)[5]、支持向量机(Support Vector Machines, SVMs)[6-7]和随机森林算法(Random Forest, RF)[8]等作为分类器,梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCCs)作为声学特征。但是随着将环境声事件识别投入实际场景应用的需求增加,环境声事件识别技术所面临的应用场景更加复杂多变,上述分类器由于对复杂数据的建模能力有限,无法满足当前的环境声事件识别的要求。

近些年,随着计算机的算力提升,深度学习受到环境声事件识别领域研究人员的广泛关注,目前主流的环境声事件识别技术常使用卷积神经网络(Convolutional Neural Networks, CNN)[9-12]作为分类器,对数梅尔谱(Log-mel spectrogram, Log-mel)作为声学特征,卷积神经网络因具有强大的特征提取能力和复杂函数建模能力而使得环境声事件识别性能得到有效的提升。然而在机器学习领域中,数据、特征和分类算法是决定机器学习性能的关键因素,文献[9-12]尽管采用了不同的卷积策略和不同的激活函数提升了分类算法的性能,但其只采用 Log-mel 特征作为卷积神经网络的输入,使得环境声事件识别性能受限。针对这个问题,许多研究人员对多特征融合进行了调研,并指出融合特征的表现要优于单一特征[13],例如文献[2]将投影特征和局部二元模式变化特征进行融合从而完成了低信噪比环境下动物声的自动识别任务。文献[14]融合梅尔频率倒谱系数(MFCC)和 Gammatone 倒谱系数(GFCC)解决了有噪声环境下的说话人识别问题。然而上述文献的特征融合方式均采用前融合方式(early fusion-based method),尽管此类融合方式已经取得一定成效,但是并不适合于卷积神经网络,因为这种融合方式存在如下缺陷:单位或尺度不同的两种特征拼接在一起会使得融合特征存在内部数值差异较大以及产生无规律的拼接边界,从而影响卷积神经网络的特征提取能力。文献[13,15]使用不同的声学特征对不同的模型进行训练,然后将训练好的模型使用 DS 证据理论(Dempster-Shafer evidence theory)进行融合,经 Urbansound8K、ESC-10 和 ESC-50 数据集评估结果表明基于 DS 证据理论的后融合方式(late fusion-based method)具有较好的识别表现。这种基于 DS 证据理论的后融合方式尽管避免了前融合方式带来的弊端,但是需要对两个模型分开训练使得识别方法更繁琐并且无法保证特征进行有效的融合。因此,寻找一种适合卷积神经网络的特征融合方式成为必要。

为解决上述问题,本文作出如下贡献:(1) 提出一种基于双输入卷积神经网络的特征融合框架,该框架的核心是为 MFCCs 特征和 Log-mel 特征匹配合适的卷积和池化策略。(2) 通过实景实验,探索了该融合框架在实际场景中应用的可行性。

1 MFCCs 和 Log-mel 特征提取

声学特征是影响环境声事件识别性能的重要因素,不同类型的声学特征可以从不同角度描述声音信号,该融合框架选择 MFCCs 特征和 Log-mel 特征作为融合对象,两种特征提取流程如图 1 所示。Log-mel 特征是经过梅尔滤波器过滤后的频谱特征,符合人耳的听觉特性,描述了声音信号频谱的全局信息,被广泛应用于环境声事件识别和声场景识别中;MFCCs 特征是 Log-mel 特征经过离散余弦变换之后得到的倒谱特征,该特征反映了信号的倒谱特征,被广泛应用于语音识别和说话人识别中。图 2 是对汽车鸣笛声、枪声和尖叫声分别提取 Log-mel 特征和 MFCCs 特征得到的特征图,可以看出,Log-mel 特征图可以更直观地看到三种声音的区别,在图像上更具辨识度,而 MFCCs 特征由于只保留了低频部分的谱包络信息无法直观地分辨出三种声音。对这两种特征进行融合不仅可以从全局的频谱信息中对声音信号进行区分,还可以通过低频的包络信息对特征进行补充,有效地提高了特征的描述能力和抗噪能力。除此之外,Log-mel 特征是 MFCCs

特征的中间产物,同时提取这两种特征时不会增加额外的计算消耗,可以满足在实际应用中对特征提取的实时性要求,因此选择这两种声学特征来描述环境声信号。两种特征的提取步骤如下[16]。

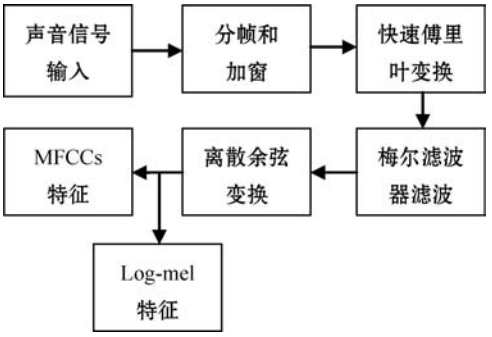

图 1　MFCCs 和 Log-mel 特征提取流程

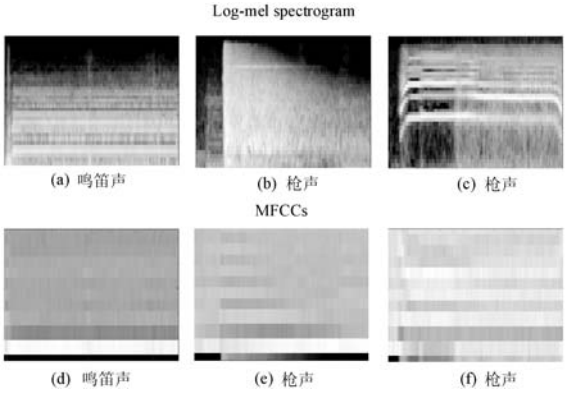

图 2　Log-mel 和 MFCCs 特征图

(1) 分帧和加窗:将一段声音信号分为一系列重叠的短帧 $s(n)$,帧长设为 1 024,帧移设为 512。然后对帧信号 $s(n)$ 加汉明窗 $\omega(n)$ 来减轻边界效应,汉明窗 $\omega(n)$ 为:

$$\omega(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right)\quad 0\leqslant n\leqslant N-1 \tag{1}$$

式中:N 为总的采样点数。

(2) 快速傅里叶变换:进行快速傅里叶变换(Fast Fourier Transform,FFT)得到其复数谱。假设输入信号为 $x(n)$,该信号的离散傅里叶变换(Discrete Fourier Transform,DFT)公式为:

$$X(k)=\sum_{n=0}^{N-1}x(n)\exp\left(\frac{-j2\pi kn}{N}\right)\quad 0\leqslant k\leqslant N-1 \tag{2}$$

式中:N 表示进行 DFT 变换的点数;$X(k)$ 表示第 k 个频率点的值。然后将得到的复数谱取模平方得到功率谱。

(3) 梅尔滤波器滤波:将功率谱通过一组梅尔滤波器,即:

$$s(m)=\ln\left(\sum_{k=0}^{N-1}|X(k)|^2Hm(k)\right)\quad 0\leqslant m\leqslant M \tag{3}$$

式中:$Hm(k)$ 为梅尔滤波器组;M 为滤波器组中三角滤波器的数量,取 $M=40$。梅尔滤波器组计算公式为:

$$H_m(k)=\begin{cases}0 & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m-1)\leqslant k\leqslant f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m)\leqslant k\leqslant f(m+1)\\ 0 & k\geqslant f(m+1)\end{cases} \tag{4}$$

式中:$f(m)$ 为第 m 个三角滤波器的中心频率,$1\leqslant m\leqslant M$。

然后将梅尔频谱取对数,得到对数梅尔谱特征。

(4) 离散余弦变换:对数梅尔谱做离散余弦变换得到 MFCCs 系数,即:

$$MFCCs(n)=\sum_{m=0}^{M-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)\quad n=1,2,\cdots,L \tag{5}$$

本文取前 12 个系数作为最终的 MFCCs 特征,即 $L=12$。

2　基于双输入卷积神经网络的特征融合框架

不同声学特征的描述能力不同,经过有效的融合可以极大地提高环境声事件识别的性能,本文采用基于双输入卷积神经网络的特征融合框架,通过双输入方式为 Log-mel 和 MFCCs 匹配不同的卷积和池化策略,然后通过展平和拼接操作对提取到的高阶特征进行融合。同时,使用 Batch Normalization、正则化、Dropout 等技巧提升了网络的训练速度以及泛化能力。

2.1　双输入卷积神经网络的网络结构及特征融合方式

本文借鉴经典的卷积神经网络[9,17]和 BP 神经网络,设计了如图 3 所示的双输入卷积神经网络。该网络有两条输入并分别使用 MFCCs 特征和 Log-mel 特征作为输入数据,其数据维度分别为 $\boldsymbol{X}_{\text{mfcc}}\in\mathbf{R}^{12\times 80}$、$\boldsymbol{X}_{\text{logmel}}\in\mathbf{R}^{40\times 80}$。详细的模型结构描述如下。

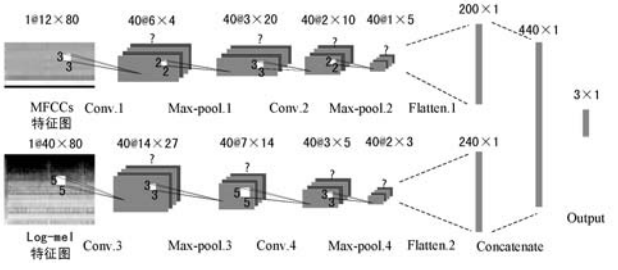

图 3　双输入卷积神经网络结构

在前向传播过程中,每次输入 $\boldsymbol{X}_{\text{mfcc}}$ 和 $\boldsymbol{X}_{\text{logmel}}$,数据从前一层网络流向下一层网络,直到输出层得到分类结果,并且前一层流向下一层网络的数据需经过非线性映射 $F(\cdot|\boldsymbol{\Theta})$,从输入层 $\boldsymbol{X}_{\text{mfcc}}$ 和 $\boldsymbol{X}_{\text{logmel}}$ 到 Max-pool2 和 Max-pool4 的操作分别为:

$$\boldsymbol{Z}_{\text{max-pool2}} = F(\boldsymbol{X}_{\text{mfcc}}|\boldsymbol{\Theta}) = f_l(\cdots f_2(f_1(\boldsymbol{X}_{\text{mfcc}}|\theta_1)|\theta_2)|\theta_l) \quad l=4 \tag{6}$$

$$\boldsymbol{Z}_{\text{max-pool4}} = F(\boldsymbol{X}_{\text{logmel}}|\boldsymbol{\Theta}) = f_l(\cdots f_2(f_1(\boldsymbol{X}_{\text{logmel}}|\theta_1)|\theta_2)|\theta_l) \quad l=4 \tag{7}$$

式中:$f_l(\cdot|\theta_l)$表示对第 l 层网络的操作,例如 $l \in \{Conv.1, Conv.2, Conv.3, Conv.4\}$ 为卷积层,其卷积运算为:

$$\boldsymbol{Z}_l = f_l(\boldsymbol{X}_l|\theta_l) = h(\boldsymbol{W} * \boldsymbol{X}_l + \boldsymbol{b}), \theta_l = [\boldsymbol{W}, \boldsymbol{b}] \tag{8}$$

式中:$\boldsymbol{X}_l$为输入的三维张量;$\boldsymbol{W}$ 为卷积核;$*$ 表示卷积操作;$\boldsymbol{b}$ 为偏置向量;$h(\cdot)$表示激活函数。然后在每层卷积层后接最大池化层 $l \in \{Max\text{-}pool.1, Max\text{-}pool.2, Max\text{-}pool.3, Max\text{-}pool.4\}$,用来减小特征映射的维度和提升训练速度。

在 Max-pool2 和 Max-pool4 后接 Flatten.1 和 Flatten.2 层将其输出的三维张量展开为一维张量 $\boldsymbol{Z}'_{\text{max-pool2}}$ 和 $\boldsymbol{Z}'_{\text{max-pool4}}$,然后在 Concatenate 层将两个一维张量串联在一起,在此时对两特征进行融合即:

$$\boldsymbol{Z}_{\text{concatenate}} = \text{concatanate}(\boldsymbol{Z}_{\text{max-pool2}}, \boldsymbol{Z}_{\text{max-pool4}}) \tag{9}$$

最后,将融合后的一维张量与输出层进行全连接,操作为:

$$\boldsymbol{Z}_l = f_l(\boldsymbol{X}_l|\theta_l) = h(\boldsymbol{W}\boldsymbol{X}_l + \boldsymbol{b}), \theta_l = [\boldsymbol{W}, \boldsymbol{b}] \tag{10}$$

式中:$\boldsymbol{X}_l$表示 Concatenate 层输出的一维张量;$\boldsymbol{W}$ 表示权重;$\boldsymbol{b}$ 为偏置参数;$h(\cdot)$表示激活函数。

基于双输入卷积神经网络的特征融合方式可归为后融合方式。而前融合方式是在卷积神经网络输入前对声学特征进行如图 4 所示的操作。这种融合方式会存在如下缺点:单位或尺度不同的两种特征拼接在一起会使得融合特征存在内部数值差异较大以及产生无规律的拼接边界的问题,从而干扰卷积核更新有效的权值,影响卷积神经网络的特征提取能力。针对这个缺点,基于双输入卷积神经网络的特征融合框架的优势在于为不同的特征匹配不同的卷积和池化策略,充分发挥卷积神经网络的特征提取能力,最后将得到高阶特征进行融合并输送到 Softmax 层,对提取到的高阶特征进行选择和非线性拟合,极大地提高了网络的分类性能。

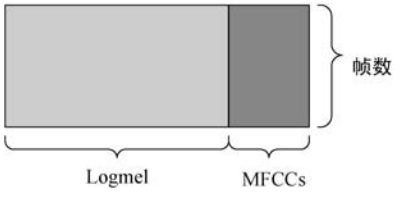

图 4 以前融合方式融合 Log-mel 和 MFCCs

2.2 网络参数分析

本文提出的卷积神经网络有两个特点:(1) 双输入结构,不同的输入经过不同的卷积层和池化层,充分发挥不同特征的描述能力,从而提高网络的分类性能;(2) 无额外的全连接层,这种结构可以有效地减少模型的参数和降低模型的复杂度,提高模型的泛化能力[18]。具体的网络参数设置如下。

(1) Conv.1 和 Conv.2:这两层卷积层均使用 40 个 3×3 的卷积核,卷积核的滑动步长为 2。这种小尺寸卷积核用于提取 MFCCs 特征图中的局部高阶特征并且有效地减少了模型的参数。然后将卷积核的输出用修正线性单元(Rectified Linear Unit, ReLU)[19]进行非线性映射,其映射关系为:

$$f(x) = \max(0, x) \tag{11}$$

同时,在每个卷积核和激活函数之间引入 Batch Normalization 技术[20],用来提高神经网络训练的速度和稳定性。

(2) Conv.3 和 Conv.4:这两层卷积层均使用 40 个 5×5 的卷积核用于提取 Log-mel 特征图的深层特征,卷积核滑动步长为 2,同样采用 ReLU 作为激活函数并且在激活函数前引入 Batch Normalization 技术。

(3) Max-pool.1 和 Max-pool.2:这两层池化层均采用 2×2 的最大池化滤波器来下采样上层输出,以达到减小输出数据的尺寸和特征选择的目的。

(4) Max-pool.3 和 Max-pool.4:这两层池化层均采用 3×3 的最大池化滤波器。

为了进一步提高模型的泛化能力,本模型在输出层前添加概率为 0.5 的 Dropout 机制,即在每批次的训练过程中,随机地让网络中的某些隐藏层节点的权重暂时失效,通过 Dropout 机制可以减轻网络节点之间的联合适应性,防止网络发生过拟合现象[21]。此外,网络还使用了 L^2 参数范数惩罚,使得权重更加接近原点,防止过拟合[21],即通过向目标函数添加一个正则项:

$$\Omega(\boldsymbol{\Theta}) = \frac{1}{2} \| w \|_2^2 \tag{12}$$

式中:向量 $\boldsymbol{w}$ 表示所有应受范数惩罚影响的权重;向量 $\boldsymbol{\theta}$ 表示所有参数(包括 $\boldsymbol{w}$ 和无须正则化的参数)。

针对多分类任务,本模型使用目标函数-分类交叉熵损失(Categorical Cross-entropy)来衡量当前训练得到的概率分布与真实分布之间的距离,交叉熵损失函数定义为:

$$C = -\sum y \log(a) \tag{13}$$

式中:y 表示期望输出;a 表示模型得到的输出,而 a =

$\sigma(z)$,其中 $\sigma(\cdot)$表示激活函数,$z = \sum WX + b$。输出层的激活函数使用 Softmax 函数,即每个神经元的输出映射为:

$$\sigma_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_j}} \tag{14}$$

而且要保证:

$$\sum_{i=1}^{J} \sigma_i(z) = 1 \tag{15}$$

式中:J 为输出层神经元个数,要求与预定义的类别数量保持一致。

在做反向传播时,采用 Adam[22] 优化器来训练网络,Adam 是一种学习率自适应的优化算法,它采用了偏置修正,修正从原点初始化的一阶矩(动量项)和(非中心的)二阶矩的估计,使得其对超参数的选择更鲁棒[20]。

3　基于环境声数据集的实验分析

3.1　环境声数据集

实验使用公开的环境声数据集 Google AudioSet[23],该数据集是目前声音种类最丰富、数量最多的声音数据集,常用于评估环境声事件识别方法。本文从该数据集中选取了三种比较典型的环境声:枪声、尖叫声和汽车鸣笛声,每种类别的声音样本数量均为 900 余条,每条声音样本均采用 44.1 kHz 采样和 16 bits位深度编码为 WAV 格式。然后按照 7∶3将声音样本随机划分为训练集和测试集。

3.2　实验设置

本实验使用公开的环境声数据集对如下十种环境声事件识别方法进行评估对比。

方法一:使用文献[11]中的识别方法作为 Baseline 方法,该方法使用对数梅尔谱作为声学特征,使用卷积神经网络作为分类算法。

方法二:采用 MFCCs 作为声学特征,单输入卷积神经网络作为分类器,卷积神经网络结构如图 5 所示。其中卷积层和池化层结构与本文设计的双输入卷积神经网络中关于 MFCCs 输入部分的卷积层和池化层结构保持一致,在 Flatten 层与输出层之间添加一层全连接层。

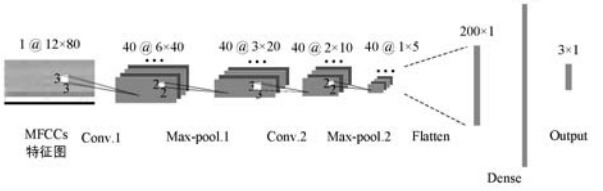

图 5　以 MFCCs 特征作为输入的卷积神经网络结构

方法三:采用 Log-mel 作为声学特征,单输入卷积神经网络作为分类器,其结构如图 6 所示。该网络中卷积层和池化层与本文设计的双输入卷积神经网络中有关 Log-mel 输入部分中的卷积层和池化层结构保持一致,同样在 Flatten 与输出层之间添加一层全连接层。

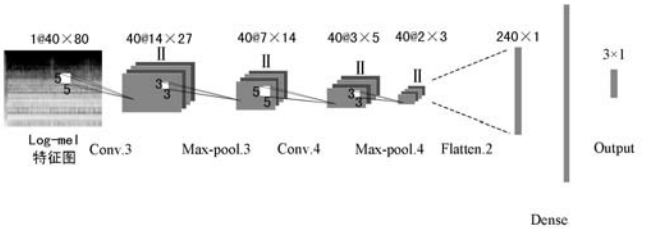

图 6　以 Log-mel 特征作为输入的卷积神经网络结构

方法四:声学特征采用前融合方式融合 MFCCs 特征与 Log-mel 特征,分类器采用 K 近邻算法。

方法五:声学特征采用前融合方式融合 MFCCs 特征与 Log-mel 特征,分类器采用支持向量机算法。

方法六:声学特征采用前融合方式融合 MFCCs 特征与 Log-mel 特征,分类器采用随机森林算法。

方法七:声学特征采用前融合方式融合 MFCCs 特征与 Log-mel 特征,分类器采用包含两个隐含层的多层感知机。

方法八:声学特征采用前融合方式融合 MFCCs 特征与 Log-mel 特征,分类器采用图 6 所示的卷积神经网络。

方法九:使用文献[13,15]中采用的 DS 证据理论对方法一和方法二中训练好的模型进行融合,以此作为基于后融合的对比方法。

方法十:即本文方法,采用 MFCCs 和 Log-mel 作为声学特征,双输入卷积神经网络作为分类器。

所有的实验均在 Windows 平台下完成,硬件设备使用酷睿 i7 6800K 处理器和 GTX1080TI 显卡,软件部分中涉及到的特征提取和分类算法的建模和应用借助 Python 语言中的 librosa、sklearn 和 TensorFlow 等模块完成。

3.3　评估指标

评估环境声事件识别方法常采用如下的评估指标[24]:

(1) 查全率(Recall):正确识别到的鸣笛声数量占鸣笛声真实发生数量的比率。

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

(2) 查准率(Precision):正确识别到的鸣笛声数量占识别到鸣笛声数量的比率。

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{17}$$

（3）F1-度量（F1-measure）：

$$F1 = \frac{2 \times Percision \times Recall}{Precision + Recall} \times 100\% \qquad (18)$$

式中：*TP* 称为真正例（True Positive）；*FP* 称为假正例（False Positive）；*TN* 称为真反例（True Negative）；*FN* 称为假反例（False Negative）。在评估指标中，查全率和查准率越高说明检测系统性能越好，但是这两者是相互矛盾的，因此引入 F1-度量来权衡两者。

3.4 实验结果分析

将实验结果以混淆矩阵图的形式呈现在图 7 中，其中图 7（a）－图 7（j）是使用十种方法得到的评估结果。并将实验结果以查全率、查准率、F1 度量的形式呈现在表 1 中。

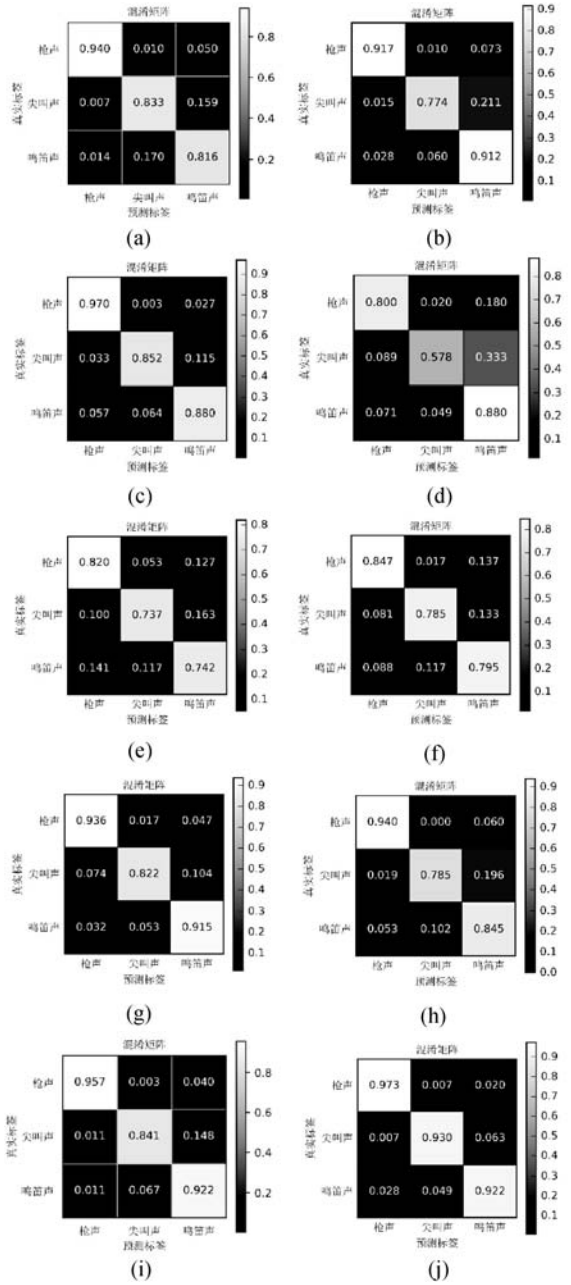

图 7 十种识别方法得到的混淆概率矩阵

表 1 十种方法的评估结果（%）

识别方法	评估指标		
	Recall	Precision	F1-measure
Baseline	86.3	86.5	86.4
方法二	86.7	87.9	87.3
方法三	90.0	90.3	90.1
方法四	75.3	78.8	77.0
方法五	76.6	76.9	76.7
方法六	80.9	81.2	81.0
方法七	89.0	89.5	89.2
方法八	85.7	88.6	87.1
方法九	90.7	91.1	90.9
本文方法	94.2	95.4	94.8

对比方法二和方法三的实验结果可以得出，使用 MFCCs 特征的方法仅对汽车鸣笛声的识别表现优于 Log-mel 特征，而从整体识别表现看，其识别表现不如使用 Log-mel 特征的方法，因此可以得出，Log-mel 特征和 MFCCs 特征对不同声音信号的描述能力不同，而且使用 Log-mel 特征的方法要优于使用 MFCCs 特征的方法，通过将两种特征进行融合可以对特征的描述能力进行互补从而提高识别方法的性能。方法二和方法三的实验结果要优于 Baseline 方法，验证了本文所设计的卷积神经网络分类性能突出。

通过比较方法四-方法八的实验结果，可以对使用前融合方式的不同分类算法进行比较。分析实验结果，使用传统分类算法的方法相比使用深度学习的方法存在一定差距。因此证明了深度学习技术更适合处理环境声信号。

通过对比 Baseline、方法二、方法三、方法八、方法九、方法十（本文方法）的实验结果，可以对单特征方法、基于前融合方式的融合特征方法和基于 DS 证据理论的后融合方法与本文提出的基于双输入卷积神经网络的方法进行对比。分析实验结果，方法二和方法三的识别结果优于方法八，因此验证了基于前融合的特征融合方式对卷积神经网络的分类性能产生了负面影响。方法九的表现优于方法二和方法三，证明了基于 DS 证据理论的融合方式是一种有效的特征融合手段。而本文方法在各项指标的表现相较于其他的方法有明显提升，因此本文提出的特征融合框架是有效且性能突出的。

4 基于实际场景的汽车鸣笛声识别实验

为了评估本文方法在实际场景中应用的性能，通

过实景实验对上述性能较好的识别方法与本文方法进行对比。

4.1　环境声数据的采集

为了保证实验的真实性,在桂林电子科技大学金鸡岭校区正门前放置声音采集设备,对过往车辆的鸣笛声进行采集,采集场景及采集设备如图 8 所示。经过长时间的采集,最终得到 1 742 条鸣笛声数据,每条声音数据持续时间为 0.6 s ~ 1.5 s,均采用 44.1 kHz 的采样频率和 16 bits 的位深度保存为 WAV 格式。使用采集到的汽车鸣笛声数据用于训练分类算法,最终使用一段未参与训练的时长为 10 min 的街道环境声数据对该网络进行评价。

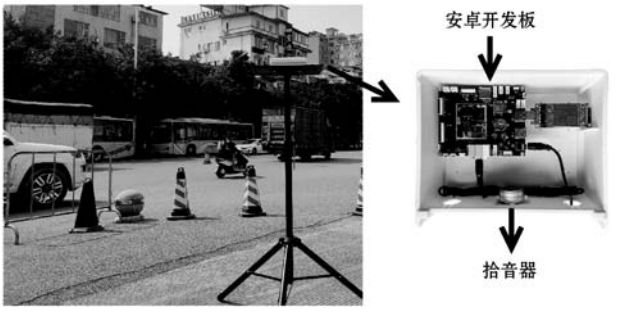

图 8　声音采集场景

4.2　评估方式

汽车鸣笛声识别类似于跌倒声识别[25]属于二分类任务,要求在一段声音信号中检测并识别出是否存在汽车鸣笛声,因此采用如图 9 所示的评估方法。图 9 中上方的黑线表示鸣笛声检测的真实结果,中间的虚线表示模型检测得到的结果,底部的粗黑线表示时间轴,凸起的线条表示有汽车鸣笛声发生。图 9 中展示了在模型的识别结果中会出现的四种情况:TP、FP、TN、FN,当模型识别结果和真实结果均为汽车鸣笛声时表示为 TP,反之表示为 TN。当模型识别结果为汽车鸣笛声而真实结果中无汽车鸣笛声时表示为 FP,反之为 FN。

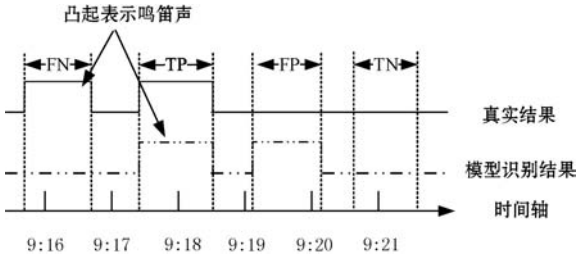

图 9　鸣笛声识别评估策略

4.3　实验结果分析

表 2 呈现了汽车鸣笛声识别的实景实验结果,基于双输入卷积神经网络的环境声事件识别方法对汽车鸣笛声的识别拥有较高的查全率(*Recall* = 87.7%),而且其查准率(*Precision* = 84.7%)相比查全率也仅仅低了 3 百分点,综合这两个指标得到的 F1-度量也能达到 86.2%,而且相比 Baseline、方法二、方法三、方法六、方法八和方法九表现也有较大提升。综合实验结果,基于双输入卷积神经网络的特征融合框架在实际环境声中仍具有较好识别性能,而且该识别方法明显优于单特征方法、基于前融合的融合特征方法和基于 DS 证据理论的模型后融合方法。

表 2　鸣笛声识别的评估结果(%)

识别方法	评估指标		
	Recall	Precision	F1-measure
Baseline	73.0	72.8	72.9
方法二	73.6	72.4	73.0
方法三	75.4	76.8	76.1
方法六	63.0	64.3	63.6
方法八	73.7	68.9	72.2
方法九	76.8	77.3	77.0
本文方法	87.7	84.7	86.2

5　结　语

本文针对前融合的特征融合方式不利于卷积神经网络提取高阶特征的问题,提出一种基于双输入卷积神经网络的特征融合框架。经公开数据集评估以及实景实验验证,所提出的融合框架是有效的,并具备在实际场景中应用的可行性。但是,本文工作仍存在不足,例如还需对特征的选择做进一步探索。在以后的工作中将对更多的特征进行研究,探索性能更优以及鲁棒性更强的融合特征,推动环境声事件识别在实际场景中的应用。

参　考　文　献

[1] Foggia P, Petkov N, Saggese A, et al. Audio surveillance of roads: A system for detecting anomalous sounds[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(1): 279 - 288.

[2] Li Y, Wu Z. Animal sound recognition based on double feature of spectrogram in real environment[C]//2015 International Conference on Wireless Communications & Signal Processing (WCSP), 2015: 1 - 5.

[3] Laffitte P, Wang Y, Sodoyer D, et al. Assessing the per-

formances of different neural network architectures for the detection of screams and shouts in public transportation[J]. Expert Systems With Applications,2019,117:29 - 41.

[4] Babaee E, Anuar N B, Wahab A W A, et al. An overview of audio event detection methods from feature extraction to classification[J]. Applied Artificial Intelligence, 2017, 31(9/10): 661 - 714.

[5] Chu S, Narayanan S, Kuo C C J. Environmental sound recognition with time-frequency audio features[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(6):1142 - 1158.

[6] Huang W, Chiew T K, Li H, et al. Scream detection for home applications[C]//2010 5th IEEE Conference on Industrial Electronics and Applications,2010:2115 - 2120.

[7] Lei B, Mak M W. Sound-event partitioning and feature normalization for robust sound-event detection[C]//2014 19th International Conference on Digital Signal Processing,2014: 389 - 394.

[8] 陈莎莎,李应. 结合时-频纹理特征的随机森林分类器应用于鸟声识别[J]. 计算机应用与软件,2014,31(1):154 - 157,161.

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

[10] Chen Y, Guo Q, Liang X, et al. Environmental sound classification with dilated convolutions[J]. Applied Acoustics, 2019,148:123 - 132.

[11] Salamon J, Bello J P. Deep convolutional neural networks and data augmentation for environmental sound classification [J]. IEEE Signal Processing Letters, 2017, 24(3): 279 - 283.

[12] Zhang X, Zou Y, Shi W. Dilated convolution neural network with LeakyReLU for environmental sound classification [C]//2017 22nd International Conference on Digital Signal Processing (DSP),2017:1 - 5.

[13] Su Y, Zhang K, Wang J, et al. Environment sound classification using a two-stream CNN based on decision-level fusion [J]. Sensors,2019,19(7):1733.

[14] Imoto K, Ono N. Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),2017,25(6):1335 - 1343.

[15] Li S, Yao Y, Hu J, et al. An ensemble stacked convolutional neural network model for environmental event sound recognition[J]. Applied Sciences,2018,8(7):1152.

[16] 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.

[17] Yann L, Bernhard B, John D, et al. Handwritten digit recognition with a back-propagation network[C]//Advances in Neural Information Processing Systems,1990:396 - 404.

[18] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:1 - 9.

[19] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//27th international conference on machine learning (ICML-10),2010:807 - 814.

[20] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// 32nd International Conference on International Conference on Machine Learning,2015.

[21] Ian G, Yoshua B, Aaron C. Deep learning[M]. MIT Press, 2016.

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

[23] Gemmeke J F, Ellis D P, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2017:776 - 780.

[24] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[25] Cheffena M. Fall detection using smartphone audio features [J]. IEEE Journal of Biomedical and Health Informatics, 2015, 20(4):1073 - 1080.

(上接第 72 页)

[14] Naderi B, Zandieh M, Balagh A K G, et al. An improved simulated annealing for hybrid flowshops with sequence-dependent setup and transportation times to minimize total completion time and total tardiness[J]. Expert Systems with Applications,2009,36(6):9625 - 9633.

[15] Tsai J T, Ho W H, Liu T K, et al. Improved immune algorithm for global numerical optimization and job-shop scheduling problems[J]. Applied Mathematics and Computation, 2007,194(2):406 - 424.

[16] 彭佳程. 冷鲜肉品质安全控制技术的研究[D]. 武汉:武汉轻工大学,2014.

[17] 周强,刘蒙佳,张宝善,等. 肉桂精油-壳聚糖涂膜协同气调包装对冷鲜肉品质的影响[J]. 浙江大学学报(农业与生命科学版),2019,45(6):723 - 735.

[18] 李素,赵冰,张顺亮,等. 高氧及 CO2 气调包装对冷鲜猪肉品质的影响[J]. 肉类研究,2016,30(11):16 - 21.