DOI：10.11830/ISSN.1000-5013.202008011

室外人体脚步声事件及环境联合识别

徐峰，李平

（华侨大学 信息科学与工程学院，福建 厦门 361021）

摘要： 为了实现室外人体脚步声事件及环境联合识别，首先，设计一个复杂相似环境下的人体跑动和行走数据集，提出一种交叉双脚步声的分割方案，对连续脚步声信号进行交叉分割；然后，从事件和环境的角度分别提取特征，并从任务平衡的角度设计两种融合特征；最后，采用 3 种深度学习模型对任务进行精确地识别. 结果表明：文中方法简化平衡了任务，使室外人体脚步声事件及环境联合识别的多任务设计不需要复杂模型就能实现精确识别.

关键词： 交叉双脚步声；联合识别；多任务学习；融合特征

中图分类号： TP 391.4 **文献标志码：** A **文章编号：** 1000-5013(2021)05-0676-08

Outdoor Human Footsteps Event and Environment Joint Recognition

XU Feng, LI Ping

(College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China)

Abstract: In order to realize the joint recognition of outdoor human footsteps events and environment, firstly, a human running and walking data set in a complex and similar environment was designed, and a cross double footsteps segmentation scheme was proposed to cross segment the continuous footsteps signals. Then, features were extracted from the perspectives of events and environment, and two fusion features were designed from the perspective of task balance. Finally, three deep learning models were used to identify the task accurately. The results showed that the proposed method simplified and balanced the tasks, and the multitask design of joint identification of outdoor human footsteps events and environment could realize accurate identification without complicated models.

Keywords: cross double footsteps; joint recognition; multitask learning; fusion features

声场景分类(ASC)和声事件检测(SED)是声音场景分析领域中的两个独立任务[1-2]. 许多研究把重点放在一般用途的环境声识别上，很少有专门用于人类活动检测的方法[3]. 一般的声音分类技术与室外人类活动、环境分类之间存在差距，需要考虑室外更加丰富、嘈杂的噪声环境，以及提取的声音特征是否融合了环境和人类活动.

传统的基于隐马尔科夫模型(HMM)[4]、高斯混合模型(GMM)[5]和支持向量机(SVM)[6]的研究方法需考虑声音的持续时间，从而便于对上、下文进行标记[7-8]，此外，使用音频中的关键信息区分场景或

收稿日期： 2020-08-10

通信作者： 李平(1981-)，女，副教授，博士，主要从事智能控制、非线性系统的研究. E-mail：pingping_1213@126.com.

基金项目： 国家自然科学基金资助项目(61603144)；福建省自然科学基金资助项目(2018J01095)；福建省高校产学研合作科技重大项目(2013H6016)；华侨大学中青年教师科技创新资助计划项目(ZQN-PY509)

事件，往往需要人工设定和精心选择，适用性较差. 随着神经网络的发展，深度学习方法比传统方法具有更好的性能[9-11]. 由于声学事件和环境密切相关，近期的相关研究已着眼于二者的联合分析[12-13].

多任务识别模型学习的关键是输入数据中是否包含不同任务的区分特征. 人的连续脚步声数据是一种近似周期信号的数据，可以提取单个周期或单个脚步声信号，即把模型建立在事件的较短持续时间上，使事件和环境的联合识别重点避开模型的复杂度. 就数据处理的角度而言，因为音频信号具有时变特性，基于短帧的特征能够逼近时不变函数和表达细节[14]，所以可将音频流切割成指定长度帧（毫秒级），并提取特征构建模型. 基于此，本文构建一个复杂室外环境下的人体活动数据集，分析交叉双脚步声分割算法和两种融合特征，提出一种室外人体脚步声事件及环境联合识别方法.

1　交叉双脚步声分割

在音频信息中区分非周期信号是音频信号处理领域最重要的问题之一[15]. 在一段特定的时间范围内，脚步声音频信号可近似为周期信号，一连串的脚步声中包含的特征具有重复性，需要将脚步声音频信号进行分割，从而降低数据冗余.

考虑到脚步声的类周期性，一些研究将分割得到的单个脚步声数据作为处理对象，用于后续任务[16]. 然而，就人的运动特点而言，双脚步声数据含有更明显的行为特征，因为人的行走和脚步运动往往以左右或者右左为一个运动周期. 同时，如果只按照双脚步分割原始脚步声，一方面，会造成部分连续性特征的损失，另一方面，背景声的切割会破坏数据，降低识别精度. 因此，提出一种基于包络波谷值的交叉双脚步声分割算法.

基于包络波谷值的交叉双脚步声分割算法如下.

输入：原始脚步声数据 data；时间窗 T_w.

输出：最小的波谷点 $\text{trough}_{\min}$；下一个波谷 $\text{trough}^i_{\text{next}}$；第 i 个脚步波谷 troughstep_i；脚步个数 n；第 j 个交叉双脚步 dstep_j.

步骤 1　初始化变量 $i,j,n,\text{throug}_{\min},\text{troughstep}_i,\text{dstep}_j$.

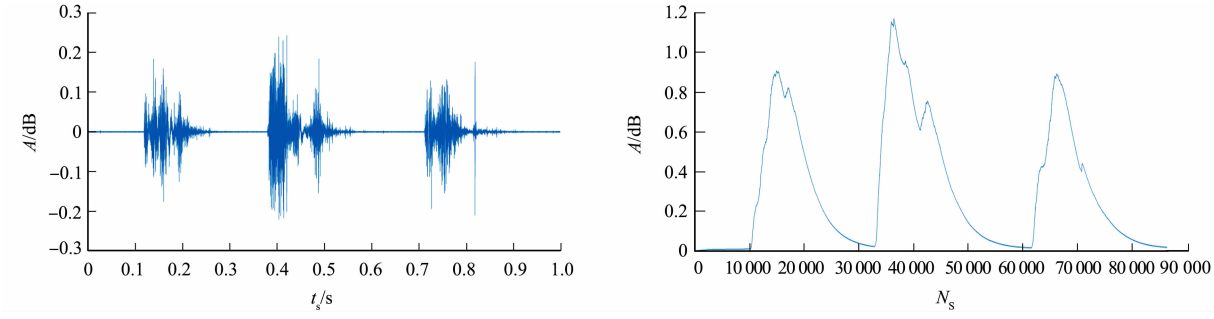
步骤 2　去掉 data 前无声的数据，遍历数据，找到最小的波谷点 $\text{trough}_{\min}$，$\text{troughstep}_i=\text{trough}_{\min}$.

步骤 3　以 troughstep_i 为起点，$i=i+1$，在 T_w 的时间范围内，查找下一个波谷 $\text{trough}^i_{\text{next}}$，$\text{troughstep}_i=\text{trough}^i_{\text{next}}$，$n=n+1$.

步骤 4　重复步骤 3，直至找不到新的脚步波谷值.

步骤 5　按照时间顺序排列 troughstep_i，再初始化 i，分割出第 1 个交叉双脚步 $\text{dstep}_0=\text{data}[0]\sim\text{troughstep}_i$，第 j 个交叉双脚步 $\text{dstep}_j=\text{step}_i+\text{step}_{i+1}$，$i=i+1$，$j=j+1$.

通过信号波形，可清晰判断脚步声的静音段，利用中间的静音分割出一个行动周期内的脚步声. 由于脚步声音频数据极不平滑，无法直接从原始数据波形中找到单个脚步声的波谷值，故先对脚步声数据取包络. 3 个脚步声的波形及包络，如图 1 所示. 图 1 中：A 为振幅；t_s 为脚步声持续时间；N_s 为脚步声采样点数. 根据包络中的单个脚步声波谷值判定单个脚步声的结束位置，再进行分割，可得交叉双脚步声.



（a）波形　　（b）包络

图 1　3 个脚步声的波形及包络

Fig. 1　Waveforms and envelopes of three footsteps

2 特征提取

交叉双脚步声切割时,完整地切割了脚步声数据,切割后的数据长度不同,因此,在特征提取时需保证特征数固定.声音可以平行地被观察到,通过室外人体脚步声数据可对活动和环境进行联合识别,用于该多任务学习的特征需随时间的推移而建立.因此,分别从事件、环境和平衡3个角度进行特征提取.

2.1 事件的角度

声音以压力波的形式存在于当前时刻,声音现象只能当作事件,而不能当作物体.从脚步声判断人的活动是对跨时间依赖事件的分析,提取的特征必须从时间维度进行考虑.

单个脚步周期 t 由脚步声持续时间 t_ε 和脚步声间隔时间 t_τ 组成;交叉双脚步中第1、第2个脚步声的持续时间分别为 t_ε^i,t_ε^j;交叉双脚步中第1、第2个脚步声的间隔时间分别为 t_τ^i,t_τ^j.脚步声的时间表示,如图2所示.图2中:t^i 为第 i 个脚步周期;footshep i,footshep j 分别为第 i,j 个脚步声.脚步周期 t 与人的活动类型关系密切,就文中人的活动形式(跑步、行走)而言,跑步的单脚步周期 t_r 小于行走的单脚步周期 t_w,且脚步声的持续时间和间隔时间的比值关系为 $\frac{t_{w,\varepsilon}}{t_{w,\tau}}<\frac{t_{r,\varepsilon}}{t_{r,\tau}}$,其中,$t_{w,\varepsilon}$,$t_{w,\tau}$ 分别为行走的脚步声持续时间和间隔时间;$t_{r,\varepsilon}$,$t_{r,\tau}$ 分别为跑步的脚步声持续时间和间隔时间.脚步声的这两个特征可以用来区分行走和跑步的活动形式.考虑到需要同时提取活动环境的特征,切割出单个脚步声,获取的环境特征缺乏连续性,且人的脚步运动周期本身就是左右或者右左的循环,因此,从交叉双脚步声数据中,提取事件特征.特征向量 $\boldsymbol{T}$ 为

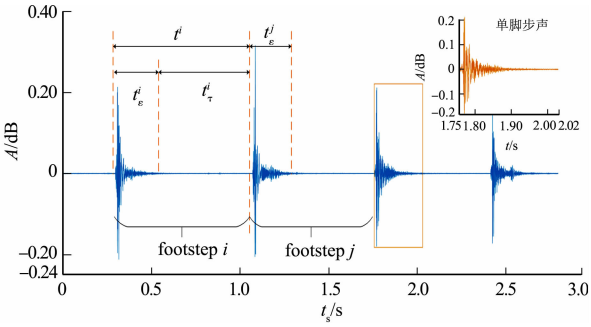


图2 脚步声的时间表示

Fig. 2 Time representation of footsteps

$$\boldsymbol{T}=\left\{t_\varepsilon^i,t_\tau^i,\frac{t_\varepsilon^i}{t_\tau^i},t_\varepsilon^j,t_\tau^j,\frac{t_\varepsilon^j}{t_\tau^j}\right\}. \tag{1}$$

2.2 环境的角度

基于人耳听觉特性提取的梅尔倒谱系数(MFCC)[17]不依赖于信号的性质,可反映语音信号的静态特征,在语音识别和环境声分析中得到了广泛的应用[11].

MFCC是利用梅尔频率(Mel)和物理频率(f)的非线性对应关系得到的物理频率特征,梅尔频率和物理频率之间的关系为

$$\mathrm{Mel}=2\ 595\times\lg(1+f/700). \tag{2}$$

MFCC特征的提取过程(图3)如下.

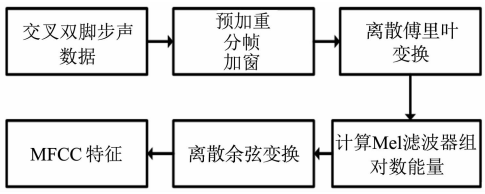


图3 MFCC特征的提取过程

Fig. 3 Extraction process of MFCC features

1) 将统一采样后的交叉双脚步声数据 $y(n)$ 通过预加重滤波器进行预加重,有

$$y(n)'=y(n)-\alpha\times y(n-1),\quad \alpha=0.95. \tag{3}$$

式(3)中:$y(n)'$ 为预加重后的数据;α 为预加重系数.

2) 将 $y(n)'$ 分成短时帧 $s(n)$,$n=0,1,\cdots,N-1$,N 为帧的大小,$N=512$;统一采样率 r_s 为22 050.经计算可得覆盖时间 t_c(单位为ms)为

$$t_c=1\ 000N/r_s\approx 23. \tag{4}$$

3) 采用汉明窗 $W(n)$ 进行加窗,窗外值设定为0,将每一帧与汉明窗相乘,可得时域信号 $s(n)'$,有

$$\left.\begin{aligned}W(n)&=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0\leqslant n\leqslant N-1,\\ 0, & \text{其他}.\end{cases}\\ s(n)'&=s(n)\cdot W(n).\end{aligned}\right\} \tag{5}$$

4) 将时域信号 $s(n)'$ 转化到频域后,进行频率分析,经离散傅里叶变换(DFT)后的频谱 $S(k)$ 为

$$S(k)=\sum_{n=1}^{N}s(n)'\exp\left(\frac{-\mathrm{j}2\pi kn}{N_{\mathrm{w}}}\right),\qquad 1\leqslant k\leqslant N_{\mathrm{w}}. \tag{6}$$

式(6)中：k 为傅里叶变换的点数；N_{w} 为加窗后的采样点数.

5) 计算功率谱，并将每帧谱线能量 $|S(k)|^2$ 通过 Mel 滤波器组后取对数，得到对数能量 $H_m(k)$ 和 $S(m)$，有

$$\left.\begin{aligned}H_m(k)&=\begin{cases}\dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\leqslant k\leqslant f(m),\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)\leqslant k\leqslant f(m+1),\\ 0, & \text{其他}.\end{cases}\\ S(m)&=\ln\left(\sum_{k=0}^{N-1}|S(k)|^2H_m(k)\right),\qquad 0\leqslant m\leqslant M.\end{aligned}\right\} \tag{7}$$

式(7)中：$f(m)$为第 m 个三角滤波器中心频率；M 为三角滤波器的个数，也表示 MFCC 的维度.

$S(m)$经离散余弦变换(DCT)，得到梅尔倒谱系数，表示为

$$C(l)=\sum_{m=0}^{M-1}S(m)\cos\left(\frac{\pi l(m+0.5)}{M}\right),\qquad 1\leqslant l\leqslant L. \tag{8}$$

式(8)中：L 为 MFCC 阶数.

MFCC 特征是经 Librosa 0.7 数据处理库调整数据维度计算得到，其中，wetleaves 和 metal 的 MFCC 特征可视化，如图 4 所示. 由图 4 可知：不同环境下脚步声数据的 MFCC 特征区别较大. 周期性声音信号通常是由一个基波和若干谐波组成，这些谐波由声源按照特定的关系隔开，谐波的混合决定了声音的音色；频率的分布是非局部分布的，信号特征能够表示当前的活动环境.

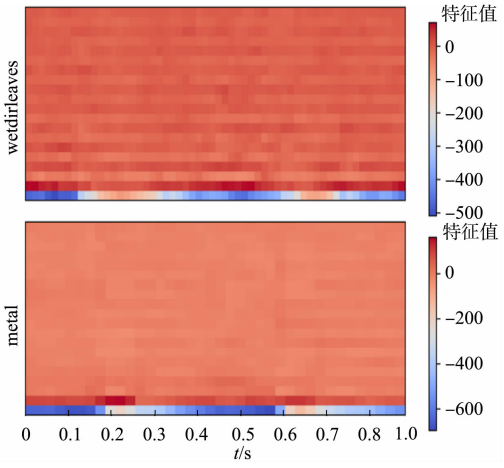


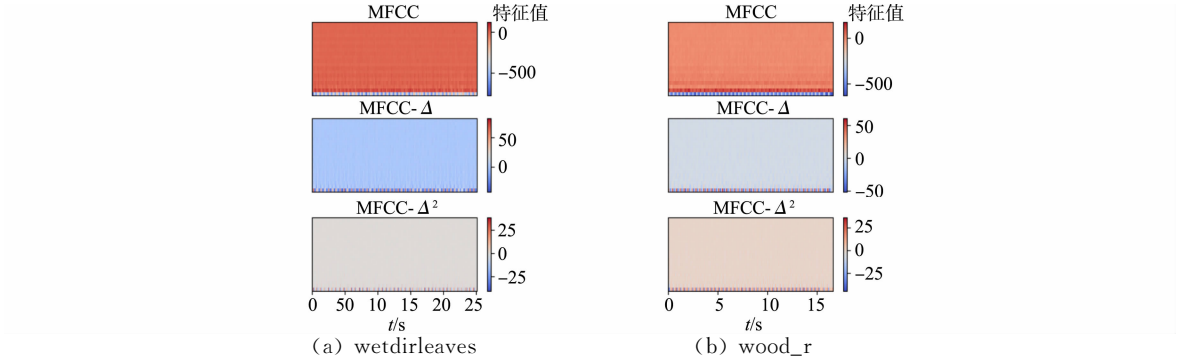
图 4　wetleaves 和 metal 的 MFCC 特征可视化

Fig. 4　MFCC visualization of wetleaves and metal

2.3　平衡的角度

虽然只采用 MFCC 特征区分环境声的效果更加突出，但会破坏联合识别多任务学习的平衡性. 因此，提取一阶差分 MFCC_Δ 和二阶差分 MFCC_Δ^2，以反映音频信号的动态特征，加大特征对事件的敏感度. MFCC_Δ 表示当前 MFCC 相邻两项的差，可体现交叉双脚步声相邻两帧的关系；MFCC_Δ^2 表示当前 MFCC_Δ 相邻两项的关系，可体现交叉双脚步声相邻 3 帧的动态关系.

单个领域的特征只代表有限信息，为使模型学习更加平衡，从 2 个融合特征方向进行考虑. 1) MFCC+T；2) MFCCs，MFCCs=MFCC+MFCC_Δ+MFCC_Δ^2. 为便于比较，融合特征维度保持一致，取 36 维. 其中，MFCC+T 由 MFCC∶T=30∶6组成；MFCCs 由 MFCC∶MFCC_Δ∶MFCC_Δ^2=12∶12∶12 组成. wetdirleaves 和 wood_r 的 MFCC，MFCC_Δ，MFCC_Δ^2 的特征对比，如图 5 所示.



(a) wetdirleaves　　(b) wood_r

图 5　wetdirleaves 和 wood_r 的 MFCC，MFCC_Δ，MFCC_Δ^2 的特征对比

Fig. 5　Characteristics comparison of MFCC, MFCC_Δ, MFCC_Δ^2 of wetdirleaves and wood_r

3 实验分析

3.1 实验数据

为了研究室外人体脚步声活动事件和环境的联合识别,构建一个数据集,其相关资料,如表1所示.通过学习数据本身的性质实现事件和环境的联合识别.

表1 数据集的相关资料

Tab.1 Relevant information of data set

数据集来源	结构	格式	描述
WARCAKESTUDIOS[18]	4.8 kHz,16 bit,00:09.240	.wav	在潮湿的石头上快速踩踏声,用T-Bone微型录音机录制
INSPECTORJ[19]	44.1 kHz,24 bit,00:16.479	.wav	穿着运动鞋在碎石车道顶部的薄碎冰上跑步的原始音频
AUDIONINJA001[20]	44.1 kHz,16 bit,04:45.948	.wav	多种环境下人的脚步声,用zoom h5,sennheiser mk600录音机录制

通过交叉脚步声分割的方法,得到模型的输入数据,切割后的脚步声数据分布,如图6所示.图6中:b为脚步声数据的数量.由于脚步声的类周期性,实验数据不需要非常大.数据类型分为11类,共586个脚步声数据.为了增加模型的泛化能力,胜任多任务识别,采用的数据类别和数量都是特定选取的,数据类标签带有_r的表示跑步状态,其他表示行走状态.该设计可保证模型能从脚步声中区分多种环境.特别的,如wetdirleaves和wetleaves这两种类别的脚步声较为相似,需仔细聆听才能区分出两种脚步声的场景.此外,选取一组wood场景下,活动类型为跑步和行走的脚步声,用于训练模型活动类别的区分能力.数据集特别选取了两对场景相似的跑步和行走状态的脚步声(wetsones_r/wetsand,ice_r/mud),以保证模型能够在区分活动类型的同时也能区分相似的活动场景.

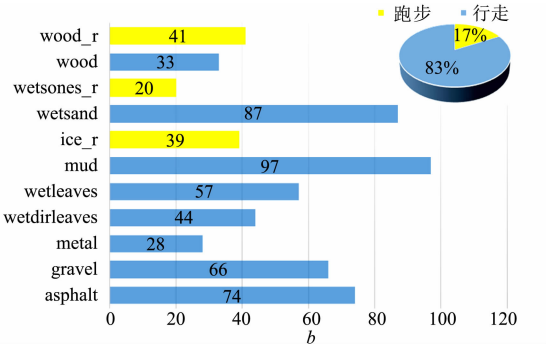


图6 脚步声数据分布

Fig.6 Footstep data distribution

3.2 实验方法

脚步声对象本身相对较简单,重点在于设计融合特征,在保证较高精度的前提下,实现事件和环境的联合识别.适合的融合特征设计能够简化任务,采用较为简单的模型即可达到多任务识别的要求.建立如下3种较为简单的深度学习模型,对室外脚步声事件和环境进行联合识别.

1)四层感知机(MLP)模型(无状态分类算法).每个隐层神经元个数为50,激活函数采用ReLU,输出层采用Softmax.

2)卷积(CNN)模型(无状态分类算法),如图7所示.图7中:Max_pooling层后Dropout设置为0.3.

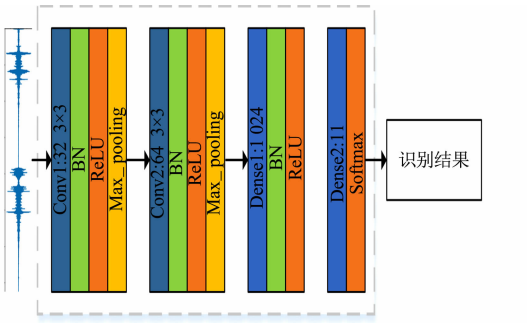


图7 卷积模型

Fig.7 Convolution model

3)循环神经网络(GRU)模型(有状态分类算法).在任务中,实验处理得到交叉双脚步声数据,针对活动事件的识别,提取的特征与时间密切相关,采用有状态分类算法门控循环单元网络完成任务,单个隐层神经元个数为50.同时,与前两种无状态分类算法进行比较.

采用3种模型进行实验对比.学习率为0.01,batch_size为200,迭代次数为1 000次.

3.3 实验结果

对交叉双脚步声数据提取两种融合特征MFCC+T,MFCCs,分别采用3种模型训练,进行联合识别.从实验结果中统计11类数据的真正例(TP)、假正例(FP)、真反例(TN)及假反例(FN),得到分类结果的混淆矩阵.定义查准率P、查全率R及F_1分数分别为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{9}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{10}$$

$$F_1=\frac{2PR}{P+R}. \tag{11}$$

不同模型的 MFCC＋T，MFCCs 融合特征分类精度，如表 2，3 所示. 表 2，3 中：δ 为分类精度. 由表 2，3 可得以下 3 点结论. 1) 两种融合特征起到很好的识别效果. 2) 相较于 MLP 模型、GRU 模型，CNN 模型的建模效果更加突出. 同时，MFCC＋T 比 MFCCs 表现得更加稳定，这是因为除了 MFCC 特征，还额外从事件的角度上提取特征 T，融合特征 MFCC＋T 一方面保证了分类精度，另一方面，使分类模型在事件和环境上建模平衡. 3) 在 MLP 模型下，MFCCs 的效果最差，其对文中数据集相似环境的区分效果不佳，这是由于增加任务复杂性后，设计的数据中部分环境容易混淆. 混淆矩阵(图 8，d 为识别结果的实际数量)中 wetleaves，wetdirleaves 的识别效果最差，两类环境非常相似. 相较而言，采用 CNN 模型后，MFCC＋T 能达到最佳效果. 实验中发现 GRU 模型在迭代多次后才开始收敛，虽然收敛速度最快，但非常不稳定，效果一般. 此外，MFCC＋T 可在无状态分类算法中起到比有状态分类算法更好的效果.

由实验可知，原始室外人体脚步声经过交叉双脚步声分割后，提取事件与环境的融合特征，采用较为简单的深度模型就能实现室外人体脚步声事件与环境联合识别.

表 2　不同模型的 MFCC＋T 融合特征分类精度

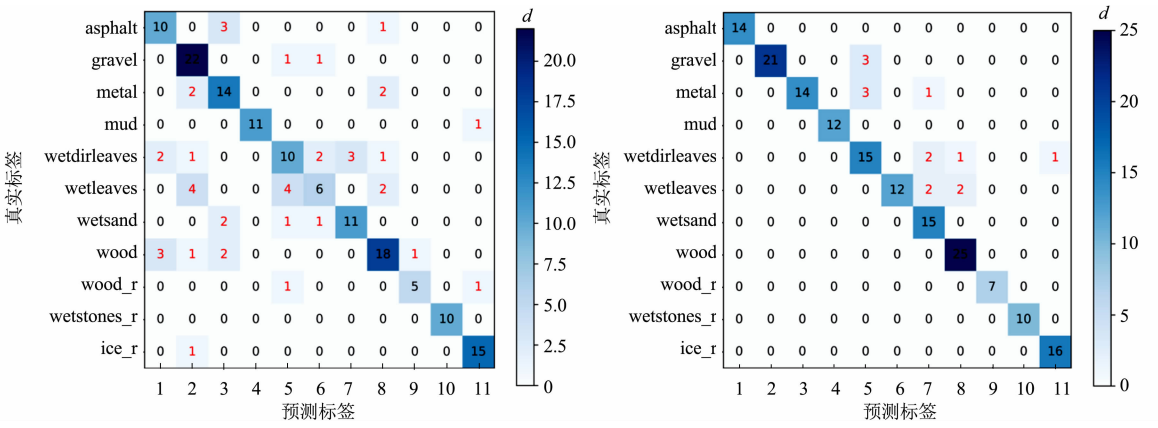
Tab. 2　Classification accuracy of MFCC＋T fusion feature of different models

数据类型	MLP 模型			CNN 模型			GRU 模型		
	P	R	F_1	P	R	F_1	P	R	F_1
gravel	0.93	0.93	0.93	1.00	1.00	1.00	0.77	0.71	0.74
asphalt	0.81	0.71	0.76	1.00	0.88	0.93	0.85	0.92	0.88
metal	0.84	0.89	0.86	1.00	0.78	0.88	0.77	0.94	0.85
mud	1.00	0.92	0.86	1.00	1.00	1.00	1.00	0.42	0.59
wetdirleaves	0.62	0.68	0.65	0.71	0.79	0.75	0.73	0.84	0.78
wetleaves	0.73	0.69	0.71	1.00	0.75	0.86	0.83	0.31	0.45
wetsand	0.94	1.00	0.97	0.75	1.00	0.86	0.83	1.00	0.91
wood	0.96	0.96	0.96	0.89	1.00	0.94	0.83	1.00	0.91
wood_r	1.00	1.00	1.00	1.00	1.00	1.00	0.86	0.86	0.86
wetstones_r	0.91	1.00	0.95	1.00	1.00	1.00	1.00	0.90	0.95
ice_r	1.00	1.00	1.00	0.94	1.00	0.97	0.89	1.00	0.94
δ/%	86.93			91.48			82.95		

表 3　不同模型的 MFCCs 融合特征分类精度

Tab. 3　Classification accuracy of MFCCs fusion feature of different models

数据类型	MLP 模型			CNN 模型			GRU 模型		
	P	R	F_1	P	R	F_1	P	R	F_1
gravel	0.67	0.71	0.69	1.00	0.71	0.83	0.93	0.93	0.93
asphalt	0.71	0.92	0.80	0.95	0.88	0.91	0.88	0.92	0.90
metal	0.67	0.78	0.72	0.89	0.94	0.92	0.94	0.89	0.91
mud	1.00	0.92	0.96	0.92	1.00	0.96	0.86	0.50	0.63
wetdirleaves	0.59	0.53	0.56	0.74	0.89	0.81	0.84	0.84	0.84
wetleaves	0.6	0.38	0.46	0.86	0.75	0.8	0.93	0.81	0.87
wetsand	0.79	0.73	0.76	0.88	0.93	0.9	0.81	0.87	0.84
wood	0.75	0.72	0.73	1.00	0.88	0.94	0.83	0.96	0.89
wood_r	0.83	0.71	0.77	0.88	1.00	0.93	1.00	0.86	0.92
wetstones_r	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ice_r	0.88	0.94	0.91	0.84	1.00	0.91	0.74	0.88	0.80
δ/%	75.00			89.77			86.93		



(a) MFCCs(MLP模型) (b) MFCC+T(CNN模型)

图8 混淆矩阵

Fig. 8 Confusion matrix

4 结束语

提出一种室外人体脚步声事件及环境联合识别的多任务学习方法,对提出的复杂相似环境下的人体跑动和行走脚步声数据设计分割算法,得到交叉双脚步声数据,从而便于脚步声事件及环境特征的提取,通过融合事件与环境特征平衡任务,能够用简单的模型较精准地实现室外人体脚步声事件及环境的联合识别.由此可知,部分联合识别任务从预处理和特征融合的角度出发,可采用简单模型实现精准识别,简化任务.

参考文献:

[1] MESAROS A, HEITTOLA T, BENETOS E, *et al*. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2017, 26(2): 379-393. DOI: 10.1109/TASLP.2017.2778423.

[2] VIRTANEN T, MESAROS A, HEITTOLA T, *et al*. Proceedings of the detection and classification of acoustic scenes and events 2017 workshop (DCASE2017)[R/OL]. (2017-09-15)[2019-12-10]. https://www.researchgate.net/publication/320409431_Deep_Sequential_Image_Features_on_Acoustic_Scene_Classification.

[3] PICZAK K J. Environmental sound classification with convolutional neural networks[C]// IEEE 25th International Workshop on Machine Learning for Signal Processing. Boston: IEEE Press, 2015: 1-6. DOI: 10.1109/MLSP.2015.7324337.

[4] MESAROS A, HEITTOLA T, ERONEN A, *et al*. Acoustic event detection in real life recordings[C]// 18th European Signal Processing Conference. Aalborg: IEEE Press, 2010: 1267-1271.

[5] YUN S, KIM S, MOOM S, *et al*. Discriminative training of GMM parameters for audio scene classification[R/OL]. (2016-03-02)[2019-12-10]. https://www.aminer.cn/pub/5f44d5c49e795ee83b76546b/discriminative-training-of-gmm-parameters-for-audio-scene-classification-and-audio-tagging.

[6] RAKOTOMAMONJY A, GASSO G. Histogram of gradients of time-frequency representations for audio scene detection[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2015, 23(1): 142-153. DOI: 10.1109/TASLP.2014.2375575.

[7] CAI Rui, LU Lie, ZHANG Hongjiang, *et al*. A flexible framework for key audio effects detection and auditory context inference[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(3): 1026-1039. DOI: 10.1109/TSA.2005.857575.

[8] HEITTOLA T, MESAROS A, ERONEN A, *et al*. Context-dependent sound event detection[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1): 1-13. DOI: 10.1186/1687-4722-2013-1.

[9] MESAROS A, DIMENT A, ELIZALDE B, *et al*. Sound event detection in the DCASE 2017 challenge[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2019, 27(6): 992-1006. DOI: 10.1109/TASLP.2019.

2907016.

[10] IMOTO K, KYOCHI S. Sound event detection using graph Laplacian regularization based on event co-occurrence [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE Press, 2019: 1-5. DOI: 10.1109/ICASSP.2019.8683708.

[11] CHAUDHURI S, RAJ B. Unsupervised hierarchical structure induction for deeper semantic analysis of audio[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE Press, 2013: 833-837. DOI: 10.1109/ICASSP.2013.6637765.

[12] TONAMI N, IMOTO K, NITTSUMA M, *et al*. Joint analysis of acoustic events and scenes based on multitask learning[C]// IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York: IEEE Press, 2019: 338-342. DOI: 10.1109/WASPAA.2019.8937196.

[13] BEAR H L, NOLASCO I, BENETOS E. Towards joint sound scene and polyphonic sound event recognition[C]// Interspeech. Graz: [s. n.], 2019: 4594-4598. DOI: 10.21437/Interspeech.2019-2169.

[14] WANG Wei, SERAJ F, MERATNIA N, *et al*. Privacy-aware environmental sound classification for indoor human activity recognition[C]// Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments. New York: Association for Computing Machinery, 2019: 36-44. DOI: 10.1145/3316782.3321521.

[15] SHOLOKHOV A, SAHIDULLAH M, KINNUNEN T. Semi-supervised speech activity detection with an application to automatic speaker verification[J]. Computer Speech and Language, 2018, 47(1): 132-156. DOI: 10.1016/j.csl.2017.07.005.

[16] HORI Y, ANDO T, FUKUDA A. Personal identification methods using footsteps of one step[C]// International Conference on Artificial Intelligence in Information and Communication. Tianjin: [s. n.], 2020: 73-78. DOI: 10.1109/ICAIIC48513.2020.9065230.

[17] 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. DOI: 10.1109/TASL.2009.2017438.

[18] WARCAKESTUDIOS. Fast steps on wet stones: Recorded with a *T*-Bonemicro[EB/OL]. (2013-01-10)[2019-12-17]. https://freesound.org/people/WarcakeStudios/sounds/173596/.

[19] INSPECTORJ. Raw audio of running on thin, cracked ice on top of a gravel driveway with trainer shoes[EB/OL]. (2018-01-30)[2019-12-17]. https://freesound.org/people/InspectorJ/sounds/416967/.

[20] AUDIONINJA001. Recorded with zoom h5 and sennheizer mk600[EB/OL]. (2018-12-27)[2019-12-17]. https://freesound.org/people/audioninja001/packs/25644/.

（**责任编辑**：钱筠　　**英文审校**：吴逢铁）