计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500091

基于深度学习的无人机声音识别算法

徐 浩 刘岳镭
长安大学电子与控制工程学院 西安 710016
(2017900299@chd.edu.cn)

摘 要 深度学习在图像识别和声音处理方面已经展现了它优越的性能和广阔的发展前景,对于在禁飞区设立的无人机侦测系统,使用深度学习的方法判断无人机的声音信号具有一定的意义。为了获得更优的侦测效果,首先列举了目前具有代表性的特征提取和分类方法,并分析其优缺点;然后提出了一种扩大可用样本数量的数据处理方式,同时在实验中使用不同组合的深度学习网络训练样本;最后通过混淆矩阵法,针对不同信噪比模型、滤波下限、拟合程度、神经网络组合和跨型号识别的实验效果进行评价。实验结果表明,适当地降低训练样本中的无人机声强可以增大系统的识别距离;使用 MFCC 提取声音特征,通过全连神经网络进行分类的样本识别的半径更远,误判率更低。

关键词:声音识别;深度学习;无人机;混淆矩阵
中图法分类号 TP391

UAV Sound Recognition Algorithm Based on Deep Learning

XU Hao and LIU Yue-lei
School of Electronics and Control Engineering,Chang' an University,Xi' an 710016,China

Abstract Deep learning has demonstrated its superior performance and broad development prospect in image recognition and sound processing. It is of certain significance for the UAV detection system established in no-fly zone to use deep learning method to judge the sound signal of UAV. In order to obtain better detection effect,the representative feature extraction and classification methods are listed at first,and their advantages and disadvantages are analyzed. Then,a method of data processing is proposed to expand the number of available samples. At the same time,different combinations of deep learning network training samples are used in the experiment. Finally,the confounding matrix method is used to evaluate the experimental results of different SNR models,filtering limits,fitting degrees,neural network combinations and cross-model recognition. The results show that reducing the sound intensity of the UAV can improve the recognition distance of the system. By using MFCC to extract the sound features,the samples classified by the fully connected neural network have a longer identification radius and a lower misjudgment rate.

Keywords Sound detection,Deep learning,Unmanned aerial vehicle,Obfuscation matrix

1 引言

近年来,无人驾驶飞行器(Unmanned Aerial Vehicles,UAV)被广泛应用于各行各业。文献[1-2]综述了无人机的发展现状和前景,指出无人机在农业、工矿业的普及使生产力得到大大提高。文献[3]研究并总结了无人机在测量和遥感方面的应用。文献[4]依托军用无人机实现了灾害管理和救援。目前,无人机已经成为推动社会进步的有力助手。

然而,无人机在推动社会发展的同时,也不可避免地带来了社会安全问题。2019 年 1 月,深圳市通过了《民用微轻型无人机管理暂行办法》,同年 3 月,浙江省人大常委会审议通过了《无人驾驶航空器公共安全管理规定》。这些政策表明了我国对无人机监管领域的高度重视,但仍有人无视法规在禁飞区使用无人机。例如,文献[5]尖锐地指出了一些不法行为:无人机被用于走私毒品、干扰民航航道和导航系统。因此,设计一个无人机侦测系统对于民生保障和国家安全具有重要意义。

在研发新型无人机识别系统方面,前人已取得了较多的成果。文献[6]系统地分析了识别无人机所面临的困难,并评价了雷达、声音、图像和射频无线电作为监测信号的优缺点。文献[7]使用摄像机阵列实现了无人机的识别功能,它利用 AlexNet 网络和 Single Shot MultiBox Detector 算法对图像进行分析,网络训练时间约为 6h。但是,受到探测手段的限制,该系统很容易将无人机与鸟等小型飞行物混淆,且不同型号或在不同角度拍摄的无人机外形差异大,因此误报率高达 20%。文献[8]选择声音作为侦测无人机的信号,并主张使用支持向量机(Support Vector Machines,SVM)对声音进行分类,凭借着从低维向高维的变换巧妙地避开"维数灾难问题"。其采用"一对一"的策略实现多分类,但是所提算法运行较慢,使得该系统并不适用于实时侦测。文献[8]中 SVM 的分类

到稿日期:2020-05-20 返修日期:2020-08-23 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:刘岳镭(276945401@qq.com)

模型如图 1 所示。

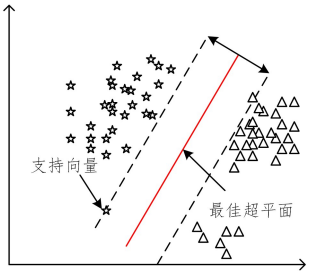


图 1 支持向量机的分类示例

Fig. 1 Support vector machine classification example

文献[9]在最小二乘支持向量机(Least Squares Support Vector Machine,LSSVM)和近似支持向量机(Proximal Support Vector Machine,PSVM)的基础上进行进一步简化,建立极限学习机(Extreme Learning Machine,ELM)对 SVM 多分类问题进行优化,运算速度显著提升,但没有根据新算法给出应用实例。除以上直接针对无人机进行的研究外,文献[10]基于卷积神经网络(Convolutional Neural Networks,CNN)对城市场景声音进行识别;文献[11]将梅尔倒谱与其 delta 导数相结合,设计了语音识别系统,且辨识率达到了 94%;文献[12]中提出使用小波包 Filter Bank 图谱与 CNN 的车内异响识别方法,这种方案的识别率优于传统高斯混合模型与支持向量机模型。文献[10-12]均使用了梅尔倒谱或相关结构作为声音特征进行分类,对本文中无人机声的特征提取和识别具有指导意义。

本文在已有研究的基础上对比了不同神经网络和音频特征对侦测系统识别效果的影响,并综合考虑了滤波和过拟合问题,最终找到了使识别距离更远、稳定性更高的算法和模型参数。

2 原理分析

2.1 MFCC 提取声音特征

梅尔倒谱系数即 Mel Frequency Cepstral Coefficient,MFCC。人耳对声音的敏感程度与频率的对数成正比,梅尔刻度利用了这一特性,它与频率的关系如下:

$$Mel(f)=2\,595\times\lg\left(1+\frac{f}{700}\right) \tag{1}$$

当两种声音的频率差小于某一带宽时,梅尔倒谱算法将不再仔细区分,这个带宽由梅尔刻度进行衡量。MFCC 对人耳的仿生使得它特别适合用于语音识别,相似的算法还有基于声道模型的线性预测倒谱系数(Linear Predictive Cepstral Coefficient,LPCC),两者都可以为分类器提供显著的声学特征。对 LPCC 的分析和应用详见文献[13-14]。

经过傅里叶变换和三角滤波得到的梅尔频谱(mel-spectrogram)也是分析声音信号的一种方式。两者的区别在于 MFCC 多执行了取对数和离散余弦变换(Discrete Cosine Transform,DCT)两个步骤。DCT 的主要任务是降低声音的特征维度,实现去冗余的效果。图 2 给出了 MFCC 提取特征的步骤,在该流程中的第六步求得了梅尔频谱。参考文献[15],下文针对图 2 中的各个步骤给出具体解释。

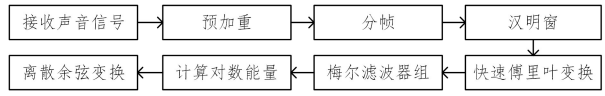


图 2 梅尔频率倒谱系数计算方法

Fig. 2 Calculation method of Mel frequency cepstrum coefficient

(1)预加重。预加重的本质是将语音信号通过一个高通滤波器。

$$H(z)=1-\mu z^{-1} \tag{2}$$

其中,μ 为常数,常取值为 0.97。预加重使输入信号功率谱中较少的高频部分得到提升,避免了低频过强高频过弱而导致的信号失真。处理后的信号为:

$$S_1(n)=s(n)-\mu s(n-1) \tag{3}$$

其中,$s(n)$为原始音频信号。

(2)分帧。将若干采样点合并为一帧,帧长度一般为 128~512,每帧时间约为 20~30 ms。为了避免相邻两帧差异过大,通常在两帧之间保留一段重叠,常取重叠率为 50%。

(3)加汉明窗。若不加处理,则每帧信号的两端会不连续,在傅里叶展开后会出现吉布斯现象。添加汉明窗可以使帧两端的信号逐渐减弱至 0,保证了两帧信号的连续性。

若规定$S_2(n,i)$为第 i 帧音频信号,则加窗操作可表示为:

$$S_2(n,i)=S_1(n,i)\times(1-a)-\operatorname{acos}\frac{2\pi n}{N-1} \tag{4}$$

其中,a 值决定汉明窗的滤波效果,a 常取值为 0.46,N 为帧长,也是汉明窗的宽度。

(4)快速傅里叶变换(Fast Fourier Transform,FFT)。对加窗后的信号进行 FFT 变换,即可得到傅氏频率信号谱。

$$F(k,i)=\sum_{n=0}^{N-1}S_2(n,i)\times e^{\frac{-j2kn\pi}{N}},0\leqslant kn\leqslant N \tag{5}$$

(5)三角带通滤波器。将傅氏频谱输入一组梅尔刻度的三角滤波器组,滤波后对信号求对数,以模拟人耳听觉。添加该滤波器的主要目的是使频谱平滑,并消除谐波,突显原先语音的共振峰,同时也可降低计算量。每个滤波器组输出的对数能量为:

$$s(m)=\ln\left(\sum_{k=0}^{N-1}|F(k,i)|^2\times H_m(k)\right),0<m<M \tag{6}$$

其中,M 为滤波器的数量,$H_m(k)$为三角滤波器的频率响应,详细信息请见文献[16]。常见的等幅三角滤波器如图 3 所示。

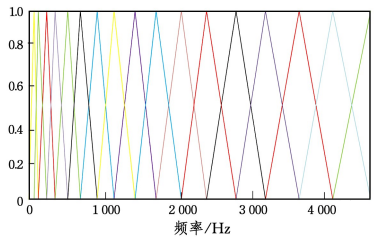


图 3 梅尔频率三角滤波器组

Fig. 3 Mel frequency triangular filter bank

(6)离散余弦变换(DCT)。DCT 的作用是对滤波后的信号进行无损的降维,去除声音中的冗余信号,用于语音识别的 MFCC 一般取前 13 位。设 $f(k,l)$为二维离散函数,$k,l=0,1,\cdots,N-1$,$F(x,y)$为返回的倒谱信号,则计算式为:

$$F(x,y)=c(x)c(y)\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}f(k,l)\cos\frac{(2k+1)x\pi}{2N}\cos\frac{(2l+1)y\pi}{2N} \tag{7}$$

其中，$c(x)=c(y)=\begin{cases}\frac{1}{\sqrt{N}}, & x=0 \text{ 或 } y=0\\ \frac{\sqrt{2}}{\sqrt{N}}, & x,y=1,2,\cdots,N-1\end{cases}$

2.2　全连神经网络分类器

随着深度神经网络的不断发展，CNN 在图像处理方面起到了不可忽视的作用。常见的 CNN 包括输入层、卷积层、池化层和全连接层。其中，卷积层和池化层在图像处理中的作用是提取图片特征，全连接层的作用是对特征进行整合，实现回归或分类，详细信息请见文献[17]。

在声音识别的深度学习模型中，特征提取由前述的 MFCC 得出，它实际上已经实现了特征提取的功能。若音频文件较多且较大，可以进一步卷积；若数据量不大，则不需要进一步加工，舍弃卷积层和池化层并不会对数据分类产生重要影响。声音频率信号求取倒谱后直接输入全连神经网络进行分类，进而得到结果。全连接层和 ReLU 激活函数的功能可表示为：

$$y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \tag{8}$$

$$f(\boldsymbol{y})=\max(0,\boldsymbol{y}) \tag{9}$$

其中，$\boldsymbol{w}^{\mathrm{T}}$ 表示待训练的系数矩阵，$\boldsymbol{x}$ 是输入矩阵，b 代表偏差，$\boldsymbol{y}$ 为输出矩阵。文献[18]指出，ReLU 函数可以将输出的值与 0 进行比较，将较大的结果作为输出，激活函数可以有效地解决非线性分类问题。

上文提到实验的目的是对声音信号进行二分类，即判断是否有无人机，但二分类并不代表全连接层的输出结果只有两种。为了保证识别系统具有较强的鲁棒性，训练时应将各种自然噪声考虑在内，如车流声和人声，对这些声音进行逐一分类，最后将这些背景声音统一归为没有无人机。

3　数据收集与处理

3.1　实验数据收集

用于训练的无人机型号为大疆 TLW004 系列，录制地点为安静的室内，内容包括悬停、上升下降和盘旋。实验所用的声音数据主要借助于 Audition 音频剪辑软件，将纯净的无人机声与自行录制或来源于公共音频集包含多种生活声音源（如 CPU 风扇转动声、人员交谈声、街边车流声、狗叫声等）的环境音混合到一起。该数据可满足训练样本的数量和多样性要求，还尽可能地保证了样本中无人机声音的准确性。

本文通过表 1 对原始数据做出了说明。我们将训练集和测试集分开来确保能够严格地评估网络性能。

表 1　原始音频数据说明

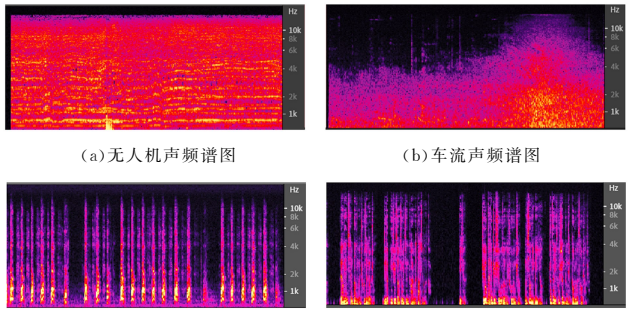
Table 1　Description of raw audio data

数据类型	总时间/s	备注
原始背景音	910	与无人机声混合之后使用
原始无人机声	904	与背景音混合之后使用
扩大后的数据	5142	混合扩充之后用于训练
测试数据	300	测试数据采集于城镇街道

在混音时调节无人机音轨的声音强度，可以得到多种信噪比的训练音频。调节信噪比时，由于背景音为实地取材，因此混音时背景音的声音强度不需要改变，为了模拟无人机与话筒距离远近的变化，只需要将原始无人机音频的声音强度分成不同的级别。这不仅能够获取更多的音频信息，还可以适应不同的背景环境。在实际应用中，无人机与探测话筒的距离有近有远，反映在声音信号上就有强有弱。我们希望系统能够在保证正确性的前提下，识别的距离越远越好。

3.2　滤波预处理

文献[19]指出，在前述的自然噪声中，人的说话声最常见，其基波频率为 60～400 Hz，高次谐波频率则可达到数千赫兹；路边杂音的成分较复杂，且具有较大的偶然性，一般不超过 3000 Hz；狗叫声约为 400～1000 Hz。与自然噪声相反，无人机声可以在 15000 Hz 以下的频段检测到明显波动，图 4 和图 5 直观地体现这一差异。另外，又考虑到自然界中可能存在 15000 Hz 以上的高频杂音干扰，因此，为了防止上述自然噪声对识别系统造成干扰，引起误判，我们为输入的音频加一个上限为 15000 Hz 且下限不低于 2000 Hz 的带通滤波器。每一个样本在测试之前都要经过滤波，另外，为了提高模型的性能，保证训练集与测试集样本一致，训练集的样本也先经过滤波再用于训练。

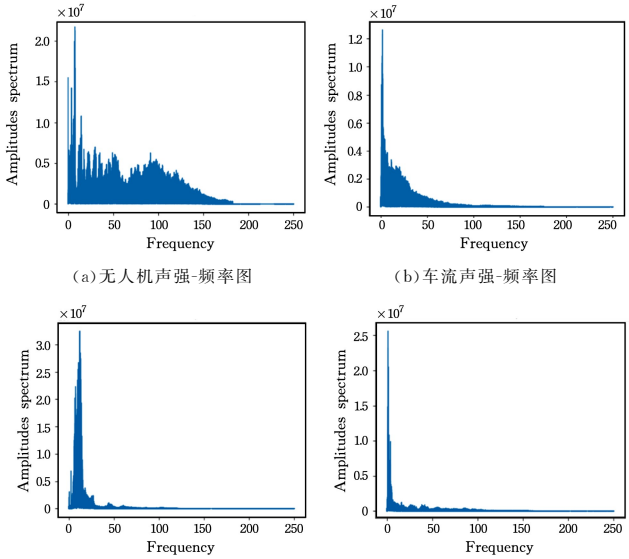


(a)无人机声频谱图　(b)车流声频谱图

(c)狗叫声频谱图　(d)人声频谱图

图 4　不同声音样本的频谱图

Fig. 4　Spectrum of different sound samples



(a)无人机声强-频率图　(b)车流声强-频率图

(c)狗叫声强-频率图　(d)人说话声强-频率图

图 5　不同声音样本频率的直观表示图

Fig. 5　Intuitive representation of the frequencies of different sound samples

4 实验设计与结果分析

4.1 特征提取算法的选择与训练网络的搭建

根据文献[20]对神经网络的研究可知，在使用正方形卷积核的情况下，CNN 网络中单个卷积层的时间复杂度为：

$$Time \sim O(M^2 \times K^2 \times Cin \times Cout) \tag{10}$$

其中，K 为卷积核(Kernel)的边长，Cin 为输入通道数，$Cout$ 为输出通道数，M 为输出特征图(feature map)的尺寸，M 的表达式为：

$$M=(X-K+2*Padding)/Stride+1 \tag{11}$$

其中，X 为输入级的长度，$Padding$ 为补零层数，$Stride$ 为卷积步长。

而在全连神经网络中，单个全连接层可以视作一种 $X=K$，$M=1$ 的卷积层，其时间复杂度为：

$$Time \sim O(X^2 \times Cin \times Cout) \tag{12}$$

参考 AlexNet 网络，即卷积步长和补零层数取 1，卷积核边长取 3，则单个全连接层的时间复杂度总是小于单个卷积层的时间复杂度，这说明深度相同的全连网络可以比卷积网络获得更快的运算速度，而鉴别的快速性在本任务中尤为重要。

表 2 列出了全连神经网络和卷积神经网络的运行时间。本文进行测试的平台为：基于 Python3.5 的 Tensorflow 1.10.0 版本，Librosa0.7.2 音频处理库。硬件方面为：四核 2.5GHz 的 CPU，固态硬盘和英伟达 GTX1050 显卡。全连神经网络使用了 MFCC 和梅尔频谱两种方式来提取特征。由于 MFCC 相对于梅尔频谱来说，返回的特征值过少(用于识别语音的 MFCC 算法只在单个时域采样点返回 13 个数据)，因此令卷积神经网络只接收梅尔频谱信息。

表 2 不同网络运算速度对比

Table 2 Comparison of different network computing speeds

神经网络类型	声音特征类型	运算操作	计算速度(秒每千次)
全连神经网络	MFCC	训练	1.4172
		验证	15.6521
	梅尔频谱	训练	3.9596
		验证	20.7467
卷积神经网络		训练	4.1795
		验证	21.3639

分析表 2 中的数据可知，无论执行训练还是验证操作，全连网络与 MFCC 的组合总可以取得最快的运算速度。因此，本文选择全连网络与 MFCC 的组合作为声音识别系统算法的核心。

用于训练的音频总是以 30s 为一段，每次训练前录入一段音频并求取梅尔频谱。这里使用的音频采样率为单声道 44100Hz，预加重参数 μ 取值为 0.97，快速傅里叶变换的窗参数 a 取值为 0.46，长度取值为 1024，窗间隔取值为 512，即保证 50%的重叠率，梅尔滤波器的个数为 128。

将频谱沿时域切割，间隔约 22.7ms 切一段，得到一个大小为 1×128 的矩阵，这就是一条训练数据，连续切割直到该段音频结束。切割样本除了在训练时可以扩充样本数量外，小样本网络可以提高系统工作时的灵敏程度并减小系统的偶然误差。

全连神经网络用于分类，它的输入就是切割后大小为 1×128 的矩阵，中间共串联 5 个 128×128 的全连接层，任意两层之间使用了 ReLU 激活函数。输出层接 $1\times N$ 的 softmax 分类器，其中 N 是分类的总数，N 增大后，系统能够识别的背景音变多。学习率取 0.001，使用 AdamOptimizer 优化器进行优化。

由于滤波环节的存在，神经网络的分类数并不是越多越好，因为许多日常声音都是低频噪声，例如绝大多数的乐器的发声频率都在 1000Hz 以下，对这种声音滤波之后求取声谱，得到的结果全都是零，而神经网络无法对两个完全相同的信号进行分类，导致作出错误的判断。在这种情况下，分类数过多反而会抑制系统的识别能力，因此需要根据目标信号的特点合理选择分类数。图 6 为系统工作流程图，无论数据是否用于训练，首先都要对数据进行滤波等前期处理。训练数据可以更新系数矩阵，测试数据则直接调出该系数矩阵来判断周围是否存在无人机声。

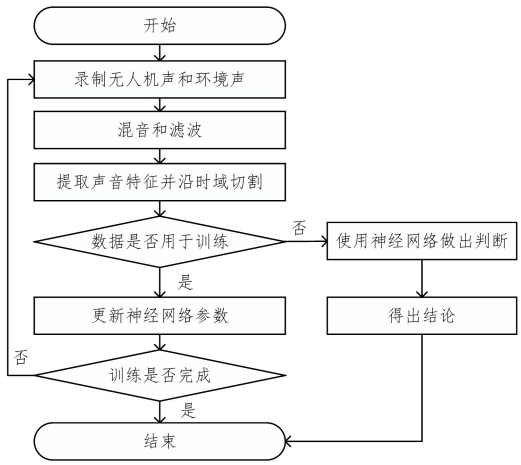


图 6 无人机识别系统工作的流程图

Fig. 6 Flow chart of UAV identification system

4.2 训练样本信噪比的分析

影响系统识别性能的最主要的因素是训练样本的质量。为了在不同环境下均可识别出无人机，我们首先需要调节样本声音信号的信噪比，寻找使系统获得最优性能的参数。可以通过混音并降低样本中无人机声强的方式，来获取多种信噪比的声音信号。在该实验中采样获得的声音信号过滤 5000Hz 以下和 15000Hz 以上的部分，滤波方法为对音频信号执行快速傅里叶变换，将滤波区间的频域信号全部改为 0，再使用傅里叶反变换将其写为新的音频文件。

实验中共合成了 3 种信噪比的音频数据，分别为：1)无人机声不减弱(数据 A)；2)无人机声减弱 15dB(数据 B)；3)无人机声减弱 30dB(数据 C)。除这 3 组数据外，从数据 A 和数据 B 各取一半组成数据 D，并使用 4.1 节提出的全连结构分别训练得到 4 个网络，最后对这些网络逐一进行实地测试。

我们使用混淆矩阵来评价无人机侦察系统的性能，若规定含有无人机的声音信号为正，不含无人机的声音信号为负，则矩阵中使用的参数如下：正阳性(TP)表示实际为正且预测为正；正阴性(TN)表示实际为负且预测为负；负阳性(FP)表示实际为负但预测为正；负阴性(FN)表示实际为正但预测为

负。将实验数据的灵敏度(sensitivity)、精确率(precision)、特异性(specificity)、正确率(accuracy)和F_1分数(F_1-score)作为实验效果的评判标准，详细情况请参考文献[21]。

实验结果如表3所列。在保证精确度和灵敏度均大于80%的情况下，系统所能识别的最远距离定义为识别距离，超过这个距离则认为系统不具备识别能力。在最远距离的一半处重新测得4项混淆矩阵值填入表中，用于表征系统的平均稳定性。通过观察，我们可以发现两个现象。

(1)当训练样本中无人机声音强度降低15dB时，侦测系统的识别距离最远；当训练样本中同时含有降低15dB和不降低声强的无人机声时，识别距离变近了，但是4项检测指标和最终的F1分数均得到了有效改善；当训练样本中只含有不降低声强的无人机声时，识别距离进一步缩短，且检测指标和F1分数并无全面提升。

(2)含有降低30 dB无人机声强的样本相应的网络指标在50%左右浮动，这个百分比和掷硬币判断的概率相似。无论无人机与话筒的距离远近，实验结果均不理想，因此这种网络不具备侦测能力。

表3　训练样本声音强度对系统识别性能的影响

Table 3　Influence of sound intensity of training samples on system recognition performance

实验地点	数据A						数据B					
	最大距离/m	准确率/%	精确度/%	灵敏度/%	特异度/%	F1分数	最大距离/m	准确率/%	精确度/%	灵敏度/%	特异度/%	F1分数
室内	8.6	96.1	93.2	100.0	92.6	0.964803	9.8	82.6	84.4	90.9	65.8	0.875295
夜晚的街道	6.1	95.2	94.0	97.4	93.0	0.956698	8.0	97.6	95.5	100.0	95.1	0.976880
喧闹的街道	4.3	87.0	90.3	89.7	81.8	0.900104	5.1	87.7	90.5	90.8	81.5	0.906498
隧道内	7.7	90.2	87.6	92.9	88.2	0.901722	9.2	89.0	79.5	81.4	92.3	0.803478
实验地点	数据C						数据D					
	最大距离/m	准确率/%	精确度/%	灵敏度/%	特异度/%	F1分数	最大距离/m	准确率/%	精确度/%	灵敏度/%	特异度/%	F1分数
室内	小于1	61.5	62.7	55.6	70.6	0.589369	8.7	92.5	97.0	92.4	92.6	0.946441
夜晚的街道	小于1	45.8	56.8	45.6	45.9	0.505875	7.1	92.8	95.7	93.0	87.2	0.943307
喧闹的街道	小于1	46.7	58.6	42.1	48.2	0.489982	3.4	86.3	82.3	90.2	80.4	0.860691
隧道内	小于1	63.3	70.2	54.2	68.2	0.611711	7.0	92.7	92.4	91.0	94.0	0.916947

4.3　滤波器参数对系统的影响

噪声在低频段对无人机声产生干扰，过滤低频信号有助于系统减少不必要的混淆，但为了防止重要的信号被当作噪声消除，也不宜滤波过多。前文所述实验的滤波下限暂定为5000 Hz，但这不一定是最佳参数，寻找滤波的安全频率下限仍十分必要。

使用MFCC和全连神经网络分析音频，这里令MFCC返回13个参数，其中DCT变换选择方式2，详细信息请见文献[22]，其他参数与前述梅尔频谱完全一致。实验场所选择室内长廊和室外空旷的广场，精确度和灵敏度仍在最大识别距离的一半处测得，实验结果如表4所列。可以看出，为使系统获得较好的性能，滤波下限则不应低于6000 Hz。为了尽可能减小噪声对识别性能的影响，后续实验均取滤波下限为6000 Hz。

表4　滤波下限对系统性能的影响

Table 4　Influence of filter lower limit on system performance

滤波下限/Hz	训练次数/次	室外最大识别距离/m	室外精确度/%	室外灵敏度/%	室内最大识别距离/m	室内精确度/%	室内灵敏度/%
0	2000	小于1	45.9	50.5	小于1	62.1	60.8
	6000	小于1	49.7	58.2	小于1	51.6	70.0
2000	2000	1.6	85.3	81.4	10.4	95.8	87.1
	6000	10.0	91.6	93.3	13.6	92.3	94.1
4000	2000	6.4	86.7	86.8	13.6	89.4	86.9
	6000	9.6	90.0	91.1	16.1	95.0	91.3
6000	2000	11.7	89.9	87.0	33.6	83.6	90.4
	6000	9.0	93.4	89.3	19.2	92.8	95.6
8000	2000	10.5	82.3	88.5	18.4	92.6	93.7
	6000	8.8	89.2	80.1	18.7	91.1	95.8
10000	2000	10.7	81.4	85.7	7.1	82.3	87.1
	6000	9.1	93.6	84.5	17.2	90.6	95.5

4.4　完整系统训练次数分析与性能对比

本文选定下限为6000 Hz、上限为15000 Hz的带通滤波器对声音信号做预处理，MFCC和全连神经网络构建分析音频网络。

对比实验中使用的卷积神经网络参考了文献[23]中使用的曾获得2012年ImageNet大赛冠军的AlexNet网络。但是，考虑到声音信号的特殊性，对结构和超参数做了删改，该网络可以用于二次提取梅尔频谱的特征。

原始数据仍然是大小为1×128的矩阵，在输入层被重塑为1×16×8，卷积核边长为5；得到的卷积层1为32×8×4，卷积核边长不变，后接一个2×2的最大池化层和ReLU激活函数；这样得到的卷积层2为64×4×2，之后不再卷积，只进行同样的最大池化操作和ReLU非线性处理；后面是连续3个全连接层，神经元结构分别为512×512，512×512和512×5；

最后接 softmax 分类器并将声音信号分为 5 类。全过程的学习率均为 0.001，优化器选择 AdamOptimizer。图 7 直观地给出了 CNN 的结构。

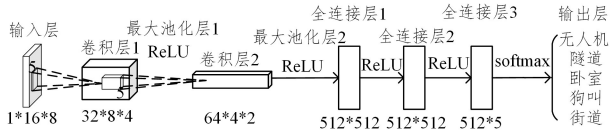


图 7 对比实验所用卷积神经网络的结构图

Fig. 7 Structure diagram of CNN used in comparative experiment

全连网络结构与 4.2 节探究最佳信噪比实验所用网络保持一致。考虑到重复训练会导致过拟合问题，需要固定学习率并在网络中限制训练次数，对训练次数不同的网络进行性能对比即可找到过拟合拐点。

表 5 列出了 3 种神经网络的识别性能随训练次数增加的变化规律。可以看出，无论在室内还是室外，使用 MFCC 提取声音特征并配合全连神经网络总能获得最远的识别距离，室内的最大距离为 33.6 m，室外为 14.3 m。且这种组合获得最大距离所需的训练次数最少，实验中只要 500～2 000 次，增加训练次数反而会减小识别距离，这就是前述过拟合现象所带来的负面影响。另外，两种神经网络的组合性能相近，在训练次数较少时稳定性差，从而不能完成识别任务，只有在超过 10 000 次训练后才能表现出识别能力，且识别距离较近。

表 5 不同网络组合对系统性能的影响

Table 5 Influence of different network combinations on system performance

神经网络类型	音频特征提取方式	训练次数/次	室外最大识别距离/m	室外精确度/%	室外灵敏度/%	室内最大识别距离/m	室内精确度/%	室内灵敏度/%
全连神经网络	梅尔频谱	100	小于 1	51.3	55.8	小于 1	47.6	50.9
		500	小于 1	58.9	62.2	小于 1	58.3	61.1
		2 000	小于 1	62.3	70.1	小于 1	61.9	68.6
		6 000	小于 1	63.7	77.3	小于 1	77.2	78.5
		9 000	小于 1	77.4	88.9	1.3	81.4	92.7
		12 000	2.2	88.6	92.3	4.7	91.4	97.0
		15 000	3.1	91.3	94.3	4.4	95.8	92.9
		20 000	2.5	95.0	92.4	2.5	96.3	99.7
	MFCC	100	5.3	82.1	75.8	8.0	80.4	81.1
		500	14.3	87.3	82.8	27.2	91.3	89.5
		2 000	11.7	89.9	87.0	33.6	83.6	90.4
		6 000	9.0	93.4	89.3	19.2	92.8	95.6
		9 000	7.3	92.9	90.9	20.0	90.8	93.1
		12 000	7.1	95.6	94.0	16.3	98.1	96.5
		15 000	6.8	98.1	100.0	18.0	97.1	98.3
		20 000	2.5	99.8	96.4	11.2	98.7	97.3
卷积神经网络	梅尔频谱	100	小于 1	47.3	52.1	小于 1	43.6	61.0
		500	1.1	79.4	85.3	2.0	90.1	88.2
		2 000	小于 1	67.2	65.9	小于 1	63.1	70.3
		6 000	小于 1	81.4	68.9	小于 1	73.1	75.8
		9 000	小于 1	70.9	67.3	1.0	80.8	81.4
		12 000	小于 1	77.3	70.0	5.0	92.6	90.3
		15 000	4.8	91.9	94.3	10.4	95.7	97.2
		20 000	3.2	99.8	99.3	8.0	100.0	97.5

以室外为例，图 8 和图 9 清晰地给出了系统的精确度和灵敏度随拟合程度变化的趋势。从图中可以清楚地看出，随着训练次数的增多，系统精确度和灵敏度都明显提高，但是不同网络组合的拟合速度仍有较大差别。当进行 500 次训练时，只有全连网络和 MFCC 的组合可以达到约 80%的精确度，另外两种组合的精确度只有 50%；训练次数达到 20 000 次时，3 种系统在识别距离内均几乎不再出错。

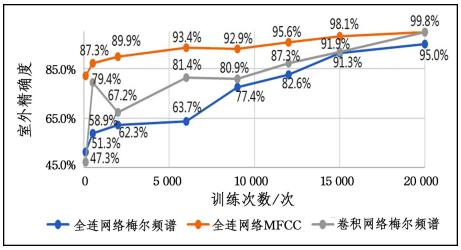


图 8 3 种网络组合在室外侦测的精确度

Fig. 8 Accuracy of three network combinations for outdoor detection

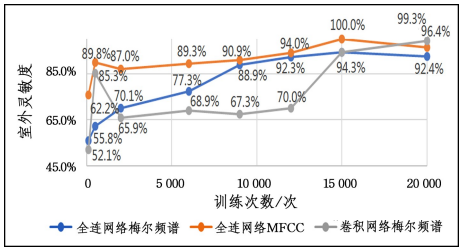


图 9 3 种网络组合在室外侦测的灵敏度

Fig. 9 Sensitivity of three network combinations for outdoor detection

由上文的分析可知，增大训练次数会导致过拟合现象，但也有利有弊，在减小识别距离的同时也可以减少在该区域内的误判。综合识别距离和误判概率两个指标，对于 MFCC 全连网络来说，本实验训练 2 000～6 000 次可以获得比较理想的效果；而对于另两种组合，则需要训练 10 000～15 000 次。

4.5 跨机型测试分析

为了更好地表征该识别算法的通用性，表 6 列出了系统

跨型号识别无人机的能力。用于训练的样本型号除 TLW004 外，还包括 Matrix100、Phantom3 和御 MavicPro，而测试时只使用 TLW004 型无人机。基于上述实验对神经网络参数的探究，在跨型号识别实验中滤波下限取 6000 Hz，拟合程度取 6000 次训练，并仍然使用全连网络与 MFCC 的组合进行实验，其他参数均不变。

表 6　跨型号识别无人机实验

Table 6　Cross-model identification experiment

无人机样本型号	Matrix100	Phantom3	御 MavicPro	TLW004
最大距离/m	6.3	6.1	5.8	9.0
精确度/%	82.7	78.3	80.8	93.4
灵敏度/%	88.1	81.4	83.7	89.3

从表 6 可以清晰地看出，在不跨型号的情况下，识别系统取得了 9.0 m 的最大识别距离，并且精确度和灵敏度均约为 90%；若进行跨型号识别，则识别距离略有减小，但仍具备一定的识别能力。

从声音识别的角度考虑，不同无人机的区别主要体现于电动机的设计。在绝大多数无人机均使用无刷电机且配用直桨的情况下，跨型号识别无人机并不困难。

4.6　实验总结

根据以上实验结果，该系统的最佳参数和性能如表 7 所列。

表 7　实验总结

Table 7　Experiment summaries

研究项目	最佳参数和性能
声音提取方式	MFCC
神经网络	全连神经网络
带通滤波器	6000～15000 Hz
优化器	AdamOptimizer
学习率	0.001
拟合程度	500～6000 次训练
室内最大距离/m	33.6(精确度为 83.6%)
室外最大距离/m	14.3(精确度为 87.3%)
最佳信噪比/dB	无人机声强降低 15
跨型号识别能力	支持跨型号识别

我们可以得出如下结论：

(1)适当降低训练样本中目标声音的强度可以有效地模拟远距离侦测的情况，但降低过多可能导致系统不能正确区分噪声和目标，失去实用价值。

(2)相比另外两种神经网络组合，MFCC 提取声音特征并经全连神经网络进行分类，可以使识别系统获得最优性能和最快的训练速度。

(3)为了防止自然噪声对识别功能造成影响，对于采集的音频信号应进行带通滤波，滤波下限与系统拟合程度有关，以 6000 Hz 为优。

(4)神经网络的拟合程度影响系统的识别距离和误判率，拟合程度越高，识别距离越近，且在该范围内误判率越低。综合考虑多种情况，本实验在取学习率为 0.001 的情况下，训练次数应在 500～6000 次之间。

结束语　本文设计了一种基于深度学习、通过捕捉螺旋桨声音来探测周围环境是否存在无人机的识别系统。主要贡献如下：

(1)提出使用全连神经网络配合梅尔倒谱系数的方法来提取无人机特征并进行分类。

(2)探究对训练样本进行滤波、切割、降低声音强度等预处理的最佳参数，使系统的识别距离更远，鲁棒性更强。

同时，该无人机识别系统在应用时仍有一些需要解决的问题。

(1)神经网络的超参数、训练样本的最佳信噪比应当在后续实验中不断完善和改进，以使系统获得更优的侦测性能。

(2)由于音频信号的特殊性，为了防止滤波后背景音重复，分类的数目不能过多也不能过少。未来应当根据不同声音的频率谱继续探究背景音种类对系统识别性能的影响。

参 考 文 献

[1] DARIO F, ROBERT J. Science, Technology and the Future of Small Autonomous Drones[J]. Nature, 2015, 521(7553): 460-466.

[2] HASSANALIAN M, ABDELKEFI A. Classifications, Applications, and Design Challenges of Drones: A Review[J]. Progress in Aerospace Sciences, 2017, 91: 99-131.

[3] COLOMINA I, MOLINA P. Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review[J]. Isprs Journal of Photogrammetry and Remote Sensing, 2014, 92: 79-97.

[4] MILAN E, ENRICO N, KAUSHIK R, et al. Help from the Sky: Leveraging UAVs for Disaster Management[J]. IEEE Pervasive Computing, 2017, 16(1): 24-32.

[5] HUTTUNEN M. Civil Unmanned Aircraft Systems and Security: The European Approach[J]. Journal of Transportation Security, 2019, 12(3): 83-101.

[6] SHI X F, YANG C Q, XIE W G, et al. Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges[J]. IEEE Communications Magazine, 2018, 9(1): 68-74.

[7] LI Q Z, XIONG R R, WANG R P, et al. Research on real-time UAV recognition method based on SSD Algorithm[J]. Ship Electronic Engineering, 2019, 39(5): 30-35.

[8] ANDREA B, FEDERICA M, EMILIANO P, et al. Drone detection by acoustic signature identification[J]. Electronic Imaging, 2017, 10(1): 60-64.

[9] HUANG G B, ZHOU H M, DING X J, et al. Exterme Learning Machine for Regression and Multiclass Classification[J]. System Man and Cybernetics, 2012, 42(2): 513-529.

[10] KAROL J P. Environmental Sound Classification with Convolutional Neural Networks[C]// 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing(MLSP). 2015: 1-6.

[11] KHAN S A, ANIL S T, JAGANNATH H N, et al. A Unique Approach in Text Independent Speaker Recognition Using MFCC Feature Sets and Probabilistic Neural Network[C]// 2015 EighthInternational Conference on Advances in Pattern Recognition(ICAPR). 2015: 1-6.

[12] WANG R P, FANG Y, CHEN D L, et al. Vehicle abnormal

sound Recognition based on wavelet packet FBank Spectrogram and CNN[J]. Journal of Congqing Uinversity of Technology (Natural Science), 2020, 34(7): 1-9.

[13] HARSHITA G, DIVYA G. LPC and LPCC Method of Feature Extracion in Speech Recognition System[C]// 2015 International Conference-Cloud System and Big Data Engineering. 2015: 498-502.

[14] MUHAMMAD Z A, ZEESHAN K, ABBAS J. Machine Learning Inspired Sound-Based Amateur Drone Detection for Public Safety Applications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(3): 2526-2534.

[15] SAHIDULLAH M, GOUTAM S. Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition[J]. Speech Communication, 2012, 54(4): 543-565.

[16] CHEN T, AO M Y, CHEN H Z. Study of speech recognition technology based on MFCC and SVM[J]. Journal of Guangxi Vocational and Technical College, 2010, 3(5): 1-4.

[17] WASEEM R, WANG Z H. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review[J]. Neural Computation, 2017, 29(9): 2352-2449.

[18] MATTHIAS H, MAKSYM A, JULIAN B. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 41-50.

[19] DARCH J, MILNER B, VASEGHI S. MAP prediction of formant frequencies and voicing class from MFCC vectors in noise [J]. Speech Communication, 2006, 6(48): 1556-1572.

[20] HE K M, SUN J. Convolutional neural networks at constrained time cost[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 5353-5360.

[21] DENG X Y, LIU Q, DENG Y, et al. An Improved Method to Construct Basic Probability Assignment Based on the Confusion Matrix for Classification Problem[J]. Information Sciences, 2016, 340: 250-261.

[22] SIRANI M P, LIU J H. Complexity Reduction, Self/Complete Recursive, Radix-2 DCT I/IV Algorithms[J]. Journal of Computational and Applied Mathematics, 2020, 379: 1-16.

[23] GU S S, DING L, YANG Y, et al. A New Deep Learning Method Based on AlexNet Model and SSD Model for Tennis Ball Recognition[C]// 2017 IEEE 10th International Workshop on Computational Intelligence and Applications (IWCIA). 2017: 159-164.

XU Hao, born in 1998, undergraduate. His main research interests include deep learning and industrial automation.

LIU Yue-lei, born in 1986, Ph.D, lecturer. His main research interests include intelligent optimization algorithm and deep learning.