小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

2018 年 9 月 第 9 期
Vol. 39 No. 9 2018

深度卷积神经网络在音乐风格识别中的应用

胡昭华[1,2], 余媛媛[1]

[1](南京信息工程大学 电子与信息工程学院,南京 210044)
[2](南京信息工程大学 江苏省大气环境与装备技术协同创新中心,南京 210044)
E-mail: zhaohua_hu@163.com

摘 要: 目前大多基于时间特征的音乐风格识别问题分类性能不佳. 鉴于卷积神经网络(CNNs)捕获信息特征能力较强,本文使用 CNN 提取音乐信号中的多种特征并进行分类. 首先采用 harmonic/percussive sound separation(HPSS)算法把原始音乐信号谱图分离成时间特征谐波分量和频率特征冲击分量,并联合原始谱图一起作为 CNN 的输入;其次对生成图像作仿射变换以及使用 PCA 改变训练图像中 RGB 通道的像素值从而扩大数据集;最后设计了 CNN 的网络结构以及研究了该网络结构中不同参数对识别率的影响. 在 GTZAN 数据集上的实验表明本文的方法可以有效改善使用单一特征的音乐风格识别.

关 键 词: 音乐风格识别;卷积神经网络;HPSS;数据增强
中图分类号: TP391 **文献标识码**: A **文 章 编 号**: 1000-1220(2018)09-1932-05

Musical Genre Recognition with Deep Convolutional Neural Networks

HU Zhao-hua[1,2], YU Yuan-yuan[1]

[1](School of Electronic & Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing 210044, China)

Abstract: At present, most of the music genre recognition problems based on the temporal features have poor performance. Considering that CNN has strong capacity to capture informative features, this paper uses CNN to extract multiple characteristics of music signals and classify music genres. First, the harmonic/percussion sound separation(HPSS) algorithm is used to separate the spectrums of original music signals into harmonic components with distinct temporal characteristics and percussive components with frequency characteristics, which input into CNN combined with original spectrograms. Then, the affine transformation of generating images and performing PCA to alter the pixel values of RGB channels in training images are used to augment data. Last, we design the structure of CNN and study the impacts of different parameters on the recognition rate. Experiments on the GTZAN dataset show that our method can effectively improve the music genre recognition.

Key words: musical genre recognition; convolutional neural network; HPSS; data augmentation

1 引 言

音乐风格分类是音乐信息检索 MIR(Music Information Retrieval)[1]领域非常具有挑战性但又很有前景的任务. 由于音乐是一种不断发展的艺术且音乐风格之间没有明确的界限,自动分类音乐风格是一个具有挑战的问题. 音乐风格分类问题的关键是音乐信息的特征提取. 多种特征提取和分类方法在最近几年相继被提出[2,3],这些分类器的性能高度依赖于按经验所选的手动提取特征的适当性. 一般地,识别任务中的特征提取和分类是两个独立的处理阶段,而本文把这两个阶段融合在一起更好地实现了信息间的交互.

最近,深度卷积神经网络 CNN(Convolutional Neural Networks)[4]在通用视觉识别任务[5,6]上不断取得显著进步,这激起了人们对于 CNN 的分类模型[7,8]的研究兴趣. CNN 包括了多级处理输入图像,提取多层和高级特征表示. 通过共享一些基本的组成部分,可以把许多手动提取的特征和相应的分类方法看作是一个近似或特殊的 CNN,然而为了保留有判别力的信息,这些特征和方法必须仔细设计和整合. 受 CNN 在通用视觉识别任务上显著成功的激励,本文将 CNN 用于具有挑战的音乐风格识别,并研究了网络结构参数的调整对识别率的影响.

Lee[9]是第一个把深度学习应用于音乐内容分析的人,特别是风格和艺术家识别,通过训练一个有 2 层隐层的卷积深度置信网络 CDBN(Convolutional Deep Belief Network)以一种无监督的方式尝试使隐层激活,产生来自预处理频谱的有意义的特征. 相比较于那些标准的 MFCCs(Mel Frequency Cepstral Coefficients)特征,其深度学习特征有更高的精准度. 对于音乐风格识别,Li[10]等人将音乐数据转化为 MFCC 特征向量输入有 3 个隐层的卷积神经网络(CNN),最终得出可用 CNN 自动提取图像特征用于分类的结论,表明 CNN 具有较强的捕获变化的图像信息特征能力.

本文主要进行了以下几个方面的工作:(1)使用 HPSS 算

收稿日期: 2017-10-09 收修改稿日期: 2018-03-05 基金项目: 国家自然科学青年基金项目(61601230)资助;江苏省自然科学基金青年基金项目(BK20141004)资助;江苏高校优势学科Ⅱ期建设工程项目资助. 作者简介: **胡昭华**,女,1981 年生,博士,副教授,研究方向为视觉跟踪、模式识别; **余媛媛**,女,1992 年生,硕士研究生,研究方向为计算机视觉、模式识别.

法分别从时间和频率方面提取了音乐信号谱图的谐波分量和冲击分量,并将其和原始谱图一起作为 CNN 网络的输入; (2) 详细设计了用于音乐风格分类的基于 CNN 的深度分类框架,对有效训练一个可靠的 CNN 所需的多个关键因素进行了研究和实验证明. (3) 本文使用对频谱图像进行仿射以及使用 PCA 改变训练图像中 RGB 通道的像素值的方法扩充数据集.

2　音乐风格识别算法

本文基于卷积神经网络的音乐风格识别的总体框架如图 1 所示. 首先使用 HPSS 算法对音乐曲目进行分离,把原始曲目分离成谐波音源和冲击音源; 然后对这两种音源及原始曲目分别作短时傅立叶变换,将变换后的谱图输入 CNN 网络中进行学习,训练以及预测,最后的输出结果即为最终的识别率.

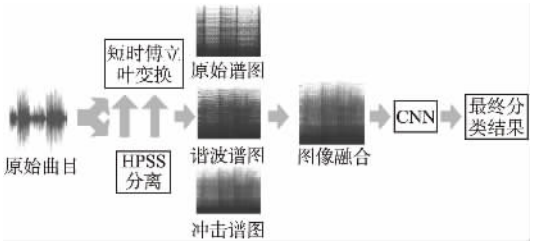

图 1　音乐风格识别的流程图

Fig. 1　Architecture for music genre classification

2.1　Harmonic/Percussive 分离算法

音乐信号通常由谐波声音成分和冲击声音成分组成,其具有非常不同的特性. 本文使用了分离音乐信号的谐波和冲击声音成分的 Harmonics /Percussion 分离算法,其本质是基于频谱图的各向异性连续性的信号分离. 这个方法的关键在于其侧重于谐波频谱和冲击频谱的连续方向的差异. 谐波频谱通常在时间方向上是连续的,冲击频谱在频率上是连续的. 图 2 是某个音乐曲目的原始频谱及分离后得到的谐波频谱和冲击频谱. 从图中可以看出分离后的谐波频谱是在某个固定的频率上沿时间轴连续平滑分布,而冲击频谱是在时间轴上很短而沿频率轴连续平滑分布,而原始频谱中包含有纵向冲击声音与横向的谐波声音.

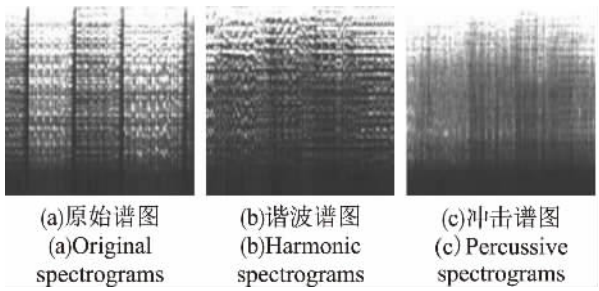

图 2　HPSS 算法后的不同谱图

Fig. 2　Different spectrograms with HPSS algorithm

2.2　网络结构

CNN 的网络结构的前几层作为特征提取器通过监督训练自动获取图像特征,在最后一层通过 softmax 函数进行分类识别.

本文的 CNN 网络结构如图 3 所示,其与传统的 AlexNet 共享基础框架. 具体来说,它包含八层,前五层是与 pooling 层交替的卷积层,剩下三层是用于分类的全连接层. CNN 网络的输入图像是使用 HPSS 分离的谐波谱图和冲击谱图以及原始音乐信号的谱图,并将输入图像大小归一化为 256* 256,然后将其输入第一个卷积滤波器. 在深层网络结构中,第一个卷积层利用 96 个大小为 11* 11、步长为 4 个像素(这是同一核映射中邻近神经元的 receptive field 中心之间的距离) 的核对输入图像进行滤波. 接下来 max pooling 层将第一个卷积层的输出作为输入并和 96 个大小为 3* 3 的核进行滤波,响应归一化后,第二个卷积层和其输出相连接并用 256 个大小为 5* 5 的核对其进行滤波. 第三、第四和第五个卷积层相互连接,没有任何介于中间的 pooling 或归一化层,第三个卷积层有 384 个大小为 3* 3 的核被连接到第二个卷积层的(归一化的、pooling 的) 输出. 第四个卷积层拥有 384 个大小为 3* 3 的核,第五个卷积层拥有 256 个大小为 3* 3 的核. 使用这五个卷积层,最终获得了 256 个大小为 6* 6 的特征图,这些特征图被馈送到分别三个含有 4096,1000 和 10 个神经元的全连接层. 最后一个全连接层的输出便是最终的识别结果.

2.3　网络训练和学习方法

本文的网络结构是一个分层的深度卷积神经网络,其通过卷积输入图像和一组核滤波器提取局部特征,卷积层通过线性卷积滤波器以及非线性激活函数(ReLU) 生成特征图. 同一层中神经元的输出形成一个平面,称之为特征图,然后通过 pooling 获得卷积特征图并将其滤波到下一层. 在 local receptive field 通过设置不同的核滤波器来获取不同的特征图. 给定 X_l^p 表示在第 l 层的第 p 个特征图,对整张特征图所进行的卷积以及使用的激活函数如公式(1) 所示:

$$X_l^q = \max\left(0, \sum_{X^p \in M_q} X_{l-1}^p \otimes k_l^{pq} + b_l^q\right) \tag{1}$$

其中 X_l^q 为第 l 层第 q 个卷积核输出的特征图,$\otimes$代表卷积运算,k_l^{pq} 为卷积核,M_q 表示特征图 X_{l-1} 的集合,$\max(\cdot)$ 是非线性激活函数 ReLU,b_l^q 是偏置,特征图 X_{l-1}^p 在卷积操作后使用激活函数. 由于发现局部响应归一化有助于网络的一般化,因此在本网络模型的某些层中在 ReLU 后进行归一化. 这种响应归一化实现了一种真正的神经元中发现的侧向抑制的形式,这种侧向抑制的效果就是使不同卷积核计算的神经元输出值之间对计算值比较大的神经元活动更为敏感.

Pooling 层使用的是 Max Pooling,在卷积神经网络中卷积层和 pooling 层都是交替出现的.

输出层与上一层完全连接,其产生的特征向量可以被送到逻辑回归层完成识别任务,所有网络中的权重使用反向传播算法[11]学习.

本文使用随机梯度下降 SGD(stochastic gradient descent) 训练网络模型,发现较小的权重衰减对模型的学习非常重要,权重衰减可以减少模型的训练误差,因此在本文的实验中微调到 0.0005. dropout 是训练神经网络过程中一种防止过拟合的技术,通常 dropout 和动量可以改善学习效果[12]. 由于所有层使用 dropout 会使网络收敛花费很长时间,因此在本文的实验中在全连接层设置 dropout 值为 0.5,α=0.9,λ=0.0005.

本文的网络结构中总共有三个全连接层,最后一个全连接层即第八层是输出层,第七层的输出即为其输入,其包含对应 m 类音乐风格的 m 个神经元,输出概率为 $p=[p_1,p_2,\cdots,p_m]^T$,使用 softmax 回归公式如下:

$$p_j=\frac{\exp(X_8^j)}{\sum_{i=1}^{m}\exp(X_8^i)} \tag{2}$$

其中 X_8 是 softmax 函数的输入,j 是被计算的当前类别,$j=1,\cdots,m$;p_j 表示第 j 类的真实输出.

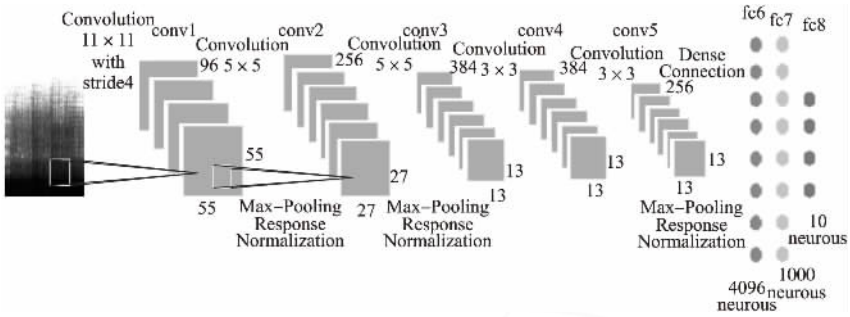

图 3 音乐风格识别的 CNN 结构图

Fig. 3 Architecture of our deep convolutional neural network for music genre classification

3 实验结果及分析

本文实验中利用 caffe 框架来训练 CNN 模型以实现音乐风格的识别.

3.1 数据集

使用识别率作为性能指标,在众所周知的 GTZAN 风格收集数据库[13] 上评估了所提出的方法. GTZAN 数据集由 Tzanetakis 和 Cook(2002) 收集,由 10 种类型(蓝调,古典,乡村,迪斯科,嘻哈,爵士,金属,流行,雷鬼和摇滚) 组成. 每个风格类别包含 100 个音频录音,长达 30 秒,共有 1000 个音乐节选.

3.2 参数优化实验

本文实验表明了正确调整超参数的重要性,超参数可分为两种: 模型相关的超参数和训练相关的超参数,如 2.2 节所示和表 1 所示.

为了调整这些超参数,数据集按 5∶1 的比例被随机分为二个子集,即 2500 个音乐曲目用于训练,500 个音乐曲目用于测试.

表 1 训练相关的超参数

Table 1 Training-relevant hyper-parameters obtained

超参数	学习率 η	Batch-size	动量系数 μ	权值衰减系数 λ	Dropout 系数
值	0.01	16	0.9	0.0005	0.5

根据 Bengio[14] 所述,参数要调整到训练集的错误率变得足够小且稳定的时候. 通过该调整过程获得的超参数总结在表 1 中.

训练相关的超参数可以显著影响网络的收敛和学习速率,它们的影响通过识别率曲线说明,如图 4-图 7 所示. 在每张图中,我们集中于一个超参数,而其他的则设置为表 1 中的最佳值.

图 4 表示对训练样本进行 20000 次迭代,学习速率 η 比较小如 0.001 时,学习过程会非常缓慢,识别率也尚不稳定. 适当提高 η 可以有效提高学习效率. 同时若 η 过大如 0.1 会

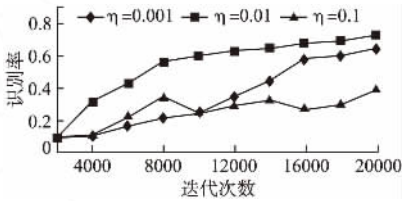

图 4 学习率 η 的影响

Fig. 4 Impact of learning rate

导致学习过程不稳定并且降低分类性能. 图 5 和图 6 分别说明了动量 μ 和权重衰减 λ 的影响. 图5表明使用动量 μ 可以

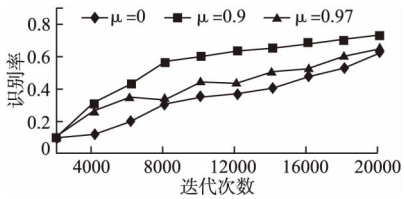

图 5 动量系数 μ 的影响

Fig. 5 Impact of momentum

很好地加快学习过程,同时,若 μ 偏大如 0.97,会在初期阶段

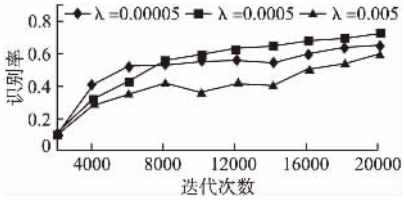

图 6 衰减系数 λ 的影响

Fig. 6 Impact of weight decay

引起振荡,收敛较慢,此外,它降低了后期阶段的分类性能. 图 6 说明了权重衰减 λ 的影响,表明较小的 λ 似乎是一个更安

全的选择,而较大的 λ 如 0.005 会破坏学习过程的稳定性.

Dropout 是训练神经网络过程中防止过拟合的一种技术. 在文献[15]中,减少 70%输出的 dropout 应用于最后一个全连接层. 本文的实验采取此技术并在每个回合中弄乱训练数据来减轻过拟合. 在这种类型的研究中,通常以一定的概率将隐层神经元的输出设置为 0,这样此类神经元对正向传播和反向传播都将不起任何作用,因此每输入一个样本,其都使用了不同的网络结构但权值又是共享的,这样求得的参数就能适应不同情况下的网络结构,提高了网络的泛化能力. 本实验将 dropout 微调为 0.5 或 0.6,当增大 dropout 值时,训练时间稍长一些,收敛较慢,该训练进行了 20000 次迭代. 基于 CNN 的分类器经过 20000 次迭代后可以产生良好的分类性能. 图 7 显示了不同的 dropout 值经过不同的迭代次数有不同的识别率.

总之,CNN 中的超参数: 学习速率 η,动量系数 μ,权重衰减系数 λ 和 dropout 值可以显著影响网络训练过程,在获得满意的分类性能之前必须仔细调整. 在本文的实验中,使用表1中设置的超参数,在数据没扩充的情况下,GTZAN数据集的识别率为 73%左右.

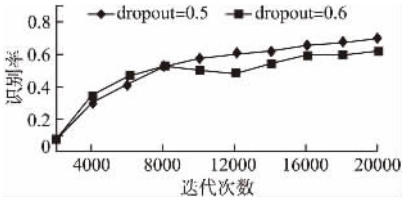

图 7　dropout 值的影响

Fig. 7　Impact of dropout

3.3　不同音乐风格的识别率比较

表 2 以混淆矩阵的形式给出了关于音乐风格分类更详细的信息,其中列对应实际的风格,行对应预测的类别,正确分类的百分比位于矩阵的对角线. 由于有些音乐风格之间的界限不分明,容易产生误判,比如有的 classical 音乐的节奏比较强烈,容易被误认为是 jazz 音乐; 而 rock 音乐由于其广泛的特性容易被误认为其他风格,所以其分类精度相比于其他风格要低一些.

表 2　GTZAN 数据集的混淆矩阵

Table 2　Confusion matrix for GTZAN dataset

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	82.5	8.3	0.0	0.0	0.0	7.3	0.0	0.0	4.7	0.0
classical	10.2	74.6	0.0	0.0	0.0	12.5	0.0	0.0	0.0	0.0
country	0.0	4.5	91.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
disco	4.0	4.1	4.3	79.1	0.0	0.0	0.0	8.3	12.0	16.7
hiphop	0.0	0.0	0.0	0.0	75.6	0.0	0.0	0.0	8.1	0.0
jazz	3.3	8.2	0.0	0.0	0.0	76.2	0.0	0.0	4.5	0.0
metal	0.0	0.0	0.0	0.0	20.2	0.0	92.0	0.0	4.4	0.0
pop	0.0	0.0	0.0	12.0	0.0	4.0	0.0	82.7	0.0	16.0
reggae	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	64.3	8.4
rock	0.0	0.0	4.3	8.9	4.2	0.0	8.0	8.1	2.0	58.9

3.4　不同特征图实验比较

表 3 表明了手工提取时间序列和频率序列的特征并以不同的组合方式放入 CNN 网络训练,最后会得到不同的效果,其中把 3 种特征图都输入网络时得到的识别率是最高的,说明只有当训练的特征更全面时才能得到更好的结果.

表 3　不同谱图的识别率

Table 3　Classification rate for different spectrograms

图片类型	识别率
Original Spectrograms	67%
Harmonic Spectrograms	58%
Percussive Spectrograms	60%
Original + Harmonic + Percussive Spectrograms	73%

3.5　数据扩充实验对比

本文主要使用仿射变换以及用 PCA 改变训练图像中 RGB 通道的像素值这两种方法对图像数据进行扩充.

表 4 是对音乐曲目的图像进行扩充后得到的实验结果图,从此表中可以看出扩充实验数据对本实验结果具有改善作用. 因此数据扩充已经成为生成更多图像样本和应对各种差异获得鲁棒性的重要方式. 然而本文的实验结果识别率偏低,经分析发现: 音乐曲目的变化是非常丰富的,因此使用 100 首曲目来代表一种特定类型的各种变体是不够的,而且相较于 8 层的网络结构来说本文的训练数据是偏少的,以至于最终的分类结果不是特别理想. 可以预见随着音乐曲目的增多,本文的识别效果将会进一步提升.

表 4　扩充实验数据后的识别率

Table 4　Classification rate with data augmentation

图片类型	识别率
Original Spectrograms	68%
Harmonic Spectrograms	60%
Percussive Spectrograms	63%
Original + Harmonic + Percussive Spectrograms	77%

3.6　与其他方法的准确率对比

当前有很多学者针对音乐风格识别提出了不同的研究方法,如表 5 所示,Gwardys[16] 也使用了 HPSS 算法获得频谱图,而后微调了一个 8 层网络,最终获得 72%的准确率,而本文使用此频谱图训练了此 8 层网络,准确率有所提升. Lee[17]

训练了一个只有 2 层的 CDBN,其识别模型的深度比本文的浅,数据量也比本文的少,但实验结果很接近本文扩充数据前的正确率,由此可见小数据集在浅层网络里也能有较好的结果. 杨松[18] 运用了传统机器学习方法中的 K 均值聚类来进行音乐识别,其识别率为 71%,本文提出的深度学习方法相比较下分类准确率有一定的提升.

表 5 本文方法与其他方法比较

Table 5 Comparison to other state-of-the-art methods

方法	准确率(%)
Gwardys[16]	72
Lee[17]	73
杨松[18]	71
本文方法	77

4 总 结

本文提出了一种基于卷积神经网络的音乐风格识别方法,详细设计了该方法使用的网络框架并研究了一些影响其分类性能的关键因素. 起初使用原始谱图进行实验时,得到的识别率只有 67%,为了改善结果,实施了 harmonic/percussive 分离实验,最后识别率提高了 6%. 此外本次实验表明了数据扩充的重要性和有效性,特别是训练数据不够时,当把本次实验数据扩充后,实验效果改善了 3%.

本文的实验取得了一定的识别率,使用了卷积神经网络的框架,再加上数据扩充等手段,使得深度学习应用于小数据集成为可能,并取得了一定的判别能力. 在将来的工作中,一方面通过搜集每种风格更多的音乐曲目来提高识别率,另一方面可以构造混合网络结构,如卷积神经网络和循环神经网络,可利用卷积神经网络提取局部特征,循环神经网络对这些提取的特征作时间上的整合,由于循环神经网络对前面的信息具有记忆功能,可使提取的特征更具有整体性和连贯性,提高识别结果.

References:

[1] Patil N M, Nemade M U. Content-based audio classification and retrieval: a novel approach [C]. International Conference on Global Trends in Signal Processing, Information Computing and Communication(ICGTSPICC) , IEEE, 2017: 599-606.

[2] Costa Y M G, Oliveira L S, Koerich A L, et al. Music genre classification using LBP textural features [J]. Signal Processing, 2012, 92 (11) : 2723-2737.

[3] Meng L, Ding S, Xue Y. Research on denoising sparse autoencoder [J]. International Journal of Machine Learning & Cybernetics, 2017, 8(5) : 1719-1729.

[4] Zeng Kai, Ding Shi-fei. Advances in image super-resolution reconstruction [J]. Computer Engineering and Applications, 2017, 53 (16) : 29-35.

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]. Advances in Neural Information Processing Systems(NIPS) , 2012: 1097-1105.

[6] Gopalakrishnan K, Khaitan S K, Choudhary A, et al. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection [J]. Construction & Building Materials, 2017, 157: 322-330.

[7] Sharif Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014: 806-813.

[8] Meng Ling-heng, Ding Shi-fei. Depth perceptual model based on the single image [J]. Journal of Shandong University(Engineering Science) , 2016, 46(3) : 37-43.

[9] Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks [C]. Advances in Neural Information Processing Systems(NIPS) , 2009: 1096-1104.

[10] Li T L, Chan A B, Chun A H. Automatic musical pattern feature extraction using convolutional neural network [C]. Proc. Int. Conf. Data Mining and Applications, 2010.

[11] Lu X, Lin Z, Jin H, et al. Rating image aesthetics using deep learning [J]. IEEE Transactions on Multimedia, 2015, 17 (11) : 2021-2034.

[12] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4) : 212-223.

[13] Tzanetakis G, Cook P. Musical genre classification of audio signals [J]. IEEE Transactions on Speech and Audio Processing, 2002, 10 (5) : 293-302.

[14] Bengio Y. Practical recommendations for gradient-based training of deep architectures [M]. Neural Networks: Tricks of the Trade, Springer Berlin Heidelberg, 2012: 437-478.

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

[16] Gwardys G, Grzywczak D. Deep image features in music information retrieval [J]. International Journal of Electronics and Telecommunications(IJET) , 2014, 60(4) : 321-326.

[17] Lee H, Yan L, Pham P, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks [C]. International Conference on Neural Information Processing Systems, Curran Associates Inc, 2009: 1096-1104.

[18] Yang Song, Yu Feng-qin. Speech/music discriminator based on sample entropy [J]. Computer Engineering and Applications, 2012, 48(23) : 125-127.

附中文参考文献:

[4] 曾 凯, 丁世飞. 图像超分辨率重建的研究进展 [J]. 计算机工程与应用, 2017, 53(16) : 29-35.

[8] 孟令恒, 丁世飞. 基于单静态图像的深度感知模型 [J]. 山东大学学报(工学版) , 2016, 46(3) : 37-43.

[18] 杨 松, 于凤芹. 基于样本熵的语音/音乐识别 [J]. 计算机工程与应用, 2012, 48(23) : 125-127.