基于多尺度残差式卷积神经网络与双向简单循环单元的光学乐谱识别方法

吴琼，李锵，关欣*
天津大学微电子学院，天津 300072

摘要 光学乐谱识别在音乐信息检索和计算机辅助教学等领域有着重要价值，针对传统框架处理步骤复杂、精度较低，而基于深度学习的算法模型训练耗时久，且对难点音符识别误差较大的问题，提出了一种改进的卷积循环神经网络以提升识别精度。首先在原始乐谱中增加不同的噪声，以扩充乐谱图像，提高训练模型的鲁棒性；随后利用多尺度残差式卷积神经网络对乐谱图像中的音符特征进行提取，提升后续识别精度；最后利用双向简单循环单元网络识别音符特征，加快训练收敛速度。实验结果表明，改进后网络模型的平均符号错误率下降至0.3234%，收敛速度加快，训练耗时约为传统卷积循环神经网络的 1/3。

关键词 数字图像处理；光学乐谱识别；卷积神经网络；多尺度特征融合；简单循环单元
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.081006

Optical Music Recognition Method Combining Multi-Scale Residual Convolutional Neural Network and Bi-Directional Simple Recurrent Units

Wu Qiong, Li Qiang, Guan Xin*
School of Microelectronics, Tianjin University, Tianjin 300072, *China*

Abstract Optical music recognition plays an important role in the field of music information retrieval and computer aided instruction. For traditional frameworks, the processing steps are complicated, and the accuracy is low. Moreover, deep learning algorithm-based model training takes a long time and shows large recognition error for difficult notes. In this work, an improved convolutional recurrent neural network is proposed. First, different noises were added to the original score to expand the score image and improve the robustness of the training model. Then, the multi-scale residual convolutional neural network was used to extract note features to improve the subsequent recognition accuracy. Finally, bi-directional simple recurrent units were adopted to recognize note features and accelerate convergence of the algorithm in the training stage. Experimental results show that the average symbol error rate of the proposed network model has been reduced to 0.3234%. Thanks to the faster converging rate, the training time is about one third of that of traditional convolutional recurrent neural network.

Key words digital image processing; optical music recognition; convolutional neural network; multi-scale feature fusion; simple recurrent units
OCIS codes 100.2000; 100.2980; 100.3008

1 引 言

乐谱能对音符、音调和音长等相关信息进行详尽描述，是音乐人学习、分享以及传播音乐最直接的方式。很多经典乐谱由于环境变化和时代变迁受到了损坏，甚至丢失，可见人为储存并不能完好无损地保留所有乐谱。随着计算机应用与图像扫描等技术的快速发展，纸质版乐谱文件经光学乐谱识别

收稿日期：2019-06-26；**修回日期**：2019-08-21；**录用日期**：2019-09-10
基金项目：国家自然科学基金(61471263)、天津市自然科学基金(16JCZDJC31100)
* **E-mail**：guanxin@tju.cn

(OMR)技术可转化为计算机能够“读懂”的电子版文件,从而广泛应用于音乐信息检索、音乐辅助教学等领域。但由于传统的乐谱识别算法结构复杂,实现难度较大,现有的商业识别软件精度较低,因此需要研究一种易实现且高精度 OMR 算法。

自 OMR 技术发展以来,大多是基于传统的框架[1]进行优化,主要包括图像预处理[2]、五线谱检测和删除[3-5]、符号识别和分类[6-7]。Vo 等[8]采用高斯混合马尔科夫随机场(GMMRF)模型实现图像二值化,对于背景复杂的乐谱可以有效去除噪声并保留音符与五线谱间的特性;dos Santos Cardoso 等[9]提出了基于稳定路径的谱线检测方法,在不需要领域知识的情况下将检测错误率下降至 1.4%.;吴天龙等[10]针对手写乐谱提出一种基于多维局部二值模式和 XGBoost 的方法删除五线谱,仅用 0.05%的训练数据就能提升模型的识别精度;Calvo-Zaragoza 等[11]利用卷积神经网络(CNN)提取特征并由 k 近邻(k-NN)等分类器实现符号分类,实验表明混合方案将分类错误率降低至 3.61%。虽然传统框架中每个单独步骤的优化效果良好,但复杂度较高且在整体精度上提升不明显。随着深度学习的不断发展[12-13],不少学者采用端对端的方法对整个乐谱图像进行处理,降低了传统框架的复杂度,也避免了单独任务之间错误的传播问题。Hajič 等[14]将 CNN 与边框回归相结合对符头进行检测,实验表明其准确率为 81%,但是在低质量二值化和模糊变形的早期乐谱图像中,后置过滤分类效果存在不稳定的问题。Choi 等[15]采用 CNN 和空间变压器网络(STN)结合的方法检测变音记号中的升记号、降记号和还原记号,检测精度达到 99.2%,但这两种方法仅针对特定符号进行识别,应用范围小且拓展性差。Calvo-Zaragoza 等[16]将 CNN 和双向长短时记忆(BiLSTM)相结合构成卷积循环神经网络 C-BiLSTM,对整个乐谱图像中的音符进行识别,达到了 2.16%的符号错误率,但模型收敛缓慢,消耗时间较长,对于难点音符如倚音、小节线等识别精度不够。Tuggener 等[17]将 ResNet-101 与 RefineNet 上采样网络相连并结合边框检测方法识别乐谱图像,对于全休止符的识别效果良好,但对于其余音符尤其是变音记号和拍号识别精度不足 50%。

针对上述问题,本文对 Calvo-Zaragoza 等[16]所用模型进行改进,结合多尺度残差式 CNN 与双向简单循环单元(MF-RC-BiSRU)的方法识别乐谱图像。主要在两个方面进行优化:1)在特征提取部分将 CNN 改进为多尺度残差式 CNN,将不同卷积层提取的特征进行像素级融合,使多层次特征集中于统一的特征图中,增强了模型的特征表示能力,提升了模型对音符的特征提取能力;2)将音符识别部分中的 BiLSTM 优化为双向简单循环单元(BiSRU),使训练部分计算转化为并行计算,从而加快收敛速度并减少训练时长。

2 MF-RC-BiSRU 光学乐谱识别方法

MF-RC-BiSRU 光学乐谱识别方法原理如图 1 所示,首先将输入的乐谱图像高度固定为128 pixel,宽度按比例放缩,加入噪声以模拟真实环境中各种不理想的乐谱图像;随后利用五层残差式 CNN 对图像中的音符信息进行不同层次的特征提取,同时将深层的语义特征信息与浅层的细节特征信息进行多尺度融合,通过多层次信息的交叉补充,为下一阶段音符识别提供更完善的特征信息;最后将提取的特征序列进行维度转换作为音符识别部分的输入,利用 BiSRU 完成音符序列的识别,采用对数据集无强制对齐要求的链式时序分类(CTC)函数实现音符分类。

2.1 用于音符特征提取的残差式 CNN 结构

乐谱图像中音符离散且分布较为均匀,主要由多个方向上的直线或曲线、实心或空心的近圆图形构成。CNN 中卷积层具有局部连接和权值共享特点,利于提取音符的边缘特征以及位置信息,因此采用 CNN 提取乐谱图像中音符的特征;激活函数层可增强 CNN 的表达能力,使 CNN 具有可微性,从而实现乐谱图像从低维简单特征到高维复杂特征的非线性映射;池化层在保留卷积层主要特征的前提下减少权值参数量,加快计算速度并防止出现过拟合问题。为提升模型的检测精度,通常需要增加 CNN 的层数宽度或深度,但在参数更新过程中易出现梯度消失/爆炸问题,导致模型不收敛。

传统 CNN 输入数据 $\boldsymbol{x}$ 后通过卷积层与非线性激活函数层后得到输出 $\boldsymbol{y}=H(\boldsymbol{x})$,但拟合函数比较困难。为解决模型不收敛问题,选用 Zhang 等[18-19]提出的残差式 CNN,引入残差学习 $F(\boldsymbol{x})=H(\boldsymbol{x})-\boldsymbol{x}$,将目标映射函数转化为 $H(\boldsymbol{x})=F(\boldsymbol{x})+\boldsymbol{x}$。在优化过程中,令 $F(\boldsymbol{x})$无限接近于 0,$H(\boldsymbol{x})=\boldsymbol{x}$。既不需直接拟合函数 $H(\boldsymbol{x})$,也没有增加新的参数和计算复杂度,同时使用随机梯度下降法进行端对端的训练,不仅解决了退化问题,还提升了模型的检测精度。设计的残差式CNN如图2所

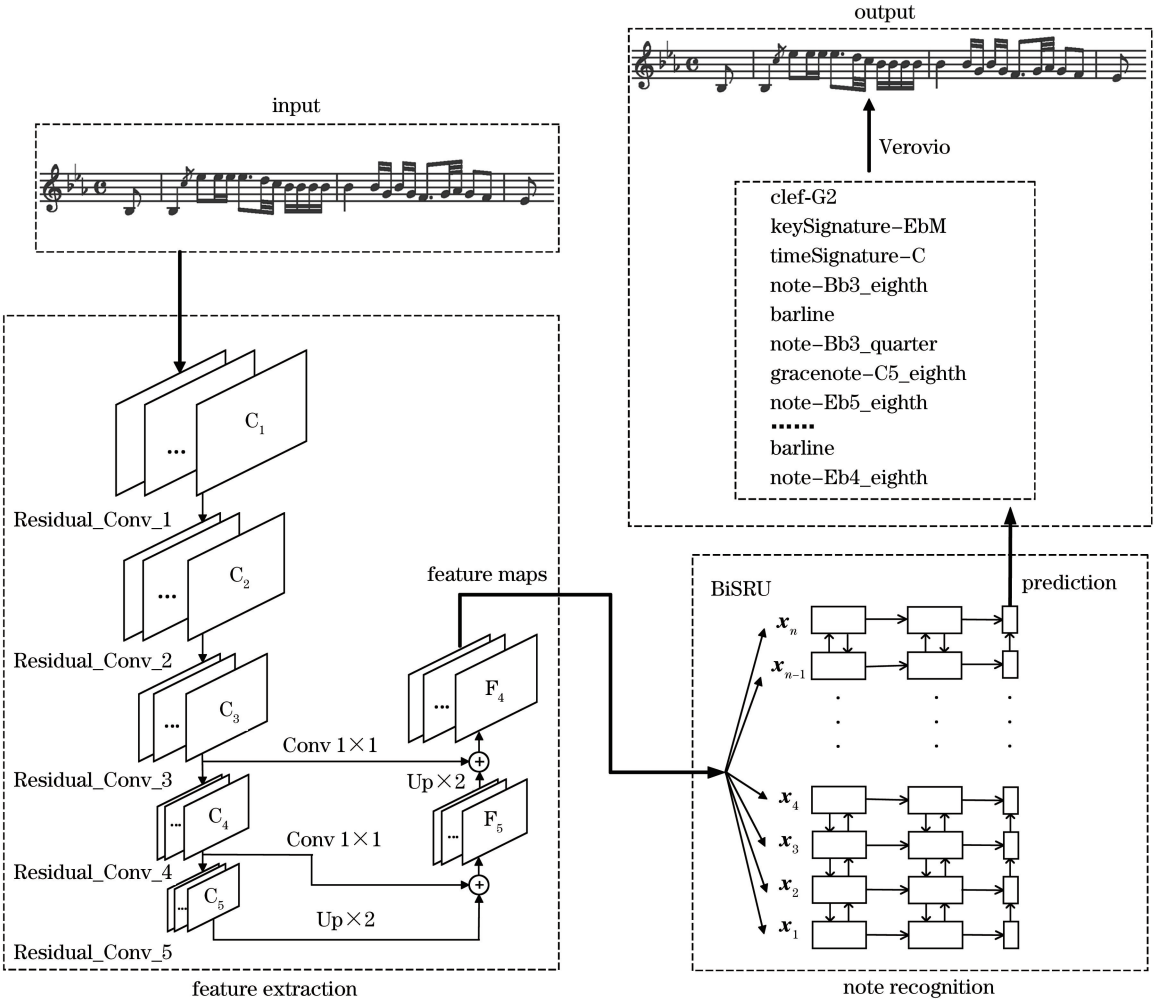

图 1 MF-RC-BiSRU 原理框图

Fig. 1 Schematic diagram of MF-RC-BiSRU

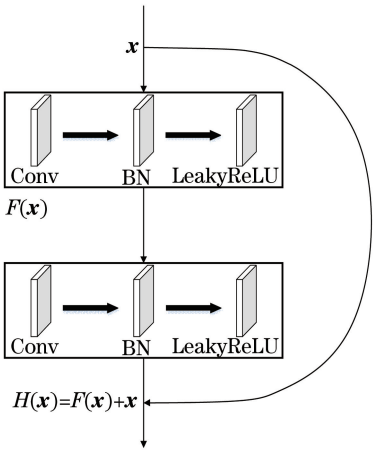

图 2 残差结构示意图

Fig. 2 Schematic diagram of residual structure

示，每一个残差块跳跃两个特征提取模块，每一个特征提取模块包含一个卷积层、一个批归一化(BN)层以及一个激活函数层。常用的激活函数包括 Sigmoid、Tanh 以及 ReLU 等，考虑到 ReLU 激活函数负半轴为 0，呈“死区”状态，梯度在更新过程中可能会消失，于是选择 LeakyReLU 函数，虽然其在负半轴时仍有很小的梯度，但能有效避免“死区”状态，可表示为

$$f(x)=\begin{cases}0.01x & x\leqslant 0\\ x & x>0\end{cases},\tag{1}$$

式中，x 为激活函数输入，$f(x)$为激活函数输出。

2.2 多尺度特征融合

利用 CNN 提取音符特征的过程中，卷积层数不断增加使模型提取到不同层次信息的特征，浅层特征一般包含音符位置、边缘信息等，深层特征虽然分辨率小但却拥有丰富的语义信息，可辅助网络更好地识别音符。但由于缺失浅层网络中的细节特征，可能会影响音符的识别精度。因此，在音符识别过程中对 CNN 深层语义信息与浅层细节信息进行了多尺度融合，具体过程如图 3 所示。

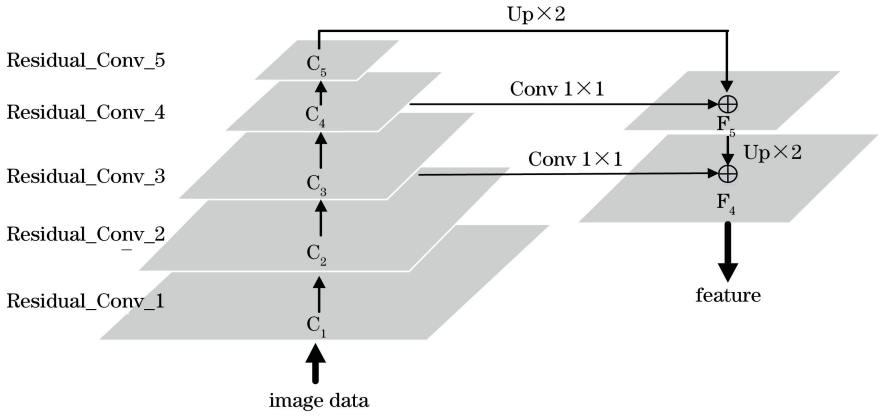

图 3 多尺度特征融合示意图

Fig. 3 Schematic diagram of multi-scale feature fusion

图 3 中，左半部分由五层残差式 CNN 自下而上对乐谱图像进行特征提取，每一层残差式 CNN 内部经过两次卷积，最大池化操作紧随每一层卷积。原始乐谱图像依次通过每一层残差式 CNN 后分别得到特征图 C_1、C_2、C_3、C_4 和 C_5，其大小逐层递减至上一层图像的 1/4。其中卷积核尺寸均为 3×3，卷积核数量分别为 32、64、128、256 和 256，具体参数如表 1 所示。右半部分结构为自上而下的特征融合部分，将含有语义信息的较深层特征图 C_5 与上一级特征图 C_4 进行像素级融合，因融合要求两个特征图在尺寸大小及维度方面保持一致，故将 C_5 通过 2 倍上采样使其与特征图 C_4 的大小一致，特征图 C_4 通过 1×1 卷积核进行卷积处理，以保证与 C_5 上采样后特征维度一致，融合后得到特征图 F_5。对特征图 F_5 和 C_3 进行相同操作，就能得到包含不同层次信息的特征实现多尺度特征融合并最终得到特征图 F_4。

表 1 改进网络的结构参数

Table 1 Structure parameters of the improved network

Input(128×weight×1)		
Part	Layer	Parameters
Feature extraction	Residual_Conv_1	(3,3,32)
	Max_Pool	(2,2,32)
	Residual_Conv_2	(3,3,64)
	Max_Pool	(2,2,64)
	Residual_Conv_3	(3,3,128)
	Max_Pool	(2,2,128)
	Residual_Conv_4	(3,3,256)
	Max_Pool	(2,2,256)
	Residual_Conv_5	(3,3,256)
	Max_Pool	(2,2,256)
Note recognition and classification	BiSRU	512
	BiSRU	512
	CTC	1780

2.3 用于音符识别的简单循环网络单元

虽然有多种网络可以对乐谱图像中的音符进行有效识别，但由于音符序列具有顺序性，每一首乐谱的音符种类和音符间的顺序是固定的，当前时刻音符与其前后时刻音符有很强的相关性。实验采用循环神经网络(RNN)识别音符，RNN 在训练过程中通常因数据长度较大容易出现梯度消失问题，因此需要通过具有“门机制”的长短时记忆网络(LSTM)或门控制单元(GRU)等模型来控制信息流，以缓解其梯度消失的潜在问题。但 LSTM 或 GRU 等模型的忘记门、输入门以及单元状态除依赖当前时刻输入外，仍需前一时刻隐藏单元的输出，很大程度上限制了并行运算的速度。因此采用简单循环单元(SRU)[20]模块，结构如图 4 所示，解除连续时刻状态间的强制约性，利用循环性较弱与较高的并行性使门状态的计算只依赖于当前时刻的输入信息。

图 4 中当前时刻 t 下隐藏单元的忘记门 $\boldsymbol{f}_t$，单元状态 $\boldsymbol{c}_t$ 定义为

$$\boldsymbol{f}_t=\sigma(\boldsymbol{w}_{\mathrm{f}}\boldsymbol{x}_t+\boldsymbol{b}_{\mathrm{f}}), \tag{2}$$

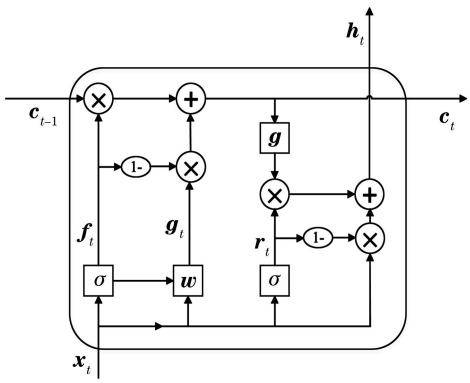

图 4 SRU 结构图

Fig. 4 Structure of SRU

$$\boldsymbol{c}_t=\boldsymbol{f}_t\cdot\boldsymbol{c}_{t-1}+(1-\boldsymbol{f}_t)\cdot\boldsymbol{g}_t,\tag{3}$$

式中，σ 为 Sigmoid 激活函数，$\boldsymbol{w}_{\mathrm{f}}$ 和 $\boldsymbol{b}_{\mathrm{f}}$ 分别为忘记门的参数矩阵与偏置，$\boldsymbol{g}_t$ 为对当前时刻输入 $\boldsymbol{x}_t$ 进行线性变换，即 $\boldsymbol{g}_t=\boldsymbol{w}\boldsymbol{x}_t$，$\boldsymbol{w}$ 为其参数矩阵。

当前时刻 t 的中间输出状态 $\tilde{\boldsymbol{c}}_t$ 可通过对单元状态 $\boldsymbol{c}_t$ 进行非线性变换得到，即 $\tilde{\boldsymbol{c}}_t=g(\boldsymbol{c}_t)$，其中 g 为 Tanh 激活函数；而重置门 $\boldsymbol{r}_t$ 用于计算中间输出状态 $g(\boldsymbol{c}_t)$ 和输入 $\boldsymbol{x}_t$ 的组合输出状态 $\boldsymbol{h}_t$，可表示为

$$\boldsymbol{r}_t=\sigma(\boldsymbol{w}_{\mathrm{r}}\boldsymbol{x}_t+\boldsymbol{b}_{\mathrm{r}}),\tag{4}$$

式中，$\boldsymbol{w}_{\mathrm{r}}$ 和 $\boldsymbol{b}_{\mathrm{r}}$ 分别为重置门 $\boldsymbol{r}_t$ 的参数矩阵与偏置，最终 SRU 的输出状态 $\boldsymbol{h}_t$ 可表示为

$$\boldsymbol{h}_t=\boldsymbol{r}_t\cdot g(\boldsymbol{c}_t)+(1-\boldsymbol{r}_t)\cdot\boldsymbol{x}_t。\tag{5}$$

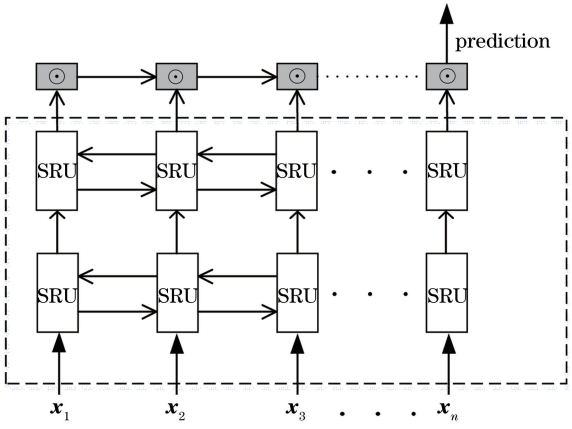

图 5 BiSRU 示意图

Fig. 5 Schematic diagram of BiSRU

音符的识别过程如图 5 所示。模型共包含两层 BiSRU，每层循环长度因乐谱图像的高度与所选卷积核数确定而保持不变，每一个 SRU 中其权重正向学习和反向传播均通过 512 个隐藏层单元实现。图5中虚线框内的矩形块表示对当前时刻输入 $\boldsymbol{x}_t$ 进行矩阵计算的 SRU，而虚线框外矩形块表示前后时刻输出间的点乘计算，利用该方式完成大多数并行计算，加速了模型的收敛速度。

利用 SRU 强大的时序建模能力对乐谱图像数据进行训练时，网络需要为序列中每一个音符提供对应的期望输出，即对应标签。但 RNN 在损失计算过程中要求音符对应标签与原始图像像素严格对齐，否则需对输入数据进行预处理操作或对输出数据进行后处理，但无论利用人工对齐还是采用开源工具对齐都会耗费大量时间，甚至影响识别精度，因此，使用基于链式时序分类(CTC)的损失函数[21]代替交叉熵损失函数。仅关注标签之间相对位置的准确性，不必强制对齐便可自动学习位置信息，大大降低了对训练集的要求。CTC 将网络输出转换为标签序列上的条件概率分布，当条件概率分布确定时，可通过标签序列概率最大化得到最终目标序列，即对于给定的乐谱图像，其输出路径由音符序列中每一位置所选择不同的音符形成，输出音符条件概率的不同导致所选路径概率也存在差异，通过遍历多条路径选取概率最大的路径，从而实现精准识别和分类音符。

3 实验结果与分析

3.1 实验所用数据集

实验所用数据为开放数据集 PrIMus Dataset (Printed images of music staves)中约 87687 例真实的谱例。如图 6 所示，每例乐谱由单行五线谱构成且由小节线划分为 4～7 个小节。大部分谱例中不仅有简单音符序列的组合，还包含了较丰富的音符种类，如谱号、拍号、升降记号、休止符以及倚音、附点音符等识别难点。

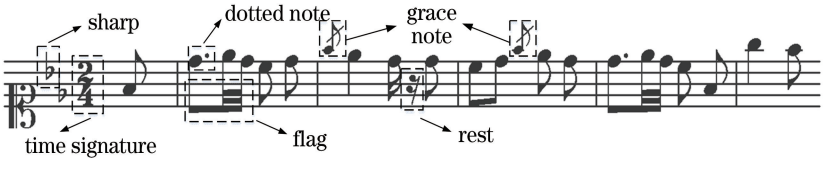

图 6 乐谱中音符识别难点

Fig. 6 Difficulties of note recognition in music score

本算法在原有数据集上，通过计算机程序加入高斯白噪声、柏林噪声以及旋转拉伸等弹性形变，模拟乐谱图像可能出现的几种不理想情况，对数据集进行扩充，以提高模型的泛化能力。以图 7(a)的乐谱为例，通过加入高斯白噪声模拟乐谱图像低质量打印或扫描的情况，结果如图 7(b)所示；图 7(c)为加入柏林噪声的乐谱，使乐谱产生局部变淡甚至褪色效果，可模拟乐谱图像中因打印墨粉不均匀或存留时间久的情况；图 7(d)中对图像进行拉伸、旋转、歪斜、缩放等弹性形变处理，模拟图像在印刷过程中出现的轻微折叠、畸变等现象，利用以上三种数据处理方法对数据集进行扩充，并以 8∶1∶1 的比例将数据集划分为训练集、验证集以及测试集。

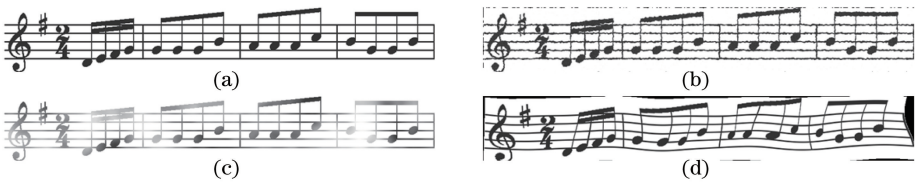

图 7 模拟不理想乐谱图像的三种数据处理方法效果图。(a)原始谱例;(b)加入高斯白噪声谱例;(c)加入柏林噪声谱例;(d)加入弹性形变谱例

Fig. 7 Three methods of data processing to simulate unsatisfactory music image. (a) Original incipit; (b) incipit of white Gaussian noise added; (c) incipit of Perlin noise added; (d) incipit of elastic transformations added

3.2 评价指标

目前 OMR 算法的评价指标并没有统一的规范,实验采用常见的序列错误率和符号错误率等指标对算法进行评估。

序列错误率(Sequence error rate)指预测错误的序列数占总序列数的比率,序列中至少有一个音符、音高、休止符等出现错误即属于错误序列。

符号错误率(Symbol error rate)指从预测的序列中产生标签序列所需的插入、修改或删除的基本编辑操作的平均数量占当前序列长度的比例。

序列错误率与符号错误率没有绝对的相关性,序列错误率对集中测试谱例的错误比例进行描述,而符号错误率是针对谱例中音符的错误情况进行总结,因此实验对音符识别精准度的衡量更侧重于符号错误率这一指标,但序列错误率在很多领域应用中仍具有指导意义。

3.3 CNN 采用残差结构前后的性能对比

实验环境:Ubuntu16.04 操作系统,Intel Core i7-8700 CPU,16 G 运行内存,Nvidia GTX1080Ti GPU,TensorFlow 深度学习框架。模型在训练过程中采用 Adam 自适应学习率算法进行优化,初始学习率设置为 exp(−3),批处理大小设置为 16,每经过 1000 次迭代算法在验证集上进行符号错误率的评估以验证模型的精度,整个过程共迭代 64000 次左右。

首先,将卷积循环神经网络 C-BiLSTM[16] 中的 CNN 改进为残差式 CNN,构成残差式卷积循环神经网络 RC-BiLSTM,同等实验条件下,对 C-BiLSTM 和 RC-BiLSTM 分别进行训练并比较其识别精度。两种网络在模型训练过程中损失函数值随迭代次数的变化如图 8(a)所示,可以看出 RC-BiLSTM 网络每一次迭代的损失值均低于 C-BiLSTM 网络,经过 15×10^3 轮训练后其损失值已降低至 5 并趋于稳定,而 C-BiLSTM 网络仅下降至 10 左右且存在较大波动。在验证集上比较两种算法中的符号错误率,结果如图 8(b)所示。可以看出,C-BiLSTM 网络错误率最低下降至 4%左右,而 RC-BiLSTM 网络的符号错误率可稳定下降至 2%以下,音符识别准确率有明显提升。这表明残差式 CNN 不仅可提升模型精度,还可解决模型退化问题,增强模型泛化能力。

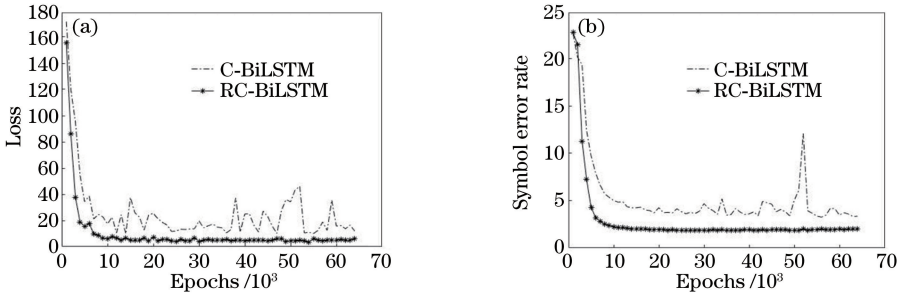

图 8 C-BiLSTM 和 RC-BiLSTM 网络的训练损失与精度对比。(a)训练损失对比;(b)符号错误率对比

Fig. 8 Comparison of training loss and accuracy for C-BiLSTM and RC-BiLSTM networks. (a) Comparison of training loss; (b) comparison of symbol error rate

3.4 多尺度特征融合有效性的实验结果与分析

为了对多尺度特征融合的有效性进行验证,提取不同层次下卷积层中的特征图,选取图 3 中 C_1、C_3、C_5 与 F_4 特征图对比分析,结果如图 9 所示。对比发现,图 9(b)所示的浅层卷积层特征图 C_1 中侧重提取符杆、符尾等音符基元的位置信息,但对拍号、音符时值以及小节线等信息的提取微乎其微。而图 9(c)所示特征图 C_3 中,增加了附点音符的补充信

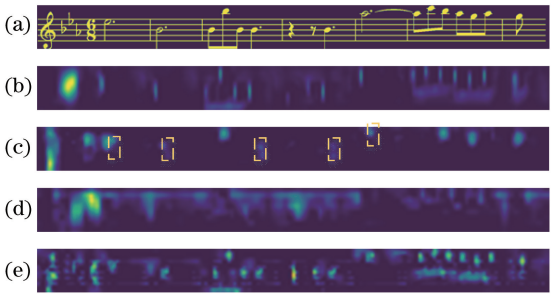

图 9 不同卷积层中特征对比。(a)原始谱例;(b)浅层特征图 C_1;(c)较深层特征图 C_3;(d)深层特征图 C_5;(e)多尺度融合后特征图 F_4

Fig. 9 Comparison of features in different convolution layers. (a) Original incipit; (b) shallow feature map C_1; (c) deeper feature map C_3; (d) deepest feature map C_5; (e) multi-scale feature fusion map F_4

息,不仅包含音符位置信息,还捕捉到谱号的相关信息。从图 9(d)可以看到深层卷积层已不再提取简单易读的信息,而是对拍号以及小节线等语义信息的提取较为突出,但丢失了音符的位置等基本信息。将不同卷积层提取的特征图融合后的效果如图 9(e)所示,从中可看出特征图 F_4 中包含更广泛的信息,这验证了多尺度特征融合对于音符的表示能力较高。

在 RC-BiLSTM 网络中结合多尺度特征融合构成 MF-RC-BiLSTM 网络,以验证其对符号识别正确率的影响。图 10 为 RC-BiLSTM 网络和 MF-RC-BiLSTM 网络在验证集上的符号错误率,可以看出 MF-RC-BiLSTM 网络的符号错误率明显

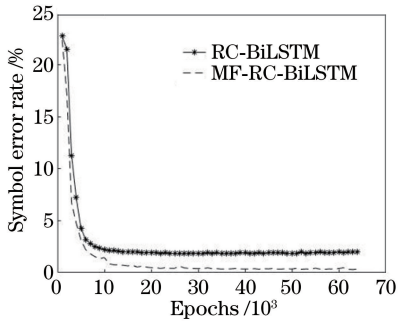

图 10 不同网络符号错误率对比

Fig. 10 Comparison of the symbol error rates in the different networks

降低至 0.5%以下,可进一步提高音符的识别正确率。

3.5 SRU 网络收敛速度对比与分析

将 MF-RC-BiLSTM 网络中的 BiLSTM 模型优化为 BiSRU 模型,提出了 MF-RC-BiSRU 光学乐谱识别算法并验证其收敛速。两次实验均迭代 64000 次,MF-RC-BiLSTM 总耗时约 16 h,平均每次约 0.92 s,MF-RC-BiSRU 总耗时约 10 h,平均每次约 0.56 s。图 11 为训练损失及验证集中符号错误率对比结果,从图 11(a)中可看出 MF-RC-BiLSTM 经历 12×10^3 次迭代损失已下降至约 1.8892,耗时约 18 min,而 MF-RC-BiSRU 在 6×10^3 次迭代损失达到约 1.5437,耗时约 6 min,仅为 MF-RC-BiLSTM 的 1/3。这表明本算法的收敛速度相较于 BiLSTM 模型来说更快,快近 3 倍。虽在精度上没有显著提升,符号错误率上仅仅下降了 0.08%,但充分验证了 SRU 模型在时序并行操作方面的高效性。

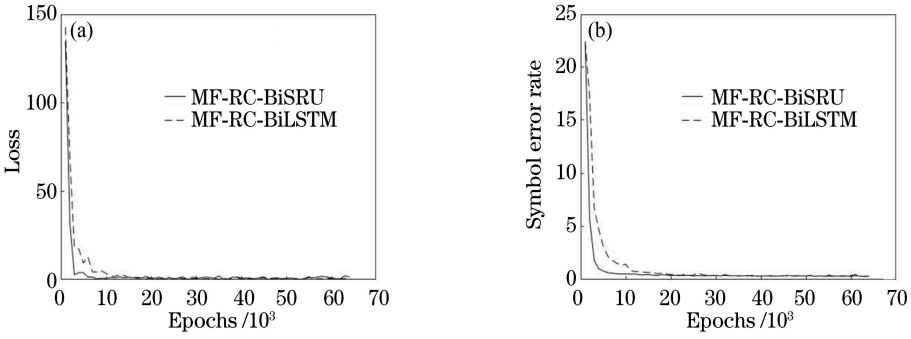

图 11 MF-RC-BiSRU 与 MF-RC-BiLSTM 的对比。(a)训练损失对比;(b)符号错误率对比

Fig. 11 Comparison of MF-RC-BiSRU and MF-RC-BiLSTM. (a) Comparison of training loss; (b) comparison of symbol error rates

表 2 为不同网络的精准度对比,可以看出在同一测试集上无论是序列错误率还是符号错误率,C-BiLSTM 网络序列错误率约 14.3498%,符号错误率约 3.2480%,而 MF-RC-BiSRU 网络的精度均达到最优,序列错误率约 1.4571%,符号错误率约 0.3234%。精准度提高了约 10 倍,提升效果显著。

表 2　不同网络精准度对比

Table 2　Comparison of accuracy in different networks

Network	Symbol error rate /%	Sequence error rate /%
C-BiLSTM	3.2480	14.3498
RC-BiLSTM	1.8440	8.1071
MF-RC-BiLSTM	0.3312	1.4637
MF-RC-BiSRU	0.3234	1.4571

考虑到网络模型测试得到的实验结果均为音符的标签信息，利用 Verovio 乐谱显示软件将其恢复为乐谱图像，对结果进行更直观的对比。将图 12(a)所示谱例通过 C-BiLSTM、RC-BiLSTM、MF-RC-BiLSTM 和 MF-RC-BiSRU 四种网络模型恢复为乐谱图像，结果分别如图 12(b)～图 12(e)所示。从图 12(b)和图 12(c)中可以看出通过加入残差式 CNN 有效识别出了第一小节中音符的符尾；图 12(d)中多尺度特征的融合校正了 RC-BiLSTM 网络对第三小节中升降记号的错误预测，也正确预测了第五小节中的附点音符；而从图 12(d)～图 12(e)可以看到模型收敛速度加快并未影响模型精度。这表明 MF-RC-BiSRU 模型能明显改善 C-BiLSTM[16] 中难点音符的识别问题，对图中音符符尾、附点音符、升降记号及倚音等识别难点可做到准确识别，且在一定程度上加快了模型的训练收敛速度，减少整体训练时间。

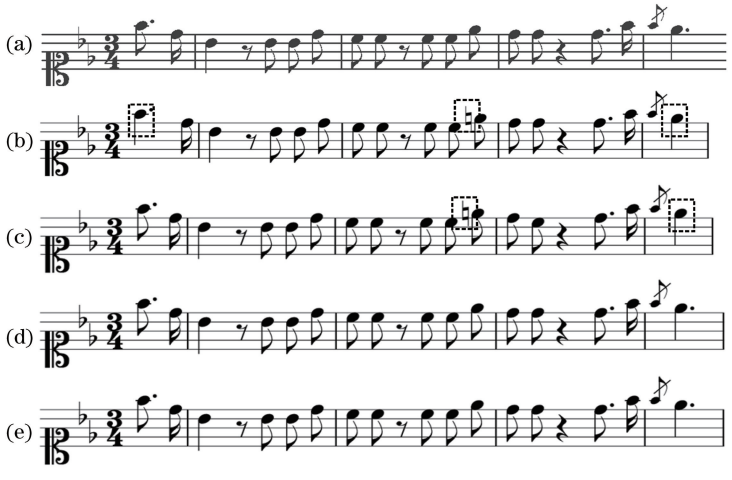

图 12　同一谱例在四个不同网络下测试结果。(a)原始谱例；(b) C-BiLSTM；(c) RC-BiLSTM；(d) MF-RC-BiLSTM；(e) MF-RC-BiSRU

Fig. 12　Test results of the same incipit in four different networks. (a) Original incipit; (b) C-BiLSTM; (c) RC-BiLSTM; (d) MF-RC-BiLSTM; (e) MF-RC-BiSRU

3.6　MF-RC-BiSRU 性能对比与分析

将 MF-RC-BiSRU 与其他方法进行对比实验，结果如表 3 所示，其中的时间为一轮迭代所用的时间，可以看出该法在符号错误率和序列错误率中均达到较好的效果。尽管 CNN-STN 算法对变音记号的识别精度较高，但将其用于整个乐谱时，识别能力受限导致乐谱图像识别精度降低。而数据仓库分层(DWD)虽然可以识别所有音符，但对不同类型音符的错误率存在较大差异性，全音符识别率不足 80%，而对于六连音的识别率达到 95%。其次，从每次迭代的平均耗时可以看到 DWD 耗时几乎是 MF-RC-BiSRU 耗时的 2 倍。由图 13 可见，MF-RC-BiSRU 方法的损失值在 6×10^3 次迭代后趋于稳定，而 DWD 在 20×10^3 次迭代后损失值降至 10 左右，CNN-STN 在 30×10^3 次迭代后损失值降至 3 左右，收敛速度明显较慢。可以发现，MF-RC-BiSRU 算法在识别精度与收敛速度上均达到较好的效果。

表 3　不同方法的性能对比

Table 3　Performance comparison of different methods

Method	Symbol error rate /%	Sequence error rate /%	Time / s
CNN-STN[15]	5.0208	16.8056	0.98
DWD[17]	8.7811	18.5609	1.21
MF-RC-BiSRU	0.3234	1.4571	0.56

4　结　　论

针对 C-BiLSTM 网络识别乐谱图像中难点音符精度不高的问题，提出了一种改进算法。首先，在预处理中通过扩充数据集提高模型泛化能力；其次，在特征提取部分采用残差式 CNN 解决模型退化问题；随后使用多尺度特征融合增强了模型的特征表

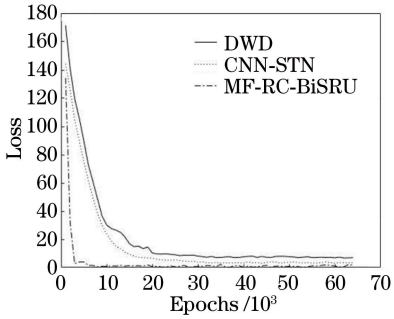

图 13 不同方法中损失对比

Fig. 13 Comparison of loss in different methods

示能力，提升了后续难点音符识别精度；最后，在音符识别部分采用 BiSRU，加快了模型收敛速度，解决了 RNN 在训练过程中耗时较多的问题。实验结果表明该算法识别精度较高、模型收敛速度较快。

由于本实验所用数据集均为单行的五线谱谱例，音符的识别也仅限于一行五线谱上的音符信息，因此对包含更多文字信息及难点音符的整篇乐谱的识别尚待进一步研究。同时该算法并未涉及难度更高的乐谱，如和弦、多声部乐谱的识别，对于这类更复杂的乐谱图像，精准找到其位置信息与音符组合关系是今后研究的难点之一。

参 考 文 献

[1] Bainbridge D, Bell T. The challenge of optical music recognition[J]. Computers and the Humanities, 2001, 35(2): 95-121.

[2] Calvo-Zaragoza J, Vigliensoni G, Fujinaga I. Pixel-wise binarization of musical documents with convolutional neural networks[C]// 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), May 8-12, 2017, Nagoya, Japan. New York: IEEE, 2017: 362-365.

[3] Timofte R, van Gool L. Automatic stave discovery for musical facsimiles[M]// Lee K M, Matsushita Y, Rehg J M, et al. Computer vision-ACCV 2012. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2013, 7727: 510-523.

[4] Gallego A J, Calvo-Zaragoza J. Staff-line removal with selectional auto-encoders[J]. Expert Systems with Applications, 2017, 89: 138-148.

[5] Visaniy M, Kieu V C, Fornes A, et al. ICDAR 2013 music scores competition: staff removal[C]// 2013 12th International Conference on Document Analysis and Recognition, August 25-28, 2013, Washington, DC, USA. New York: IEEE, 2013: 1407-1411.

[6] Pacha A, Eidenberger H. Towards a universal music symbol classifier[C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), November 9-15, 2017, Kyoto, Japan. New York: IEEE, 2017: 35-36.

[7] Rebelo A, Capela G, Cardoso J S. Optical recognition of music symbols[J]. International Journal on Document Analysis and Recognition, 2010, 13(1): 19-31.

[8] Vo Q N, Kim S H, Yang H J, et al. An MRF model for binarization of music scores with complex background[J]. Pattern Recognition Letters, 2016, 69: 88-95.

[9] dos Santos Cardoso J, Capela A, Rebelo A, et al. Staff detection with stable paths[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(6): 1134-1139.

[10] Wu T L, Li Q, Guan X. Lightweight staff removal method based on multidimensional local binary pattern and XGBoost[J]. Laser & Optoelectronics Progress, 2019, 56(6): 061006.
吴天龙, 李锵, 关欣. 基于多维局部二值模式和XGBoost的轻量谱线删除法[J]. 激光与光电子学进展, 2019, 56(6): 061006.

[11] Calvo-Zaragoza J, Gallego A J, Pertusa A. Recognition of handwritten music symbols with convolutional neural codes[C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), November 9-15, 2017, Kyoto, Japan. New York: IEEE, 2017: 691-696.

[12] Wang J, Wu X S. Medical image fusion based on improved guided filtering and dual-channel pulse coupled neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(15): 151004.
王建, 吴锡生. 基于改进的引导滤波和双通道脉冲耦合神经网络的医学图像融合[J]. 激光与光电子学进展, 2019, 56(15): 151004.

[13] Ma Y J, Ma Y T, Chen J H. Vehicle recognition based on multi-layer features of convolutional neural network and support vector machine[J]. Laser & Optoelectronics Progress, 2019, 56(14): 141001.
马永杰, 马芸婷, 陈佳辉. 结合卷积神经网络多层特征和支持向量机的车辆识别[J]. 激光与光电子学进展, 2019, 56(14): 141001.

[14] Hajič J Jr, Pecina P. Detecting noteheads in handwritten scores with convnets and bounding box regression[EB/OL]. (2017-08-05)[2019-06-25]. https://arxiv.xilesou.top/abs/1708.01806.

[15] Choi K Y, Couasnon B, Ricquebourg Y, et al.

Bootstrapping samples of accidentals in dense piano scores for CNN-based detection[C]// 2017 14th IAPR International Conference on Document Analysis and Recognition, November 9-15, 2017, Kyoto, Japan. New York: IEEE, 2017: 19-20.

[16] Calvo-Zaragoza J, Valero-Mas J J, Pertusa A. End-to-end optical music recognition using neural networks[C]// Proceedings of the 18th International Society for Music Information Retrieval Conference, October 23-27, 2017, Suzhou, China. [S.l.: s.n.], 2017: 23-27.

[17] Tuggener L, Elezi I, Schmidhuber J, et al. Deep watershed detector for music object recognition[C]// Proceedings of the 19th International Society for Music Information Retrieval Conference, September 23-27, 2018, Paris, France. [S.l.: s.n.], 2018: 271-278.

[18] Zhang K, Zuo W M, Chen Y J, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.

[19] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[20] Lei T, Zhang Y, Wang S I, et al. Simple recurrent units for highly parallelizable recurrence[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, October 31-November 4, 2018, Brussels, Belgium. Brussels: Association for Computational Linguistics, 2018: 4470-4481.

[21] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// Proceedings of the 23rd International Conference on Machine Learning ′06, June 25-29, 2006, Pittsburgh, Pennsylvania, USA. New York: ACM, 2006: 369-376.