DOI:10.11784/tdxbz202402001

强化音符位置及方向先验信息的多音光学乐谱识别

关　欣[1]，刘津津[1]，刘　辉[2]，李　锵[1]
（1. 天津大学微电子学院，天津 300072；2. 天津师范大学音乐与影视学院，天津 300382）

摘　要：由于同一个时刻包含多个音符的多音乐谱其音符符头距离近、符号间依赖关系复杂，使得多音光学乐谱识别极具挑战. 传统基于卷积和序列建模的方法，由于经典卷积存在移不变性难以精确表示音符的纵向位置信息，而传统针对上下文序列建模的方法难以有效表征调号中变音记号与五线谱内符头的空间相关性，存在符头音高识别不准、变音记号作用范围有限的问题，从而影响音符音高、时值标注的准确性. 针对以上问题，提出了一种强化音符位置及方向先验信息的多音光学乐谱识别方法. 首先，提出一种纵向位置编码方法，将纵向位置信息嵌入乐谱图像，以更精确地表示符头的纵向位置信息，从而能明确区分多音乐谱中的不同音高. 其次，提出了变音记号位置注意力，以明确建立变音记号和符头的空间依赖关系. 最后，针对多音符头纵向分布、音符序列横向排列、音符符头、符干和符尾呈现的局部方向性特点，提出了方向注意力模块，更好地捕捉音符特征分布的方向性. 在多音乐谱数据集上开展实验，实验结果表明，该方法对时值识别的符号错误率为 1.14%，对音高识别的符号错误率为 2.14%. 与当前基准方法卷积递归神经网络相比，该方法时值识别的符号错误率降低了 0.67%，对音高识别的符号错误率降低了 1.14%，对多音乐谱具有良好的识别效果.

关键词：光学乐谱识别；位置编码；位置注意力；方向注意力
中图分类号：TP18　　**文献标志码**：A　　**文章编号**：0493-2137（2025）01-0101-10

Polyphonic Optical Music Recognition with Enhanced Prior Information on Note Position and Direction

Guan Xin[1]，Liu Jinjin[1]，Liu Hui[2]，Li Qiang[1]
（1. School of Microelectronics，Tianjin University，Tianjin 300072，China；
2. School of Music and Film，Tianjin Normal University，Tianjin 300382，China）

Abstract：Polyphonic optical music recognition of notation is exceptionally challenging due to the proximity of noteheads and complex dependencies between symbols in polyphonic music. Traditional convolution methods struggle to represent the vertical position information of notes accurately due to the inherent shift-invariance of classical convolution. Moreover，conventional methods for context sequence modeling face difficulties in effectively representing the spatial correlation between accidentals and noteheads within the staff. This results in the inaccurate recognition of note pitch and a limited scope of accidental effects. As a result，the annotation accuracy of the pitch and length of notes is compromised. A method for enhancing the prior information on the note position and direction in polyphonic optical music recongnition is proposed to address these issues. First，a vertical position encoding method is proposed to embed vertical positional information into music score images，enabling precise differentiation of pitches in polyphonic music. Second，a coordinate attention mechanism is introduced for accidentals to establish the spatial dependency between accidentals and noteheads. Finally，to address the vertical distribution of polyphonic noteheads，the horizontal arrangement of note sequences，and the directional characteristics presented by noteheads，stems，and tails，

收稿日期：2024-02-01；**修回日期**：2024-04-16.
作者简介：关　欣（1977—　　），女，博士，副教授.
通信作者：关　欣，guanxin@tju.edu.cn.
基金项目：国家自然科学基金资助项目（62071323）；天津市自然科学基金资助项目（23JCZDJC00020）.
Supported by the National Natural Science Foundation of China（No. 62071323），the Natural Science Foundation of Tianjin，China（No. 23JCZDJC00020）.

a directional attention module is proposed to capture the directional distribution of note features better. Experimental evaluations conducted on a polyphonic dataset demonstrate that the proposed method achieves a symbol error rate of 1.14% for length recognition and 2.14% for pitch recognition. Compared with state-of-the-art convolutional recursive neural networks, the proposed approach reduces the symbol error rate by 0.67% for length recognition and 1.14% for pitch recognition. These findings highlight the superior performance of this method in polyphonic optical recognition.

Keywords: optical music recognition (OMR); position encoding; coordinate attention; direction attention

光学乐谱识别(optical music recognition, OMR)是指将乐谱图像转录为计算机可以理解的音乐语义符号形式[1]. 通过光学乐谱识别使计算机能够理解和编辑乐谱,这对乐谱数据库的建立、音乐信息检索、音乐智能化教学以及音乐的自动演奏与生成等领域具有重要意义[2-3]. 传统上,将乐谱图像转录为计算机可理解的音乐形式是通过手工转录的方式完成的,这不仅需要消耗大量的时间,而且由于音符相似性高,很容易出现对乐谱图像转录错误的情况,面对数以百万计的乐谱图像,急需一种高效准确的光学乐谱识别方法.

传统光学乐谱识别方法是分步进行的,包括乐谱图像预处理[4]、符号检测[5]、符号分类[6]. 然而,由于分步方法存在误差积累现象,使得乐谱识别的最终效果不理想. 随着深度神经网络的发展,基于联合优化的端到端方法成为光学乐谱识别的主流方法[7-14]. 端到端的光学乐谱识别方法采用卷积神经网络(convolutional neural network, CNN)对乐谱特征信息进行提取,使用循环神经网络(recurrent neural network, RNN)对特征序列上下文信息进行捕捉,此方法在单音乐谱识别领域取得了良好的效果,但就同一时刻包含多个音的多音乐谱识别领域鲜有研究.

多音乐谱识别需要对乐谱图像中多音音符的音高和时值语义信息进行提取表达,多音在乐谱图像上表现为同一时刻具有多个纵向排列的音符,多音乐谱符号间依赖关系复杂,符号与空间位置的相关性增强,使得多音乐谱识别成为光学乐谱识别领域一个具有挑战性的问题. 针对多音乐谱的识别, Edirisooriya 等[15]提出一种 RNNDecoder 模型,该模型通过对乐谱图像同一时刻的特征进行循环解码得到多音符号的音高和时值信息,对排列密集的多音乐谱取得了较好的识别效果. 然而,由于该方法在特征提取阶段采用卷积运算的方法,存在不能充分表达音符纵向位置信息的问题,使得特征提取过程中针对不同纵向位置的符头特征表现为同一特征值,对音高相近的符头信息不能很好地进行区别表达. 针对这一问题,本文提出了纵向位置编码的方法,通过在图像中融入纵向位置信息,使得卷积核在特征提取的过程中不仅能关注音符图像的模式信息,同时关注到符号所处的纵向位置信息,从而对不同纵向位置的音高特征进行区分.

音符音高信息不仅与符号符头特征的纵向位置有关,还与乐谱中的调号、变音记号、音符符头与空间位置的相关性有关. 调号是位于谱号之后,用于改变整行五线谱中对应位置的音符音高;变音记号位于乐谱音符符头左侧,用于临时改变本小节相应位置的音符音高. Li 等[16]将自注意网格引入乐谱识别领域,利用自注意网格的全局上下文感知能力,对音符序列间的全局上下文信息进行感知,然而由于该模型对空间位置信息表达能力较弱,不能对调号、变音记号、音符与空间位置的相关性进行准确表达. 因此,本文提出使用位置注意力[17]机制的方法通过位置信息嵌入和注意力生成使得在序列建模之前的特征包含对位置的感知能力.

最后,音符时值信息主要由符号局部内符头和符尾、符杠的关系确定,多音音符和单音符号的区别主要体现在局部符头分布的方向性,现有方法没有对符号局部方向性进行表达,使得对部分符头和符尾、符杠的关联性表达不准确,符号作用范围表达不明,使得音符识别错误. 针对以上问题,本文提出方向注意力模块,通过对十字方向信息的强化提升了模型对特征局部方向性的表达.

综上所述,本文提出了强化音符位置及方向先验信息的多音乐谱识别方法. 首先,为了增强模型的纵向位置表达能力对图片进行了纵向位置编码;其次,为了获得位置信息与各通道信息的全局依赖关系,增强位置信息与特征的相关性,引入了位置注意力模块;最后,提出了一种方向注意力模块,用于提升局部特征的方向性识别. 通过对空间与方向特征的提取与关注,在特征提取过程中对多音特征进行区分,在解码阶段选用单次解码的方法,实现了多音乐谱的识别.

1 网络模型

1.1 整体网络框架

本文提出的强化音符位置及方向先验信息的多音

乐谱识别方法是针对单行多音乐谱展开的，目标是识别出给定多音乐谱图像中的音符语义信息序列[18]. 网络由纵向位置编码、特征提取模块、特征强化模块以及序列信息提取模块组成，整体网络框架如图 1 所示.

为了保证特征维度的一致性，网络输入图像的高度固定为 128 像素，对图像宽度进行等比例缩放. 在对输入图像进行特征提取之前，采用纵向位置编码方法，将纵向位置信息融入输入图像，对图像的纵向位置进行强化；特征强化模块由两部分组成，通过位置注意力模块提取特征与空间信息的依赖关系，使用方向注意力模块强化符号的方向性表达，得到强化全局空间位置依赖性和局部符号方向信息的特征；最后，利用双向长短时记忆[19](bi-directional long short-term memory，BLSTM)网络挖掘特征的上下文关系，使用链式时序分类[20](connectionist temporal classification，CTC)损失函数对模型进行训练，根据贪婪解码方法解码出音符语义信息序列，得到最终识别结果.

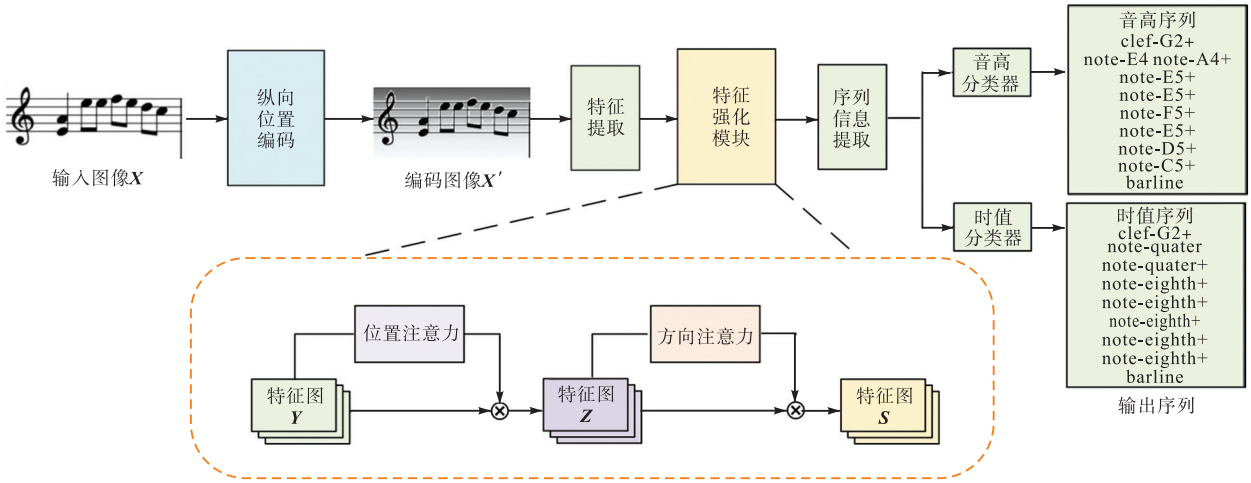


图 1　整体网络框架

Fig.1　Overall network framework

1.2　位置编码模块

卷积神经网络具有良好的特征信息提取能力，被广泛应用于乐谱识别的特征提取阶段，然而针对多音音符的符头信息，符头的纵向位置信息影响着音高语义信息的表达. 由于卷积存在平移不变性，在卷积特征提取过程中，处于不同纵向位置的符头信息得到相同的特征值表达，使得多音的音高信息识别不准确. 本文受到位置卷积[21]的启发，提出了基于位置编码的特征提取方法，将纵向位置信息编码在原始图片中，以实现对不同纵向位置符头的差异性特征表达.

由于符头特征差异性仅与符头所处的纵向位置有关，为了强化纵向位置差异性，本文提出了纵向位置编码方法. 由于输入图像纵向特征尺寸固定为 128 像素，且五线谱图像谱线间存在水平分布特点，本文选择对纵向位置信息进行线性编码的方法，线性编码公式可表示为

$$x'_{i,j} = x_{i,j} + j / H_{\mathrm{dim}} \tag{1}$$

式中：$x'_{i,j}$ 表示位于 (i,j) 处像素点经过纵向位置编码后的像素值；$x_{i,j}$ 表示位于 (i,j) 处像素点原始图像的像素值；H_{dim} 表示图像纵向总长度，为固定值，本文为 128；j 表示像素点的纵向坐标.

纵向位置编码示意如图 2 所示，经过纵向位置编码后的图像就如同在原始图片中加入横向辅助线，使得图片中每一个像素点不仅包含乐谱图像本身的模式信息，还包含了纵向位置信息，因此，在对经过纵向位置编码后的图像进行卷积运算的过程中，卷积核感受野内的信息不仅包含乐谱图像本身的模式信息，同时包含符号的纵向位置信息，从而在特征提取的过程中增强了模型对纵向位置的敏感性，强化了模型对符号纵向位置信息的特征表达.

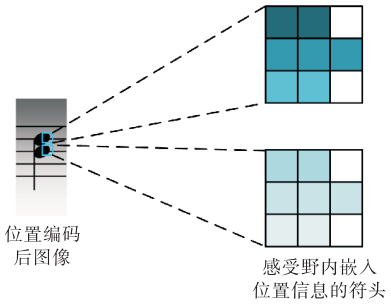


图 2　纵向位置编码示意

Fig.2　Illustration of vertical position encoding

1.3　位置注意力模块

位置注意力模块通过位置信息嵌入和位置注意

力生成两个步骤,进行了空间位置信息的凝练与特征通道信息的融合,对符号特征和空间位置的依赖关系进行了显示表达.

位置注意力模块主要通过使用一对正交的一维池化方法完成位置信息嵌入,通过卷积操作完成通道间位置注意力的生成,如图 3 所示. 正交池化核分别沿水平坐标和垂直坐标方向对每个通道进行编码,对于多音乐谱图像,水平坐标方向表达了乐谱中音高的空间分布信息,垂直坐标方向表示了乐谱符号之间的序列关系,位置信息嵌入可表示为

$$u_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}y_c(i,h) \tag{2}$$

$$u_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}y_c(w,j) \tag{3}$$

式中:u_c^h 表示第 c 通道水平位置为 h 处沿水平坐标方向的池化;u_c^w 表示第 c 通道垂直位置为 w 处沿垂直坐标方向的池化;$y_c(i,h)$ 表示第 c 通道位置为 (i,h) 处特征;$y_c(w,j)$ 表示第 c 通道位置为 (w,j) 处特征;W 为特征总宽度;H 为特征的总高度.

位置注意力生成是利用 1×1 卷积对通道的压缩扩展变换得到的. 首先,通过卷积降维操作进行通道间的信息交互,融合了不同通道之间的水平方向变音记号和音符所处的基本音位置以及垂直方向各乐谱符号分布位置的关键信息,依据位置关键信息对通道特征进行特征降维,通过非线性操作得到水平坐标方向和垂直坐标方向对空间信息编码的中间特征映射. 其次,通过卷积升维操作将位置信息通道数宽展为输入特征通道数,对语义信息表达有作用的特征通道赋予较大的权重,分别得到沿水平坐标方向不同通道对不同基本音的水平位置注意力和沿垂直坐标方向不同通道对不同序列符号的垂直位置注意力. 最后,将水平位置注意力、垂直位置注意力与输入特征相乘,得到包含对各个通道基本音位置和符号序列位置关键特征的特征图. 在公式上可表示为

$$v=\delta\left(F_1\left(\left[u^h,u^w\right]\right)\right) \tag{4}$$

$$g^h=\sigma\left(F_h\left(v^h\right)\right) \tag{5}$$

$$g^w=\sigma\left(F_w\left(v^w\right)\right) \tag{6}$$

$$z_c(i,j)=y_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{7}$$

式中:v 表示水平方向和垂直方向对空间信息编码的中间特征映射;F_1 表示通过 1×1 卷积进行降维操作;$[*,*]$ 表示沿着空间维度的拼接操作;δ 为非线性激活函数;g^h 和 g^w 分别表示水平位置注意力权重和垂直位置注意力权重;F_h 和 F_w 分别表示沿水平方向和垂直方向的卷积升维操作;σ 表示 sigmoid 函数;$z_c(i,j)$ 表示第 c 通道位置为 (i,j) 处经过位置注意力处理后的特征.

1.4 方向注意力模块

多音乐谱语义信息的表达与局部空间信息及符号间的方向性有关,位置注意力模块通过对水平坐标方向和垂直坐标方向信息分别编码,能够更好地挖掘通道特征与空间坐标的依赖关系,关注的是不同特征通道与全局空间信息之间的关系. 为了在表达不同特征通道与全局空间信息依赖关系的基础上,更好地表达局部特征的符号方向性,本文提出了方向注意力模块.

方向注意力模块由两个部分组成,即池化层和基于非对称卷积的十字方向特征强化层. 如图 4 所示,为了强化局部方向性信息,本文首先使用了平均池化和最大池化两种池化方法,生成了聚合的双通道拼接池化信息;其次,为了更好地强化注意力模块对局部空间信息中方向性的表达,本文采用一个卷积核大小为 3×3 的非对称卷积对池化层进行处理,生成方向注意力;最后,将方向注意力与输入特征相乘得到强化局部方向性表达的特征图,可表示为

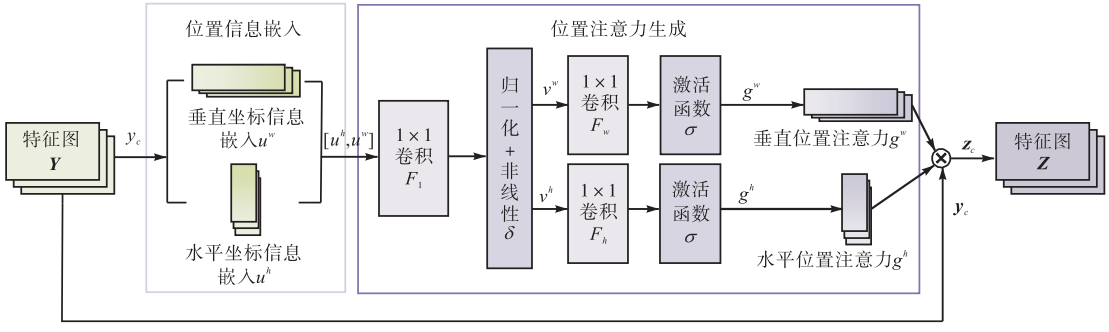
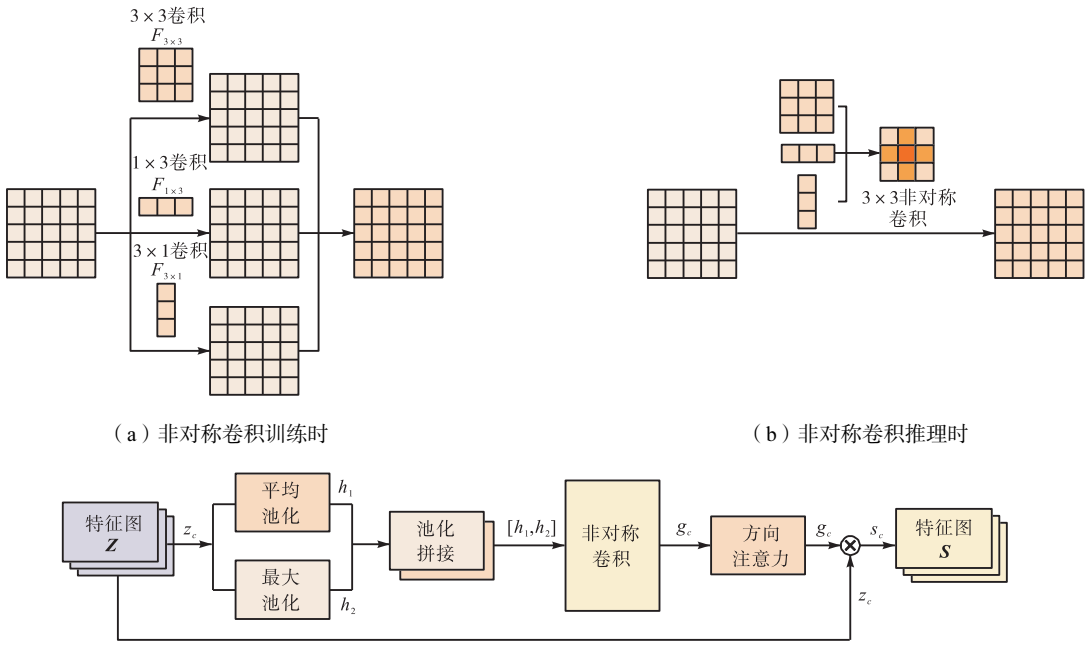


图 3 位置注意力模块

Fig.3 Coordinate attention module



（a）非对称卷积训练时　　（b）非对称卷积推理时

（c）方向注意力

图 4　方向注意力模块

Fig.4　Directional attention module

$$h_1(i,j)=\frac{1}{C}\sum_{0\leqslant c\leqslant C} z_c(i,j) \tag{8}$$

$$h_2(i,j)=\max_{0\leqslant c\leqslant C}\{z_c(i,j)\} \tag{9}$$

$$g_c=\sigma\left(F_{3\times3}([h_1,h_2])+F_{1\times3}([h_1,h_2])+F_{3\times1}([h_1,h_2])\right) \tag{10}$$

$$s_c(i,j)=z_c(i,j)\cdot g_c(i,j) \tag{11}$$

式中：$h_1(i,j)$ 表示位置为 (i,j) 处各通道平均池化结果；$h_2(i,j)$ 表示位置为 (i,j) 处各通道最大池化结果；$z_c(i,j)$ 表示第 c 通道位置为 (i,j) 处方向注意力模块输入特征；g_c 表示基于非对称卷积的强化局部方向性的方向注意力权重结果；$s_c(i,j)$ 表示第 c 通道位置为 (i,j) 处经过位置注意力处理后的特征.

2　实验设置

2.1　实验环境设置

本文实验使用 PyTorch 深度学习框架，服务器为 CPU Inter® Core i9-9900x（3.5 GHz），GPU Nvidia RTX1080Ti（11 GB），Ubuntu 16.04. 实验采用的是随机梯度下降法和 Adam 优化算法进行训练，实验学习率设置为固定值 1×10^{-4}，批量大小为 8，最大训练轮数为 300.

2.2　数据集与评价指标

由于现阶段没有公开的多音乐谱数据集，本文制作了一个单行多音乐谱数据集. 此数据包含 15 000 张单行多音五线谱图像以及通过“前进编码”方式编排对应图像的真值序列标签.

如图 5 所示，由于关注的焦点是乐谱符号，所以在编码过程中本文选择使用了最小的符号集，这个符号集足以准确地表示音高和时值信息，对乐谱的谱号、调号、拍号、小节线以及音符本身进行了编码，不关心连音、重音、力度等乐谱符号. 为了能够将多音乐谱编码为一维序列信息，本文采用了“前进编码”，将顺序出现的音符之间用“+”符号分离，同一时刻出现的多音音符从下到上对每个音符进行排列.

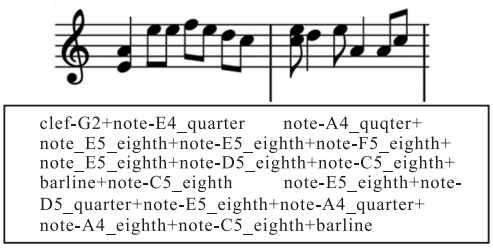


图 5　数据集示意

Fig.5　Dataset illustration

乐谱识别常用的评价指标为符号错误率（E_{ser}）和序列错误率（E_{er}）. 其中 E_{ser} 指的是将预测结果序列转换为真值标签结果所需要的基本编辑操作（插入、删除或替换操作）的数量占序列总数量的比例，用以评估模型预测的序列与真值标签序列之间的相似程度. 其计算公式为

$$E_{\mathrm{ser}}=\frac{\sum_{i=1}^{n}\left(I_i+D_i+S_i\right)}{\sum_{i=1}^{n}N_i} \tag{12}$$

式中：n 表示序列的总数目；I_i 表示第 i 个序列插入操作的次数；D_i 表示第 i 个序列删除操作的次数；S_i 表示第 i 个序列替换操作的次数；N_i 表示第 i 个序列的符号总个数.

E_{er} 表示所有序列中出现至少一处错误序列占总序列数的比例，其计算公式为

$$E_{\mathrm{er}}=\frac{\sum_{i=1}^{n}\Theta\left(I_i+D_i+S_i\right)>0}{n} \tag{13}$$

式中 Θ 表示指示函数，用于表示条件 $I_i+D_i+S_i>0$ 的成立情况，条件成立时返回 1，条件不成立时返回 0.

为了充分利用乐谱符号的二维特性并对模型结果进行评估，本文考虑使用音高符号错误率(pitch symbol error rate，PSER)、时值符号错误率(length symbol error rate，LSER)、音高序列错误率(pitch sequence error rate，PER)、时值序列错误率(length sequence error rate，LER)4 个评价指标来评估本文所使用的创新方法.

3　实验结果与分析

3.1　创新方法有效性分析

为了验证本文所提出的网络模型中各创新方案的有效性和必要性，本节以卷积递归神经网络为基准网络，通过对各创新模块进行消融实验对比分析本文模块的性能，并将本网络模型与其他乐谱识别方法进行对比，证明本文所提方法的有效性.

3.1.1　纵向位置编码分析

本小节对纵向编码方法的选择进行了实验对比分析，以便选择最适合多音乐谱识别的纵向位置编码方法，提高模型在特征提取阶段对纵向位置的学习表征能力. 文中对比了线性编码、余弦编码和通道编码 3 种纵向位置编码方法，测试不同编码方法对实验识别结果产生的影响，结果如表 1 所示.

表 1　不同纵向编码方法的识别错误率

Tab.1　Recognition error rates of different vertical encoding methods　　%

编码方法	符号错误率		序列错误率	
	时值	音高	时值	音高
基准方法	1.81	3.28	28.69	40.14
通道编码	1.70	3.19	25.66	40.25
余弦编码	1.69	3.12	26.98	37.61
线性编码	1.49	2.80	21.21	32.65

由表 1 可见，通过采用纵向位置编码强化模型对纵向位置表达的方法有助于模型对音符音高和时值信息的表达. 在 3 种位置编码方法中，采用通道编码和余弦编码的方法对模型效果影响不大，采用线性位置编码方法的网络模型效果最好，对音符时值和音符音高的识别符号错误率和序列错误率均有所下降. 与基准网络相比，采用线性编码方法音符时值符号错误率下降了 0.32%，音符音高符号错误率下降了 0.48%，音符时值序列错误率下降了 7.48%，音符音高序列错误率下降了 7.49%. 实验结果表明，线性编码方法相比于余弦编码方法和通道编码方法更适用于乐谱识别任务.

3.1.2　位置注意力模块分析

为了证明位置注意力模块的有效性，本文首先将位置注意力模块作用前后的特征图进行了可视化，结果如图 6 所示.

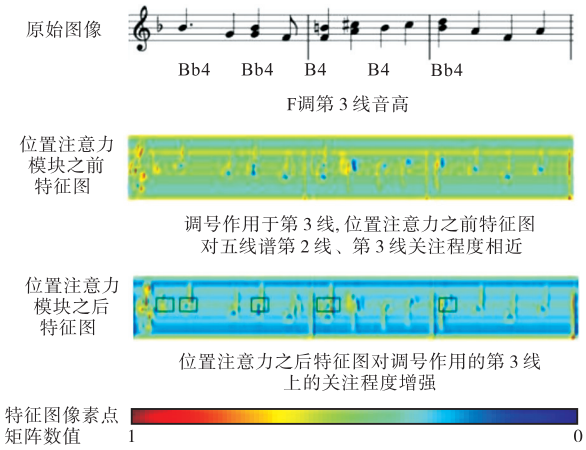


图 6　位置注意力特征图可视化结果

Fig.6　Visualization results of coordinate attention feature maps

由图 6 可见，原始乐谱图像中乐谱调号 F 大调中的降调符号 b 作用于整行五线谱的第 3 线，第 2 小节中第 3 线上的还原记号♮作用于第 2 小节内. 通过位置注意力前后特征图可视化结果可以发现，经过位置注意力模块之后的特征图对调号作用的第 3 线，尤其是对第 3 线上受调号和第 2 小节还原记号影响的符号位置关注程度增强，对背景的关注程度减弱. 特别地，位置注意力还能够表征两次复杂的降调复原转换关系：调号中的降调与第 2 小节 B4 中复原的转换，及复原符号作用消失后第 3 小节 Bb4 恢复降调. 说明位置注意力模块有助于增强调号位置信息及相关符号之间复杂关系的表征.

其次，本文将具有能够强化位置信息与符号特征相关性的位置注意力模块与仅能够进行通道间特征信息交互的两种经典的通道注意力方法进行对比分

析，证明强化空间位置与符号特征的依赖关系对多音乐谱识别的有效性. 本文选取了 SE[22]、ECA[23]两种典型的通道注意力模块与位置注意力模块作用效果进行对比，进一步比较不同注意力模块对实验识别结果产生的影响，结果如表 2 所示.

表 2　不同通道注意力模块的识别错误率

Tab.2　Recognition error rates of different channel attention modules　%

通道注意力	符号错误率		序列错误率	
	时值	音高	时值	音高
基准方法	1.81	3.28	28.69	40.14
SE	1.50	3.00	21.61	34.24
ECA	1.55	2.84	23.81	33.31
位置注意力	1.47	2.50	20.42	29.04

由表 2 可见，使用不同通道注意力方法，模型效果均有所提升，说明对通道间特征信息进行交互，学习不同通道间信息相关性、保留关键通道信息的方法对多音光学乐谱识别有效. 其中，使用位置注意力模块的方法相比于经典通道注意力的方法在对音符音高识别任务中有了进一步的提升，与使用 SE 方法相比使用位置注意力方法音符音高符号识别错误率下降了 0.50%，音高序列错误率下降了 5.20%；与使用 ECA 方法相比使用位置注意力方法音符音高符号识别错误率下降了 0.34%，音高序列错误率下降了 4.27%. 进一步表明通过强化空间与特征相关性表达，对多音乐谱特征和空间坐标的依赖关系进行显示建模，能够更好地捕捉位置信息与符号的相关性，提升多音乐谱音高识别的准确性.

3.1.3　方向注意力模块分析

为了证明方向注意力模块的有效性，首先将方向注意力模块作用前后的特征图进行了可视化，结果如图 7 所示.

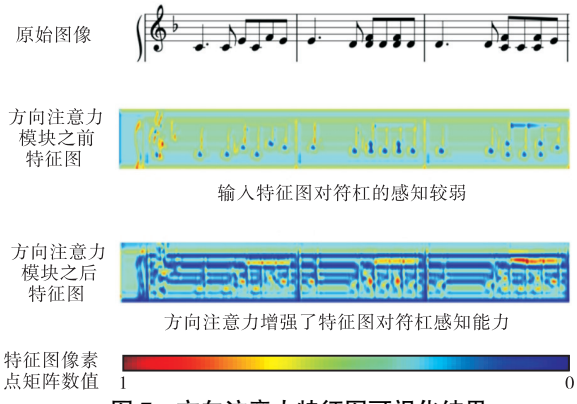


图 7　方向注意力特征图可视化结果

Fig.7　Visualization results of direction attention feature maps

由图 7 可见，乐谱中方向注意力前的特征图对音符的符杠部分的关注程度较弱，而符杠与音符的时值语义密切相关，符杠信息捕捉不准会造成音符时值识别错误. 经过方向注意力之后的特征图明显增强了对音符的符头和符杠信息的感知能力，说明方向注意力模块有助于对符号内符头和符杠特征的提取.

其次，本文通过将方向注意力模块与空间注意力模块[24]进行对比，进一步证明强化空间方向性表达对多音乐谱识别的有效性，实验结果如表 3 所示.

表 3　不同空间注意力模块的识别错误率

Tab.3　Recognition error rates of different spatial attention modules　%

空间注意力	符号错误率		序列错误率	
	时值	音高	时值	音高
基准方法	1.81	3.28	28.69	40.14
标准空间注意力	1.39	2.96	24.82	35.53
方向注意力	1.30	2.68	21.17	32.53

由表 3 可见，通过使用空间注意力，对空间局部信息进行强化使得对音符时值识别错误率明显降低. 与基准方法相比，使用标准空间注意力方法时值符号错误率下降了 0.42%，时值序列错误率下降了 3.87%. 通过使用方向强化的方向注意力模块，符号在音符时值和音符音高识别准确性均有所提高，与标准空间注意力方法相比，使用方向注意力模块音符时值符号错误率下降了 0.09%，但时值序列错误率下降了 3.65%，音高符号错误率下降了 0.28%，音高序列错误率下降了 3.00%，进一步说明通过强化局部特征的方向性能够提高模型对多音序列分布的感知能力，更好地识别多音符号的分布方向性特点.

3.2　消融实验分析

本文通过一系列消融实验的比较分析来验证本文所提网络的有效性. 本节将本文提出的纵向位置编码方法、位置注意力模块和方向注意力模块应用于基础模型框架上，对比不同方法下的实验结果，从而进一步验证本文所提方法的有效性，消融实验识别错误率如表 4 所示.

表 4　模型消融实验识别错误率

Tab.4　Recognition error rates from model ablation experiments　%

消融方法	符号错误率		序列错误率	
	时值	音高	时值	音高
基准方法	1.81	3.28	28.69	40.14
线性编码	1.49	2.80	21.21	32.65
位置注意力	1.47	2.50	20.42	29.04
方向注意力	1.30	2.68	21.17	32.53
线性编码+位置注意力	1.34	2.47	19.61	28.44
线性编码+方向注意力	1.18	2.52	20.99	31.20
位置注意力+方向注意力	1.28	2.40	18.70	27.55
线性编码+位置注意力+方向注意力（本文方法）	1.14	2.14	18.58	27.42

由表 4 可见,使用线性编码、位置注意力和方向注意力 3 种方法中的任意一种方法,音符时值识别符号错误率和音高识别符号错误率均有所降低,并且音符时值和音高识别的序列错误率明显降低,说明本文所提方法使一部分原本存在少量识别错误的乐谱图像不再出现识别问题,改善了模型对音符的识别效果. 通过加入单一创新方案的方法对比可以发现,加入位置注意力的方法对音高识别符号错误率和序列错误率改善显著;加入方向注意力的方法对时值识别符号错误率改善显著. 进一步对使用两种创新方案的方法进行实验,发现相比于使用单一创新方案的方法使用两种创新方案的方法识别效果更好,其中,线性编码方法有效利用了图像中的纵向位置信息,改善了模型对音符的识别效果;位置注意力方法通过提取位置信息与符号特征的依赖关系能够更好地挖掘变音记号与符号音高的作用关系,有效提升符号音高识别准确率;方向注意力模块,通过增强模型对音符局部方向性的捕捉能力,有效改善了多音乐谱时值识别的准确率. 最后,实验证明使用 3 种创新模块的本文方法与使用任意两种创新模块的方法相比,音符时值识别和音高识别的准确率进一步提升,有效改善了多音乐谱识别的准确率.

3.3 与其他方法对比分析

本节将本文提出的网络与其他端到端乐谱识别方法 CRNN-lite[14]、R2-CRNN[13]、RNNDecoder[15]和 TrOMR[16]进行对比. 由于以上方法评价指标中只考虑了符号错误率,并没有对序列错误率进行衡量,因此,本节只采用时值符号错误率和音高符号错误率作为评价指标. 模型对比实验识别错误率如表 5 所示.

表 5 模型对比实验识别错误率

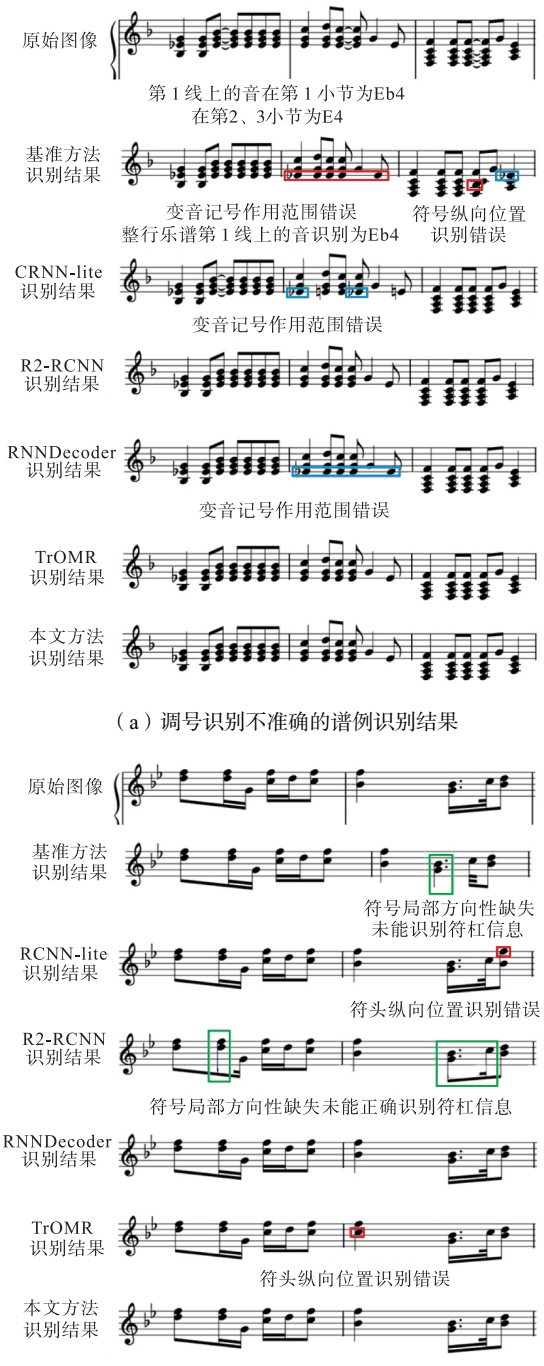
Tab.5 Recognition error rates from model comparison experiments %

对比方法	符号错误率	
	时值	音高
基准方法	1.81	3.28
CRNN-lite	2.67	5.67
R2-CRNN	3.00	8.55
RNNDecoder	1.53	2.73
TrOMR	1.24	2.54
本文方法	1.14	2.14

由表 5 可见,本文方法与 CRNN-lite 方法相比,时值符号错误率下降了 1.53%,音高符号错误率下降了 3.53%;本文方法与 R2-CRNN 方法相比,时值符号错误率下降了 1.86%,音高符号错误率下降了 6.41%;本文方法与 RNNDecoder 方法相比,时值符号错误率下降了 0.39%,音高符号错误率下降了 0.59%;与 TrOMR 方法相比,时值符号错误率下降了 0.10%,音高符号错误率下降了 0.40%,在音高识别准确率上有较为明显的提升,达到了很好的效果.

3.4 实验结果可视化展示

为了直观显示模型的识别效果,本节将识别出的音符序列展示于五线谱上,如图 8 所示,图中用方框标识出了识别错误的部分.



(a) 调号识别不准确的谱例识别结果

(b) 符杠识别不准确的谱例识别结果

图 8 识别结果可视化图像

Fig.8 Visualization images of recognition results

图 8 中红框显示了由于纵向位置表达不准造成

的音符临近音高识别错误的情况，图 8(a)中基准方法将原和弦符号识别成了转位和弦符号，出现了符头方向的变化. 蓝框显示了由于对符号与空间位置相关性表达不足临时变音记号作用范围错误的情况，临时变音记号作用范围为符号出现的同一小节内，图 8(a)中基准方法将第 1 小节的临时变音记号的作用范围识别为整行乐谱，使第 2、3 小节的 E4 错误识别为 Eb4；CRNN-lite 方法对第 1 小节的临时变音记号作用范围不准确，使第 2 小节中部分 E4 错误识别为 Eb4；RNNDecoder 方法将第 1 小节的临时变音记号的作用范围识别为前两小节乐谱，使第 2 小节的 E4 错误识别为 Eb4. 绿框显示了由于对局部符号方向性捕捉缺失造成的符号时值识别错误的现象，图 8(b)中基准方法未能识别符杠信息，使符号时值识别错误，将十六分音符识别为四分音符；R2-CRNN 方法对符杠的作用方向和数目识别不准确，误将第 1 小节的十六分音符识别为八分音符，将第 2 小节的十六分音符和三十二分音符分别识别为八分音符和十六分音符. 由其他方法和本文提出方法结果对比可以看出，本文提出的方法有效改善了基准方法中对局部位置相近的符头音高识别不准确、变音记号和符号作用关系识别错误、多音符号音符时值识别错误的问题，进一步证明了本文提出方法的有效性.

4　结　语

本文针对多音乐谱识别任务中传统方法存在的临近音高识别不准、音符变音记号作用范围识别不佳、音符局部方向性表达不清的问题，提出了一种强化音符位置及方向先验信息的多音乐谱识别方法，此方法能够有效改善上述问题，降低多音光学乐谱的识别错误率. 首先为了改善临近音高识别不准的问题，本文提出了纵向位置编码的方法，通过对图像进行纵向位置编码的方法增强了模型在特征提取的过程中对特征纵向位置的感知能力. 其次针对音符变音记号作用范围识别不佳的问题，提出了使用位置注意力的方法，对位置信息和符号的依赖关系进行提取. 最后提出方向注意力模块，通过关注特征的局部方向性关系，有效对音符的符头与符尾、符杠关系进行提取.

实验数据表明，本文提出的强化音符位置及方向先验信息的方法具有识别准确率高的特点. 本模型在多音乐谱数据集上音符时值识别错误率为 1.14%，音符音高错误率为 2.14%，同基准模型相比音符时值错误率降低了 0.67%，音符音高错误率降低了 1.14%，更加适用于多音乐谱识别.

参考文献：

[1] Calvo-Zaragoza J，Jan H Jr，Pacha A. Understanding optical music recognition[J]. ACM Computing Surveys(CSUR)，2020，53(4)：1-35.

[2] Calvo-Zaragoza J，Martinez-Sevilla J C，Penarrubia C，et al. Optical music recognition：Recent advances，current challenges，and future directions[C]// International Conference on Document Analysis and Recognition. San José，USA，2023：94-104.

[3] Shatri E，Fazekas G. Optical music recognition：State of the art and major challenges[EB/OL]. https：//arxiv.org/abs/2006. 07885，2020-06-14.

[4] 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.

[5] Huang Z Q，Jia X，Guo Y F. State-of-the-art model for music object recognition with deep learning[J]. Applied Sciences，2019，9(13)：2645.

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

[7] Baró A，Riba P，Calvo-Zaragoza J，et al. From optical music recognition to handwritten music recognition：A baseline[J]. Pattern Recognition Letters，2019，123：1-8.

[8] 黄志清，贾　翔，郭一帆，等. 基于深度学习的端到端乐谱音符识别[J]. 天津大学学报(自然科学与工程技术版)，2020，53(6)：653-660.
Huang Zhiqing，Jia Xiang，Guo Yifan，et al. End-to-end music note recognition based on deep learning[J]. Journal of Tianjin University(Science and Technology)，2020，53(6)：653-660(in Chinese).

[9] Alfaro-Contreras M，Ríos-Vila A，Valero-Mas J J，et al. Decoupling music notation to improve end-to-end optical music recognition[J]. Pattern Recognition Letters，2022，158：157-163.

[10] Alfaro-Contreras M，Valero-Mas J J. Exploiting the two-dimensional nature of agnostic music notation for neural optical music recognition[J]. Applied Sciences，2021，11(8)：3621.

[11] Ríos-Vila A，Calvo-Zaragoza J，Iñesta J M. Exploring the two-dimensional nature of music notation for score recognition with end-to-end approaches[C]//2020 17th

International Conference on Frontiers in Handwriting Recognition (ICFHR). Dortmund, Germany, 2020: 193-198.

[12] He R C, Yao J F. End-to-end optical music recognition with attention mechanism and memory units optimization[C]//Chinese Conference on Pattern Recognition and Computer Vision(PRCV). Xiamen, China, 2023: 400-411.

[13] Liu A Z, Zhang L P, Mei Y Q, et al. Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores[EB/OL]. https://arxiv.org/abs/2010.13418, 2020-10-26.

[14] 蒋凌云, 鞠金恒, 徐 佳, 等. 一种基于改进 CRNN 的轻量化乐谱识别方法[J]. 电子学报, 2023, 51(11): 3167-3175.
Jiang Lingyun, Ju Jinheng, Xu Jia, et al. A lightweight music recognition method based on improved CRNN[J]. Acta Electronica Sinica, 2023, 51(11): 3167-3175(in Chinese).

[15] Edirisooriya S, Dong H W, McAuley J, et al. An empirical evaluation of end-to-end polyphonic optical music recognition[EB/OL]. https://arxiv.org/abs/2108.01769, 2021-08-03.

[16] Li Y X, Liu H P, Jin Q, et al. TrOMR: Transformer-based polyphonic optical music recognition[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Rhodes Island, Greece, 2023: 1-5.

[17] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 13713-13722.

[18] Calvo-Zaragoza J, Rizo D. End-to-end neural optical music recognition of monophonic scores[J]. Applied Sciences, 2018, 8(4): 606.

[19] Wang S X, Wang X, Wang S M, et al. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting[J]. International Journal of Electrical Power & Energy Systems, 2019, 109: 470-479.

[20] Graves A. Studies in Computational Intelligence[M]. Berlin: Springer-Verlag, 2012: 61-93.

[21] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the CoordConv solution[EB/OL]. https: //arxiv. org/abs/1807. 03247, 2018-06-09.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141.

[23] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11534-11542.

[24] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV). Munich, Germany, 2018: 3-19.

(责任编辑: 孙立华)