技术报告：基于复调光学音乐识别的五线谱-简谱翻译任务

# 摘要

光学音乐识别（Optical Music Recognition, OMR）是一项受到广泛关注的技术，旨在将乐谱图片转化为机器可读的书面乐谱。本文介绍了一种简单而高效的方法，旨在将五线谱翻译成简谱，以帮助那些更容易阅读简谱而不是五线谱的人。该方法包括两个主要阶段：通过逐行处理的模板匹配方法获取每个音符的位置和识别特定图片，然后对这些图片进行分类以获取节奏类型。该方法具有较好的效果和良好的泛化能力，适用于光学音乐识别中的其他符号。。所有的源代码可以 [https://github.com/xuxiran/translate-staff-to-simple-musical-notation](https://github.com/xiranxu/standard_notation_to_numbered_musical_notation%20) 被获取。

# 关键词

复调、光学音乐识别、模板匹配、五线谱、简谱

# 介绍

五线谱是一种广泛使用的记谱方式，但在中国，有大量的音乐爱好者更喜欢简谱。因此，本文旨在为这些爱好者设计一种五线谱翻译成简谱的软件，以解决手动转换的麻烦。该软件的输入和输出均为PDF格式，并将简谱翻译结果放置在每行五线谱下方，示例见图1。 

图1 五线谱翻译软件的输入和输出。原始图片来源自<https://musescore.com/thatonepiano/scores/1787706>

从直觉上来看，将五线谱翻译成简谱似乎是光学音乐识别的后续工作。然而，现有的光学音乐识别软件无法满足高准确率的五线谱翻译简谱的需求，这是因为识别准确率较低所导致的。以前的研究主要使用了两种方法来实现乐谱识别，一种是将其处理成两个阶段的任务，先分割每个音乐符然后识别它们之间的关系[5-9]，另一种则是使用机器学习的端到端方法。

但是在目前的数据量下，作者认为我们还没有办法实现一种跨平台的端到端方法。即使在单个平台上拥有足够的数据，也很容易局限于仅能识别该单一平台下的谱面，这也是神经网络的普遍问题。因此，在本文中，我们依然采用了两个阶段识别方法，但是我们的识别方法和以前的研究略有不同，是一种改良的两个阶段识别方法：逐行模板匹配+分类算法。

逐行模板匹配和两个阶段任务的第一个阶段相似，但我们不会分割每个音乐符号。相反，我们将五线谱的每一行和每一间分隔开来，在相应的行或间上查找单个音符。这将多声部音乐的多音识别问题转变为单音识别问题。同时，通过定位五线谱的位置[??]，我们能够确定需要进行模板匹配的具体行图像。算法流程如图2所示。

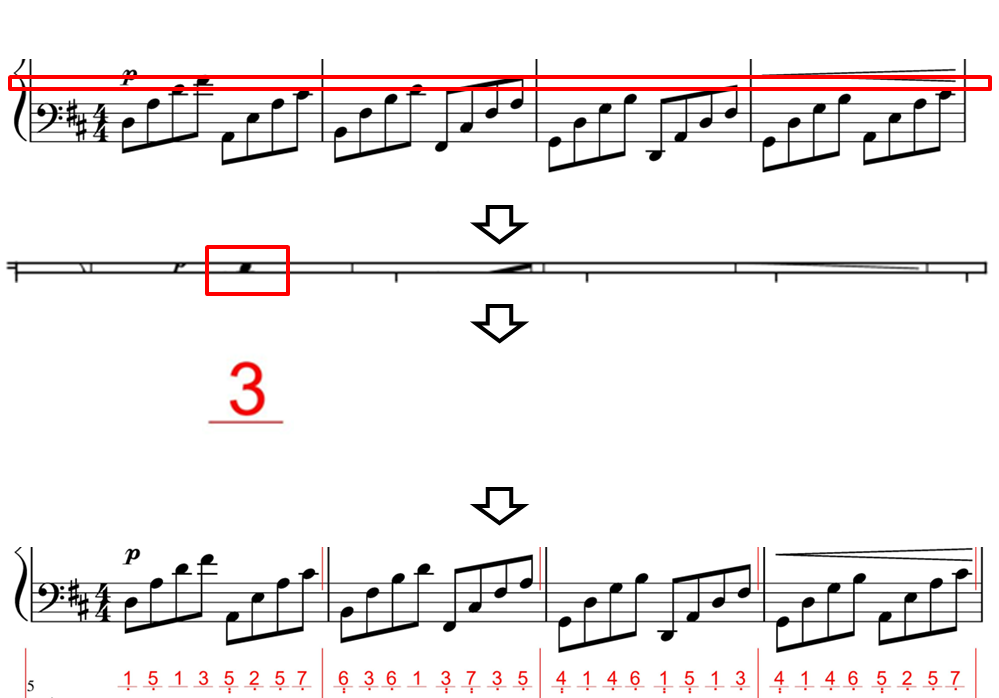


图2显示了逐行模板匹配算法的流程图。 从图中可以看出，算法从上到下逐行扫描，每次扫描宽度为五线谱两条线的一半。如果在扫描期间发现目标，则标识出其对应的简谱翻译结果。对于每行五线谱，我们共进行了19次这样的操作，从上到下扫描。如果需要更广泛的音域，则可以进一步增加扫描次数，但对于我们的任务而言，19次已经足够。

接下来，我们根据每个识别到的音符，确定其节奏型。我们截取包含音符的图像，并将其送入我们已经训练好的分类模型中（在方法中详细描述），以确定该音符的节奏型。

通过逐行模板匹配+分类，我们几乎可以找到所有的音符，不会有多、错、漏的情况发生。同时，将音符所在的图像送入神经网络也大大提高了其识别准确率。通过将端到端任务拆分为两个步骤，更好地发挥了机器学习算法的能力，类似于视频理解中使用光流抽取的思想。

经过实验，我们通过逐行模板匹配+分类的算法，成功识别出谱面中黑色实心单元（即四分音符、八分音符、十六分音符和三十二分音符），并且通过直接分类，还可以识别该行五线谱的调式和谱号。如图1和图2所示，这既展示了我们的思路，也展示了我们的有效识别结果。

# 方法

这里我们以黑色实心的单元为例，详细介绍我们是如何进行逐行模板匹配以及分类的。

## 确定每一行五线长度

我们首先确定了五线谱的每一根线。这意味着每一行乐谱我们将识别出五根弦。识别方法和前人类似[??]。接着我们关注某个特定行的五线谱，其中包括五条线在图中的高度。比如，对于图1这一张高度为7017，宽度为4959的图片，通过算法确定其第一行五线的位置分别是(549,583,618,652,687)。确定第一行五线的结果如图3(a)所示。为了后续更好的识别，我们将图片中识别的五线全部删除，删除后的图片如图3(b)所示。

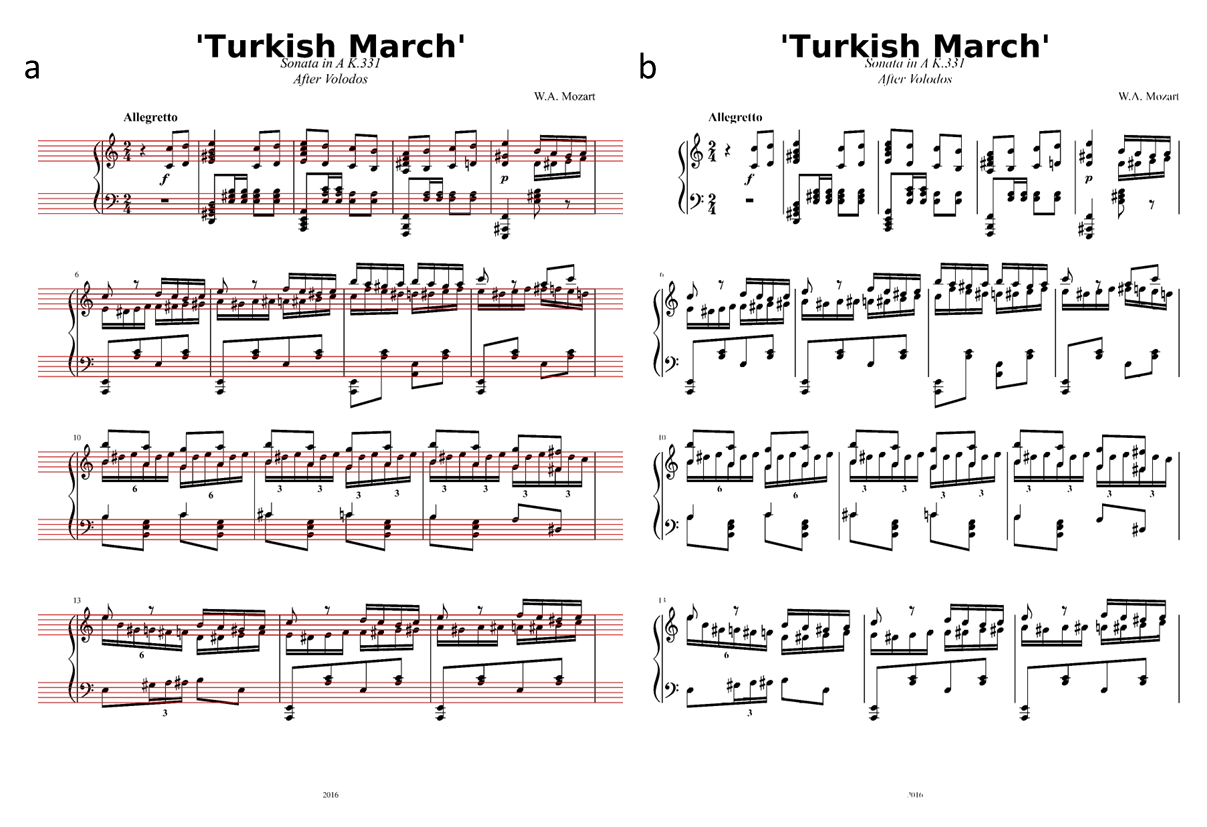


图3 预处理示意图。a，确定五线的位置；b，删除五线后的图片

## 逐行模板匹配

在2.1节中，我们已获得一张不包含五线谱的图片，如图3b所示。在本节，我们逐行截取了五线谱。以第一行为例，如图4a所示。 我们将该行五线谱分为20行（20行可满足大部分翻译需求，未来可继续扩展）。如图4b所示，红色表示我们关注的一行，每个音符占两行。该“两行”的长度被定义为“height”，例如在图3所示的谱面中，“height”为43。由于扩展后的图片高度为8个height，而五线谱高度为4个height，因此图片被称为“两倍高度-五线谱”。 该分割方法可识别19种音符，包括高音（d3-g）和低音（f1-B1），如图4c所示，我们扫描这些音符的位置，并截取图4d所示的区域。 在进行“模板匹配”前，需要有“模板”。我们从标准五线谱中截取了两种经典黑色实心单元作为模板，两个图片大小均为873×440。实际运算时，我们读取这两个模板并相加，通过cv2.resize()将高度调整为“height”，宽度等比例调整。在本例中，模板大小调整为(43,85)，如图4g所示。

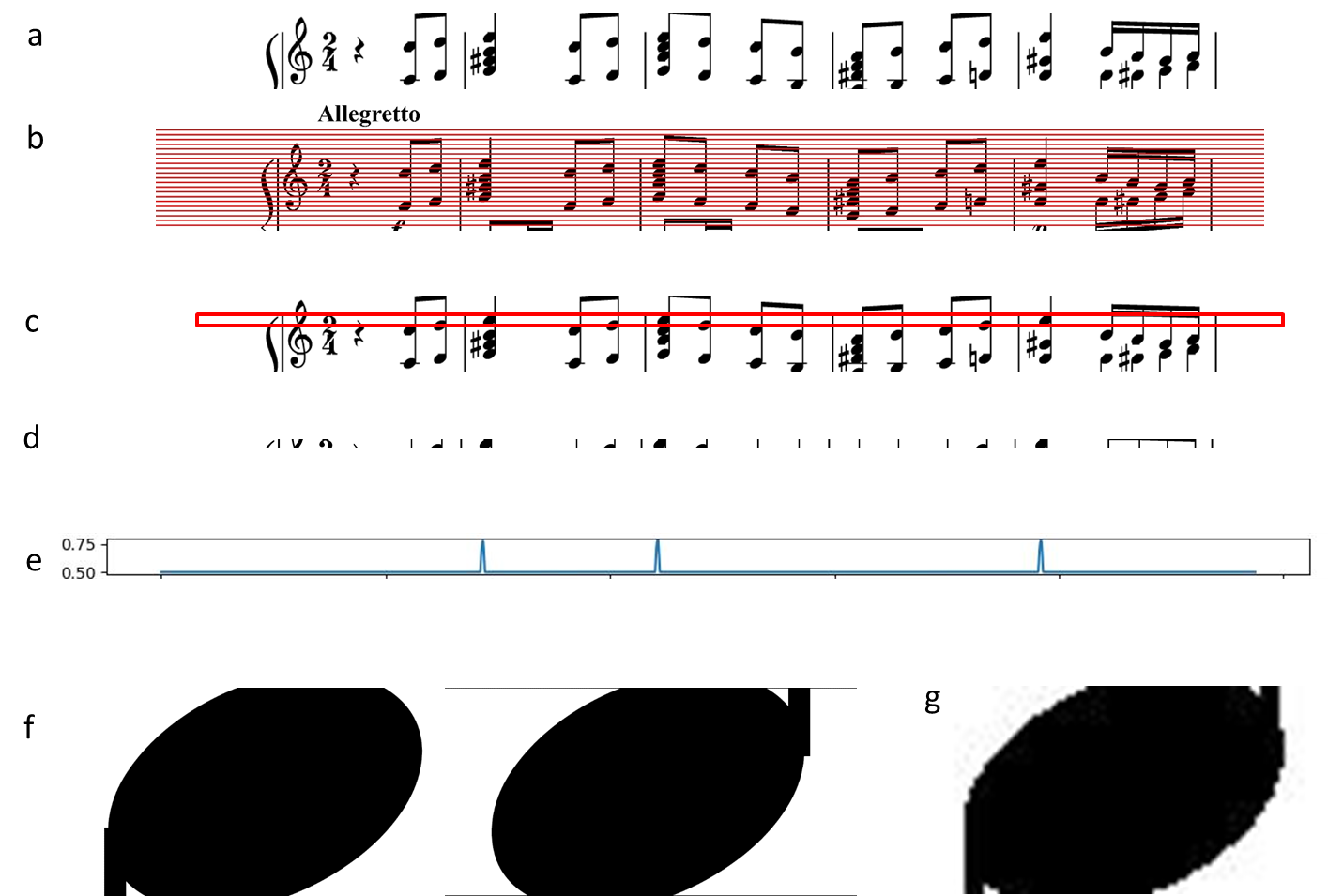


图4 逐行模板匹配。a，一行不被去除了五线区域的五线谱。b，

有了模板，我们就可以进行模板匹配了。图4e显示了模板匹配的结果。小于0.5的值被置于0.5以方便观看。很容易发现，三个峰值非常明显。这三个音全部都是e2，也就是简谱的高音3。

这个想法非常简单，但是效果非常的好。我们可以很容易检测到谱面所有的黑色实心单元。仅仅通过这一很简单的操作，我们只需要简单的判断一下所有黑色单元的位置，将同一列的黑色单元合并打印出来即可。

通过设计不同的模板，我们也可以检测出二分音符、全音符、升号、降号等。这里不再一一介绍。

## 分类

### 分类思路

虽然检测到了黑色实心单元，但是此时我们并不能区分四分音符、八分音符、十六分音符和三十二分音符，因为这些音符不能通过黑色实心单元区分。相反，需要更大的区域才能确定具体的类别。

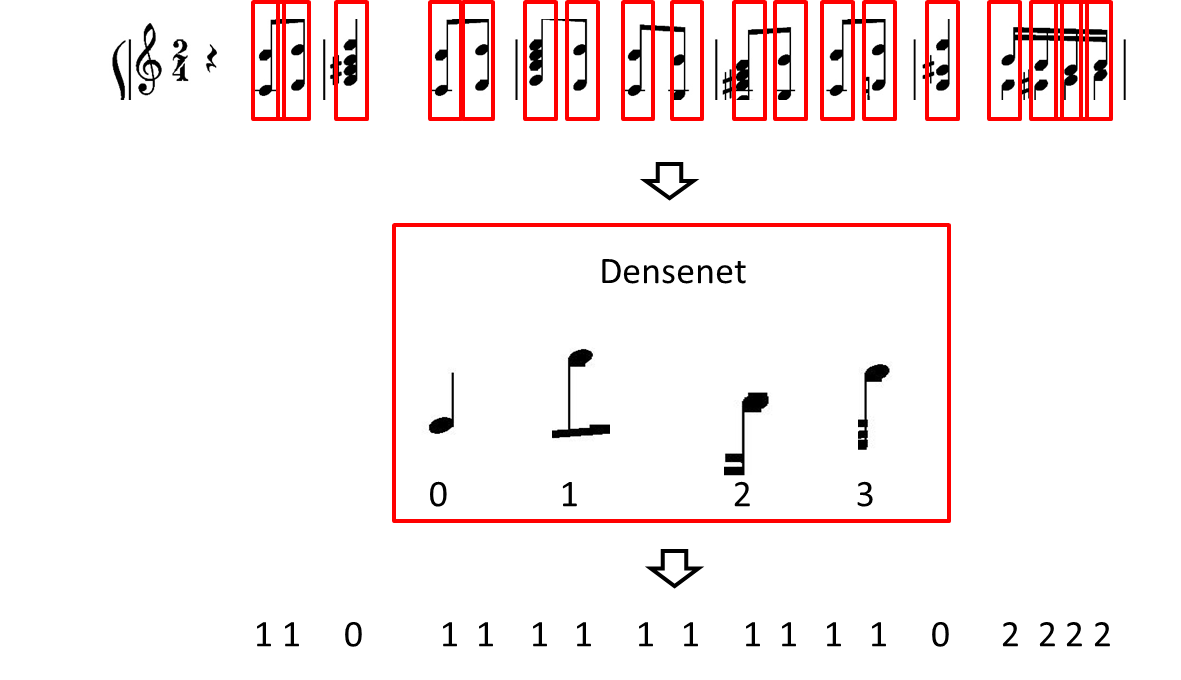


图5 分类流程图。首先根据模板匹配确定每个子图。每个子图在每一行乐谱中有着完全相同的高和宽。将子图放入神经网络获得四分类的结果，再将其绘制进入简谱，就可以获得最终的结果。

由于已经经过模板匹配，我们可以很轻松获得子图的位置。将子图送入已经训练完毕的Densenet中，就可以获得分类的结果。我们将在图6 中描述我们训练的网络。使用的densenet有四个模块，每个模块不断级联。模型完整的可视化见附件pdf。

### 获取数据

为了获取训练模型的数据，我们使用了primus数据库，尽管它只有单音数据，而我们的目标是复调音乐。使用2.2节中提到的模板匹配，可以获取特定的黑色实心单元的位置，接着截取特定的黑色实心单元，再从图片对应的“.semantic”文件中获取该黑色实心单元所对应的节奏型即可。对总共87689张图片，我们共获得了约95万张子图。这些子图构成了我们的数据库。每张子图都是以检测到黑色实心单元的位置向左0.5个height，向右2.5个height。（左右不相等是因为模板匹配到的像素对应黑色实心圆的最左侧）子图的高度将覆盖完整的“两倍高度-五线谱”的高度。由于primus数据库中height均为18，因此我们最终的图片高度和宽度将分别是180\*54。

图6a-图6d分别是获得的四种节奏类型的数据库。

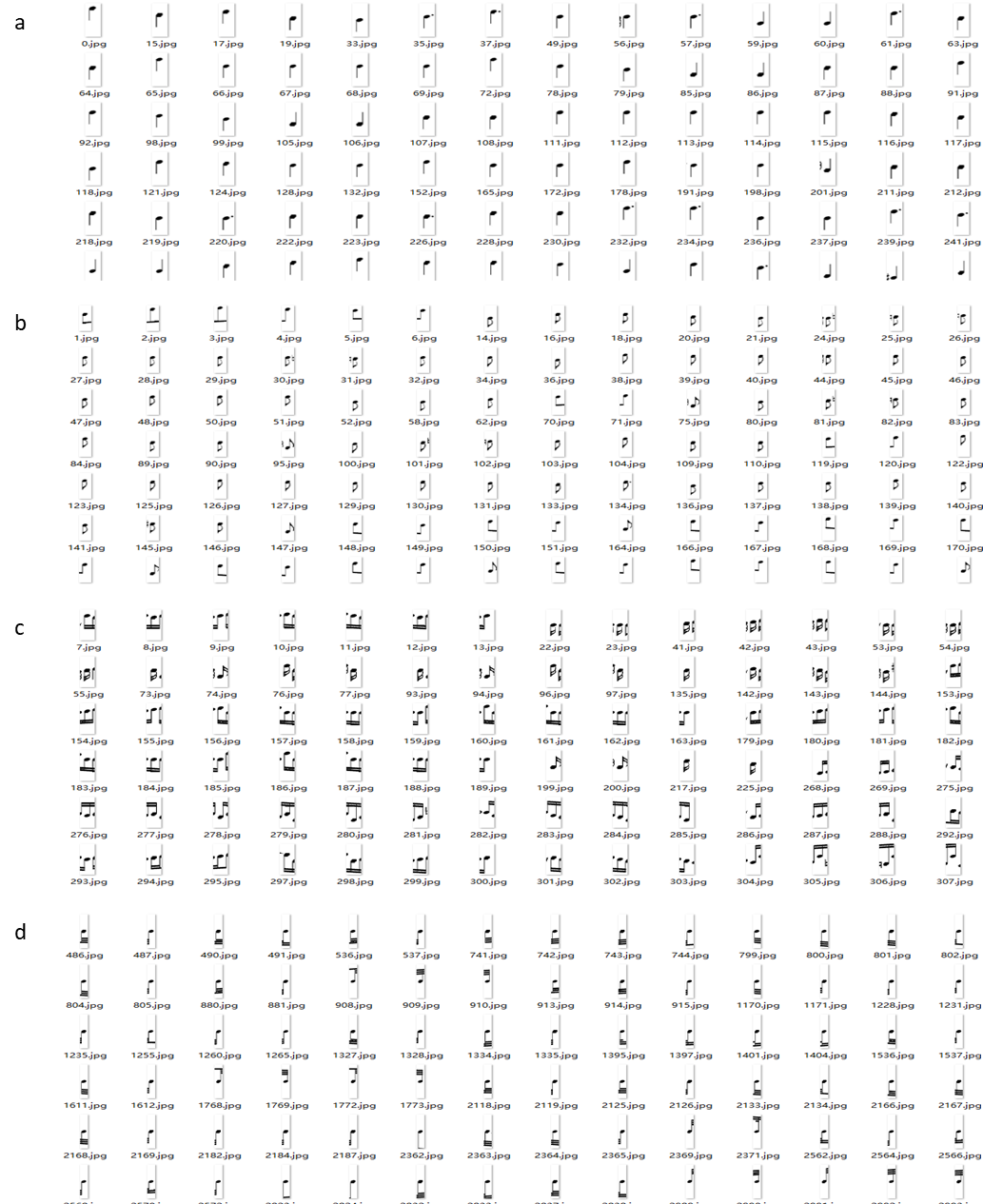


图6 四种数据库

### 数据增强

经过实验，使用原始数据训练出的模型对实际谱面中出现的各种元素非常敏感。因此，为了增强模型鲁棒性，我们对原始数据进行了数据增强。不同于经典的CV，我们的数据增强主要在增加一些无规则的线段、弧线以及一些五线谱中常常出现的元素（比如升号、降号、还原记号等等）。这些元素的位置、大小在特定范围随机，具体范围可以关注代码。

除了无规则添加一些元素之外，我们沿着该图的中线处随机添加一些黑色实心元素。这样确保模型在复调音乐的场景也可以获得高准确率。数据增强后的图片如图7所示

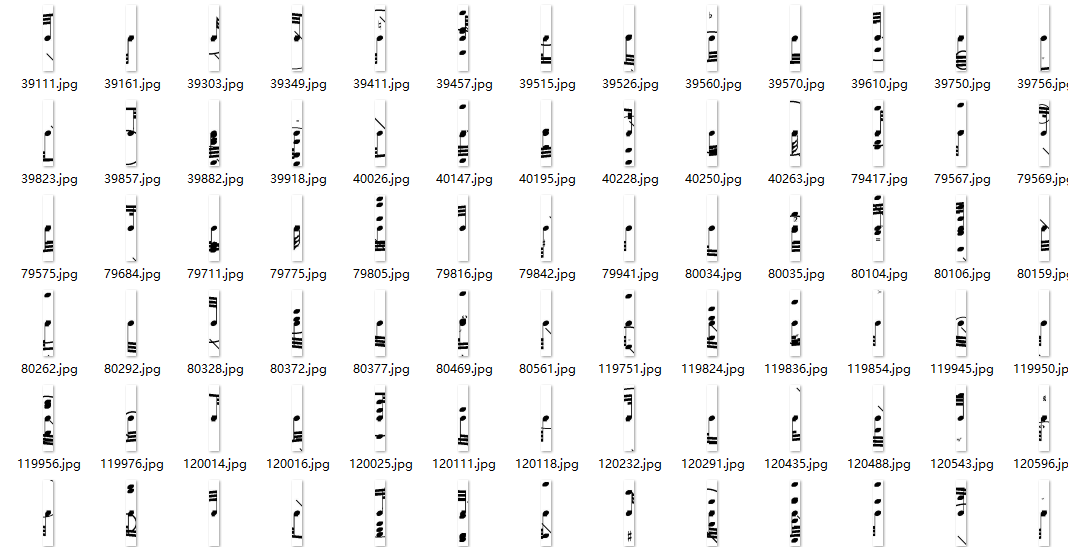


图7 数据增强后的数据库

### 模型

我们使用了densenet模型来完成这一任务。我们也使用了res50尝试完成这一任务，但是效果不如densenet模型。模型总览图见附件。

## 调式和谱号

为了实现五线谱到简谱的翻译，我们还需要知道这里的调式和谱号。我们直接将primus数据库中的一张图片作为输入。同样使用densenet架构。为了获得数据集，我们仍然在primus数据库上进行加工，图8显示了最典型的谱号（G2和F4），图9显示了最典型的4种调性（C大调、D大调）。由于数据分布过于不均匀，换言之一些谱号数量实在太少，对谱号我们只保留了C1、F4、G2三种谱号。而调性也只保留了A、bA、bB、C、D、E、bE、F、G九种常用调性。

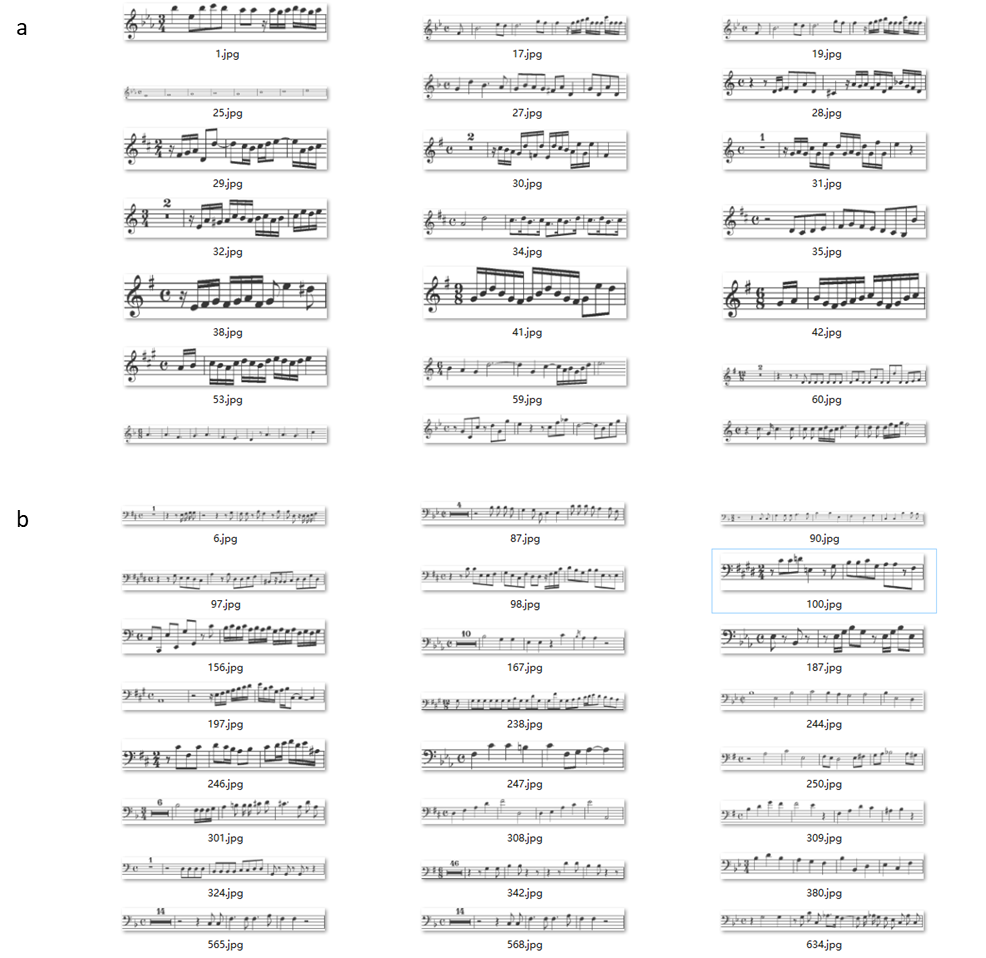


图8 示例的两种谱号，分别是高音谱号和低音谱号

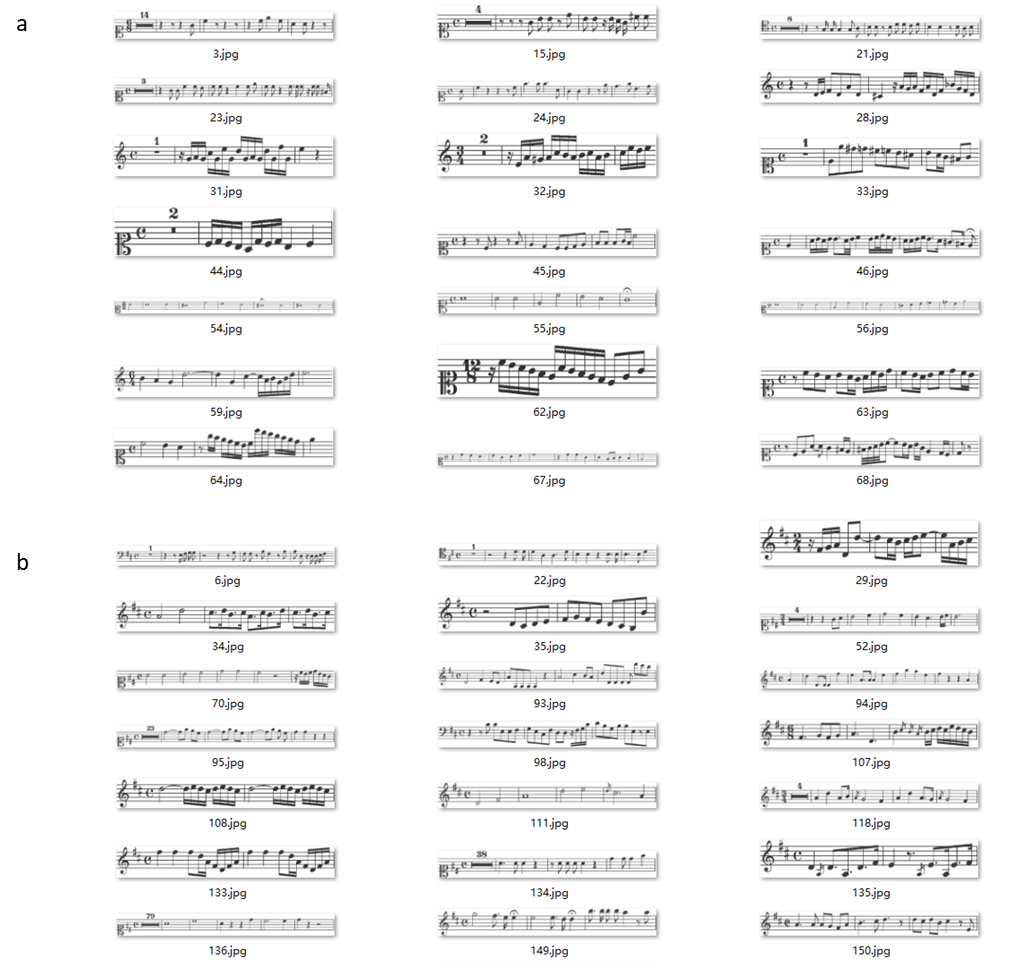


图9 示例的两种调式，分别是C大调和D大调

# 结果和讨论

我们将数据按8：2分为了训练集和验证集。对数据增强后的节奏的识别准确率在验证集上达到98.41%。

对谱号和对调式的识别准确率并没有达到可以实用的程度，还需要进一步整理。

本文作为一篇技术报告，提出了一种新的两阶段谱面识别的思路，即逐行模板匹配再过分类模型。对待一些较小的元素，比如音符，用2行足够。对待一些占据行数更多的元素，比如升号、降号、还原号等等，可以使用4行。接着，对更大的元素可以用更多的行数。

总之，本文实现了一种两阶段的识别方法，所有代码已经开源，并且整合为exe格式，无需安装任何包即可运行。