北京航空航天大學
BEIHANG UNIVERSITY

第二十六届“冯如杯”学生学术科技作品竞赛项目论文

基于卷积神经网络的五线谱识别应用

摘要

五线谱是音乐的重要载体，然而熟练识谱的能力却需要较强的专业背景，这就在业余音乐爱好者和许多优秀音乐作品之间产生了一道障碍；卷积神经网络是新兴的深度学习技术，在图像识别领域表现空前优异。本文描述了一个识别五线谱的移动应用，它以卷积网络作为核心算法，辅助音乐爱好者进行五线谱的辨识。

文中分析了识谱需要解决的基本问题，提出了自主设计的音符分类卷积网络 noteNet-1；对监督式学习模型之下大量训练样本的获取问题，创新采用了自动生成仿真样本的解决方案，取代了低效的人工收集、标注的传统方法。

关键词：卷积神经网络 五线谱识别

Abstract

Stave is a kind of musical note, a format in which a lot of musical productions are recorded. However, it’s very hard to master it without professional training, which formed an obstacle between music hobbyists and the nice productions.

Convolutional neural network, CNN, is a creative deep learning method, which turns out to be amazingly efficient on image recognization in recent few years. This paper describes an mobile application based on CNN, which helps the music amateurs recognize the staff.

The article gives a series of analyzation of the basic problems about recognizing a stave in digital image, mentions a newly designed CNN architecture for recognizing musical notes, together with a creative method of automatically generating training samples.

Key words: CNN, stave recognization

目录

第一章 引言

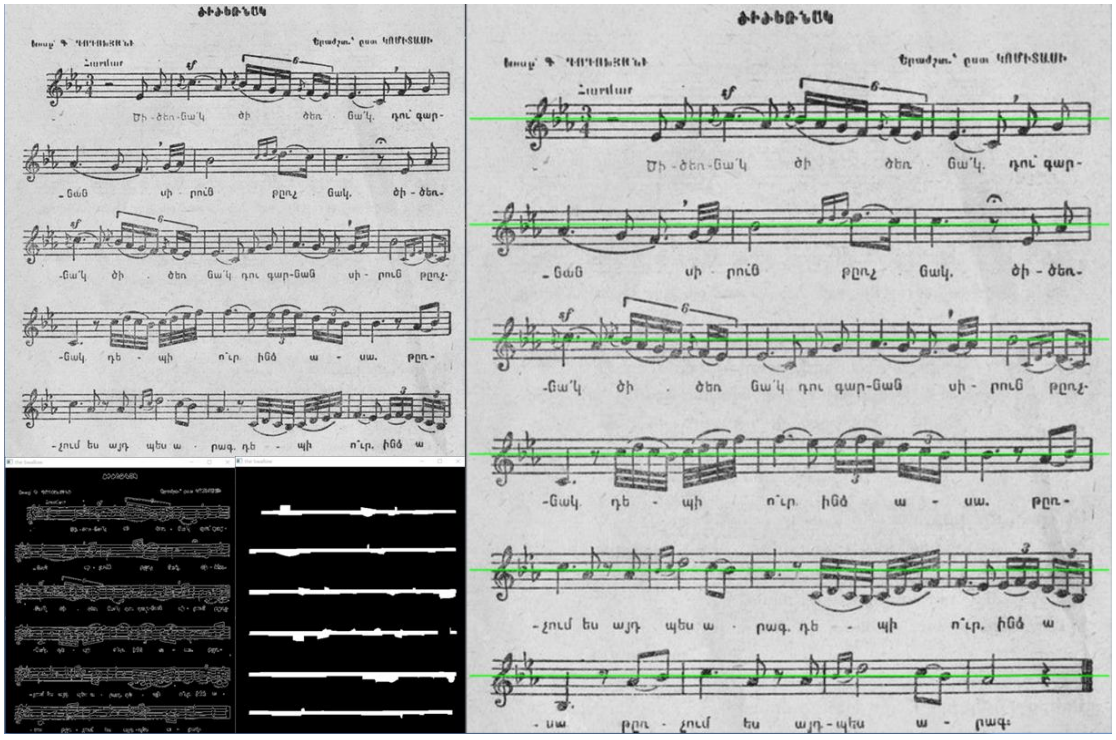
辨识五线谱的能力，没有经过专业训练很难获得。许多优秀的音乐作品除了五线谱之外，根本没有其他载体，十分不利于普及和推广；这一项目的初衷之一就是将音乐艺术的推广变得更加容易。另一方面，卷积网络还从未被应用到这一领域，这个项目也是借机对此进行探索。

第二章 识别五线谱

2.1 标定行的位置

想要识别乐谱上的音符，先要找到它们的位置。如果能够先找到乐谱的每一行，就可以缩小搜索的范围。

如图所示，这一步采取的是相对传统的图像处理方法。

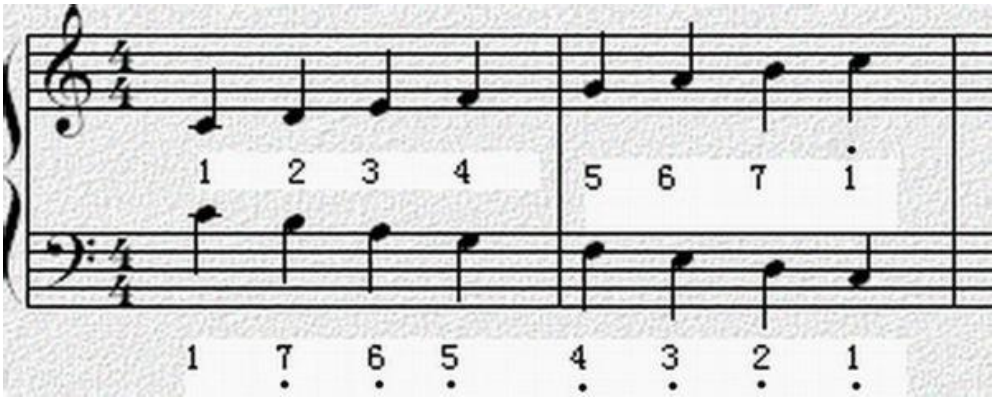


左上：原图 左下：二值化，膨胀-腐蚀操作 右：经过投影-阈值操作得到行的位置

标定行的位置之后，我们就可以把每一行的图像分割出来。这样，后续的工作都将围绕单独一行展开，降低了处理的复杂度。

2.2 识别音高

如图所示，在五线谱中，一个音的音高是由相应音符的头部和五条谱线的相对位置决定的。所以，定位了音符的头部就能够获得相应的音高。

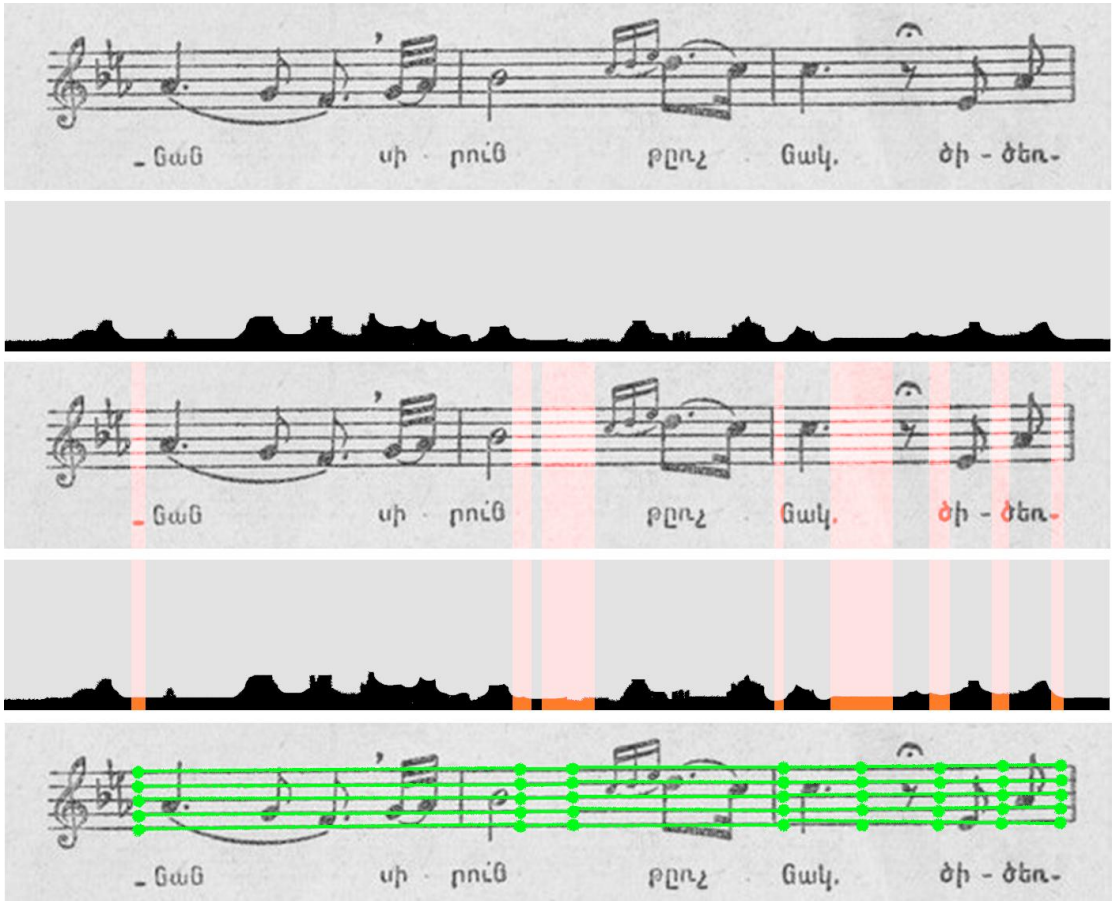


音符代表的音高由其位置决定

我们没有直接寻找音符的位置，而是从精确标定五条谱线开始。因为在五线谱中，音符相对五条谱线的位置是有限的；标定五条线的位置之后，沿着这五条线就能找到几乎所有的音符。

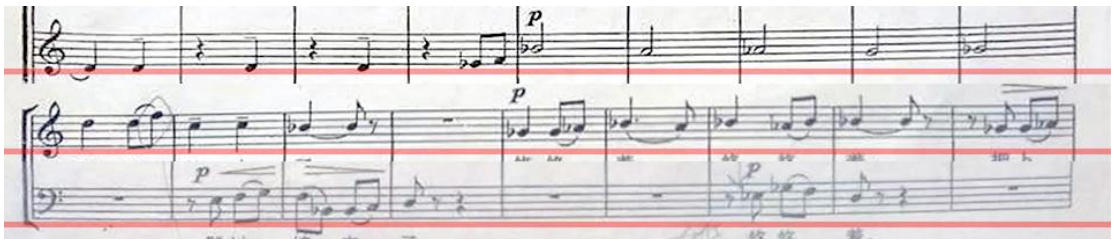
2.3 标定谱线

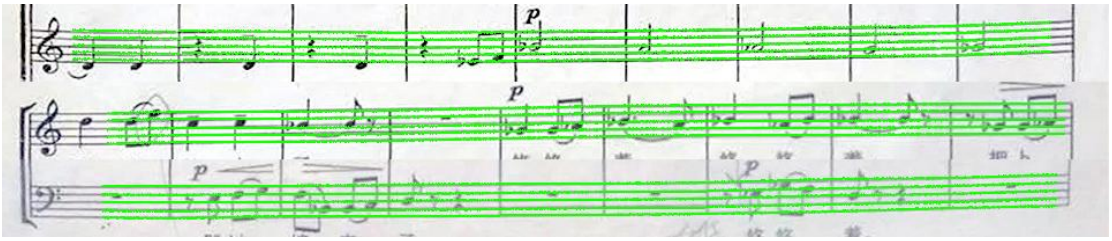
标定谱线的流程如图所示：



首先向竖直方向投影，把那些数值明显小于平均水平的区间挑选出来——这些图像区段往往只包含五条谱线。分析这些区段，可以生成一系列对应的锚点，每个锚点都包含这一竖直邻域内五条谱线纵坐标的最佳估计。把这些锚点连接起来，得到的线条就可以认为是谱线的精确近似。

这一方法原理简单，实验效果良好。值得注意的是，一些印刷质量较差的乐谱，其谱线可能有一定程度的倾斜、弯曲，用这种方法仍能比较准确地对谱线进行定位。



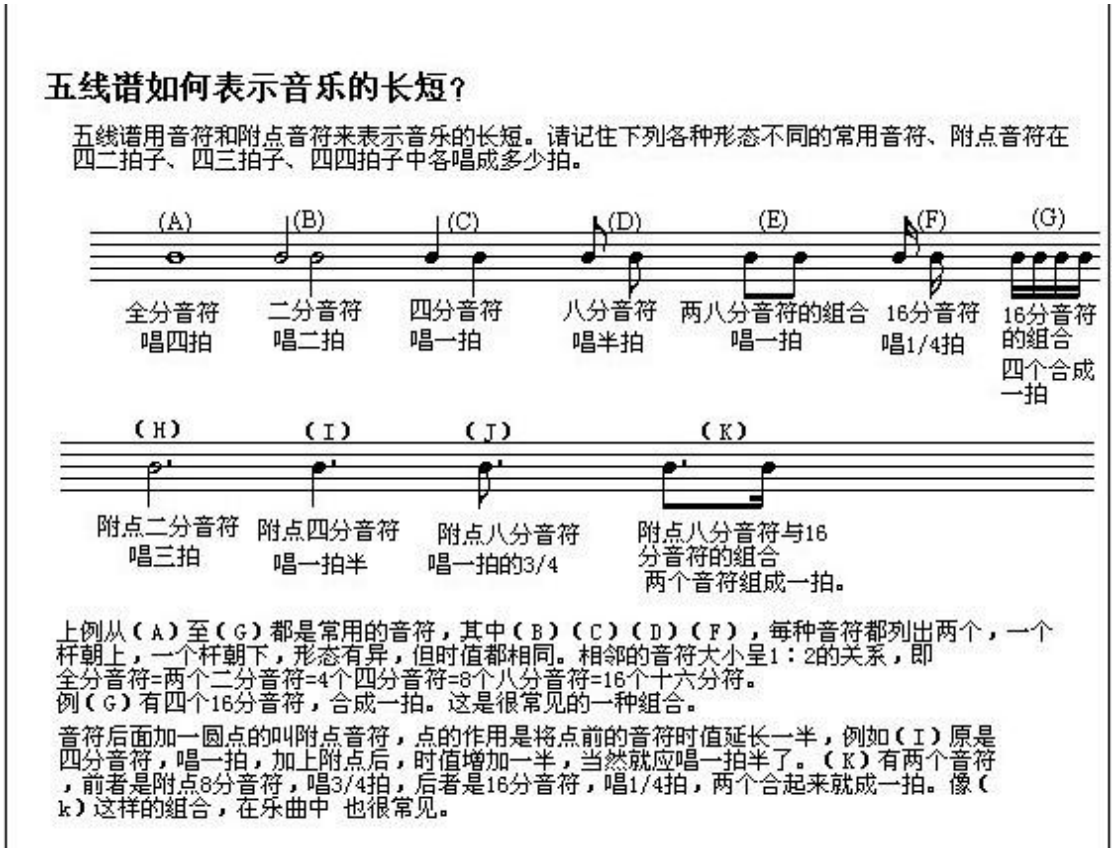


对弯曲谱线的定位示例

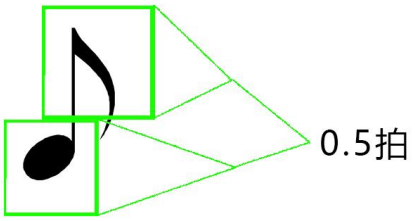
如果能够正确标定五条谱线的位置，之后就可以沿着这五条线把乐谱中可能“有内容”的小区块提取出来。如果一个小区块中是一个音符，由于我们已经获得了谱线的精确定位，那么只需要计算该音符和谱线的相对位置就能够得到它的音高。

2.4 识别时值

“拍”是描述一个音的时长的基本单位。一个音的拍数，也就是它的时值，是我们需要从五线谱中获取的另一项重要信息。在五线谱中，一个音的时值由相应音符的形状决定。如图所示：



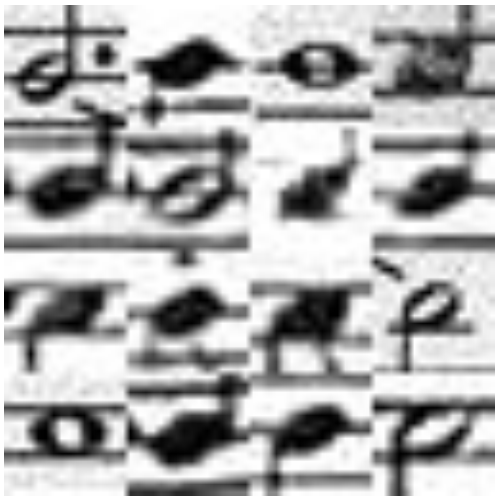
所以我们需要设计一系列分类器，它们的输入是音符的某一部分，而它们要根据这些“部件”最终综合出有关这个音的完整时值信息。



由音符的形状判断其时值

2.5 难点

在乐谱印刷质量、设备参数、拍摄环境等因素共同作用下，输入到分类器的图像可能会有诸如亮度波动、音符印刷字体差异、几何畸变、清晰度差异、印刷质量差异等等问题。如图所示：

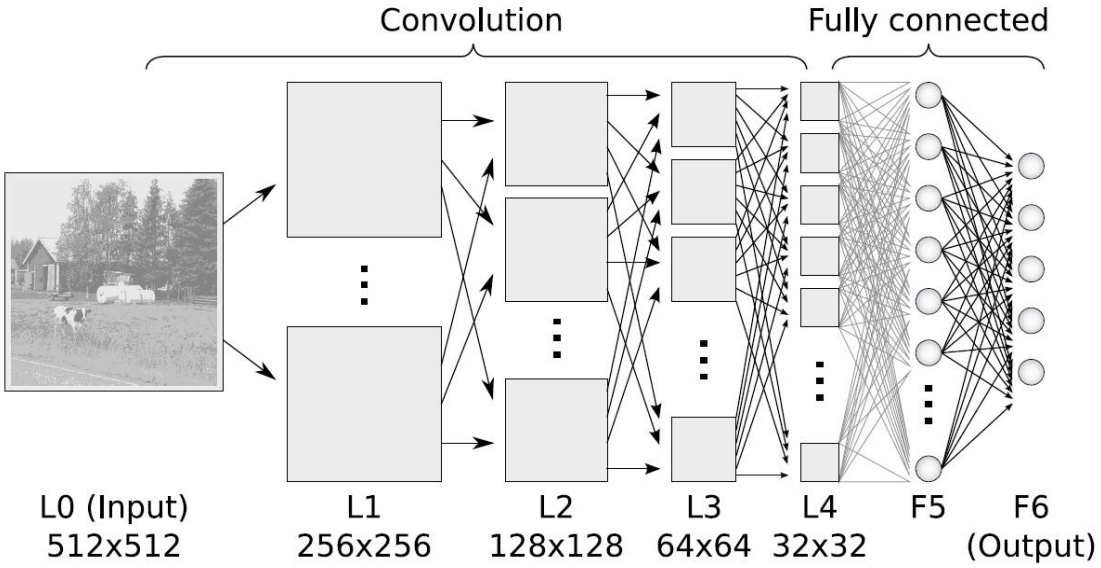


来自真实照片的样本

另一个问题是：之前用来参照音高的谱线，在当前场景下却变成了一种强烈的干扰。它们具备一定结构，不是简单的随机噪声，强度很大，无法忽略。

所以在训练分类器的过程中，必须使其具备排除上述这些干扰的能力。

2.6 卷积神经网络概述



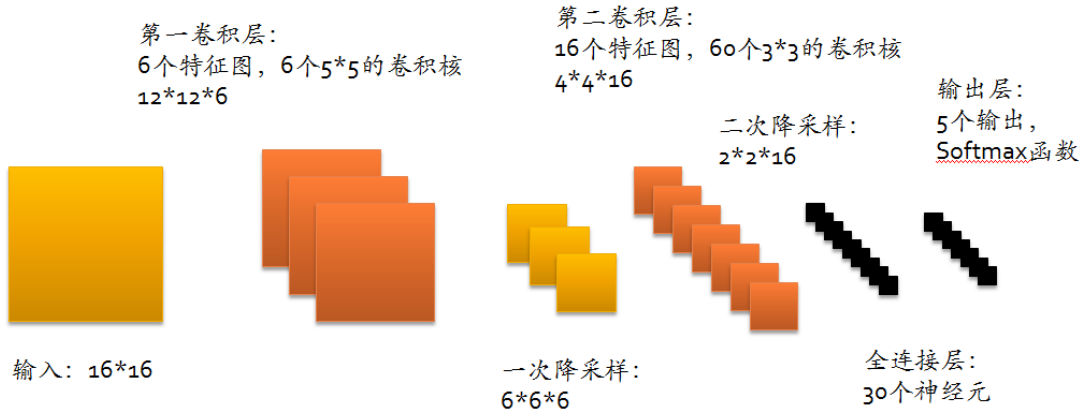
卷积神经网络是从传统前馈神经网络发展而来的一种深度学习模型。局部连接、权值共享的特性使得卷积神经网络具备了提取图像局部特征的能力，而降采样层的存在更使它具备了相当程度的空间抗畸变性；卷积-降采样交替的深层结构成就了卷积网络出众的性能[1]。在图像识别领域，卷积网络早已超越了许多传统机器学习算法[2]，并仍然呈现出广阔的发展空间。卷积网络的这些优良特性能够解决我们在识别音符形态上遇到的问题。

2.7 noteNet-1

这个项目中用到了不止一个神经网络，其中有判定音符头部形状的，有判定尾部形状的，还有判定是头部还是尾部的…… 但它们的原理是相同的，只是训练集不同，或者网络的配置有一些区别。简明起见，后面就以专门识别音符头部的 noteNet-1 网络为例进行说明。

2.7.1 网络结构

noteNet-1 结构如图所示：



以识别音符的头部为例，我们简要介绍 noteNet-1 的工作流程。

我们把音符头部分为 5 类，如图所示：

16*16 的小块二值图像作为网络输入，在第一卷积层与 6 个 5*5 的卷积核作卷积操作，得到 6 个 12*12 的特征图；然后经过一次平均降采样，把图像空间分辨率缩减至原先一半；之后，把 6 个降采样的特征图在第二卷积层与 60 个 3*3 的卷积核按下图所示规律作卷积操作，得到 16 个 4*4 的特征图；经过二次降采样，特征图分辨率再次减半；然后，降采样的特征图与 30 个单元全连接，这部分就和普通的前馈神经网络是相同的；最后，输出层由 5 个单元构成，用 softmax 函数进行规约。

考虑到移动设备的性能问题，这一网络规模设计得比较小；而且由于充分利用了五线谱的先验知识，也没有将其做得很庞大的必要。

我们把之前检测到的、有可能是音符的区域输入到网络，第一卷积层负责提取少量初级特征；降采样层降低了空间分辨率，对网络的抗干扰性有重要作用；第二卷积层提取的特征更高级，更局部，数目更多；全连接层增强了网络的拟合性能；而 softmax 输出层把五个输出值规约为五项的离散概率分布，即全都是 0-1 之间的实数，加起来总和为 1；最后进行类别判定时，只需按输出从大到小的顺序进行判断即可。实际上，在多分类问题上，softmax 输出层是一种比较普遍采用的设计方案。[3]

2.7.2 另一个难点

卷积网络属于有监督学习算法，这类算法需要大量的带标签数据做为训练样本——传统的解决方案是，手工截取，手工标注，录入数据集——这种方法过于费力，过于低效，我们没有采用。

2.7.3 Note_head_samples_50k 自动生成的标签数据集

我们生成训练样本的过程可以由下图概括：

Note_head_samples 训练集，50000个样本

我们用这种方法生成了一个具有 50000 个样本的训练集，命名为 Note_head_samples_50k。它从具备基本特征的模板出发，添加模拟的随机谱线干扰，以及仿真的拉伸、位移、扭曲等形变，与前文中来自真实照片的样本对比，具备比较高的仿真度，可以用于 noteNet-1 的训练。

2.7.4 noteNet-1 训练策略

训练集和测试集

采用 Note_head_samples_50k 数据集的前 45000 个样本作为训练集，后 5000 个样本作为测试集。

损失函数

由于使用了 softmax 输出函数，使用对应的交叉熵损失函数评价网络输出误差。

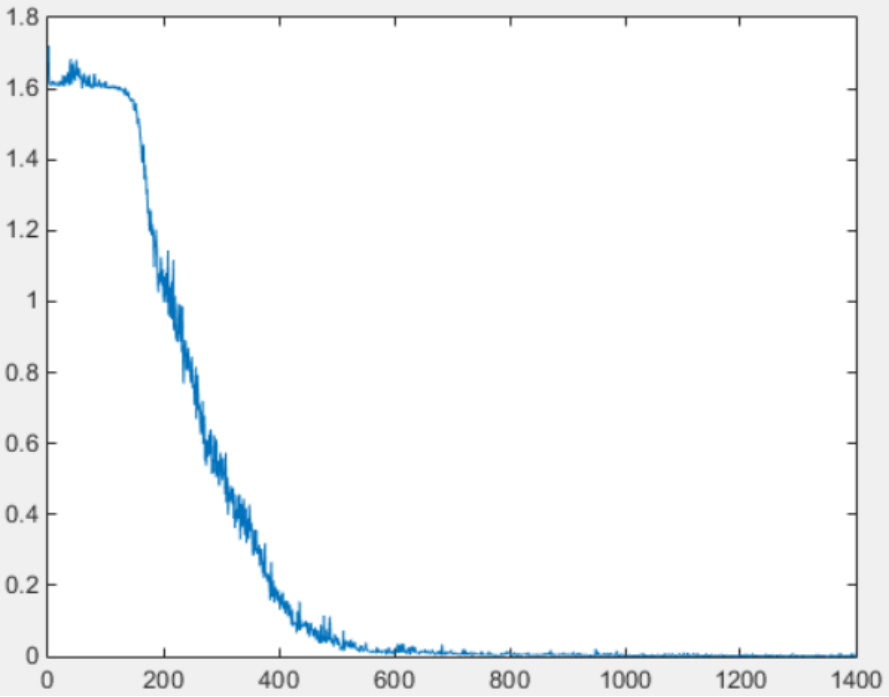
梯度下降策略

采用批量随机梯度下降法来调整网络参数。每 256 个样本为一批，每批进行一次参数调整，用每一批的总误差来计算网络参数的梯度。

学习速率调整策略

初始值为 0.2，每次遍历完整个训练集后，学习速率除以 1.05。

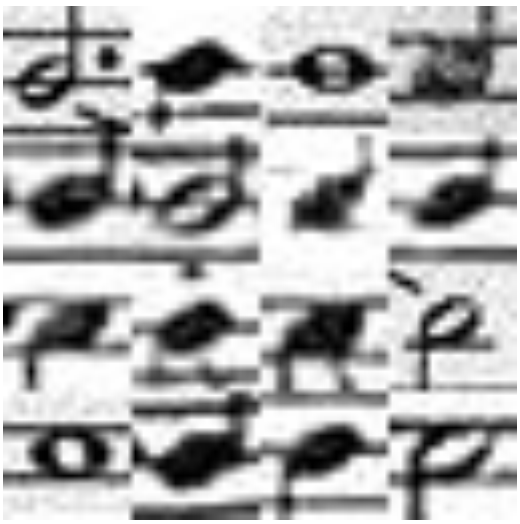
训练过程记录



横轴：迭代次数 纵轴：误差函数值

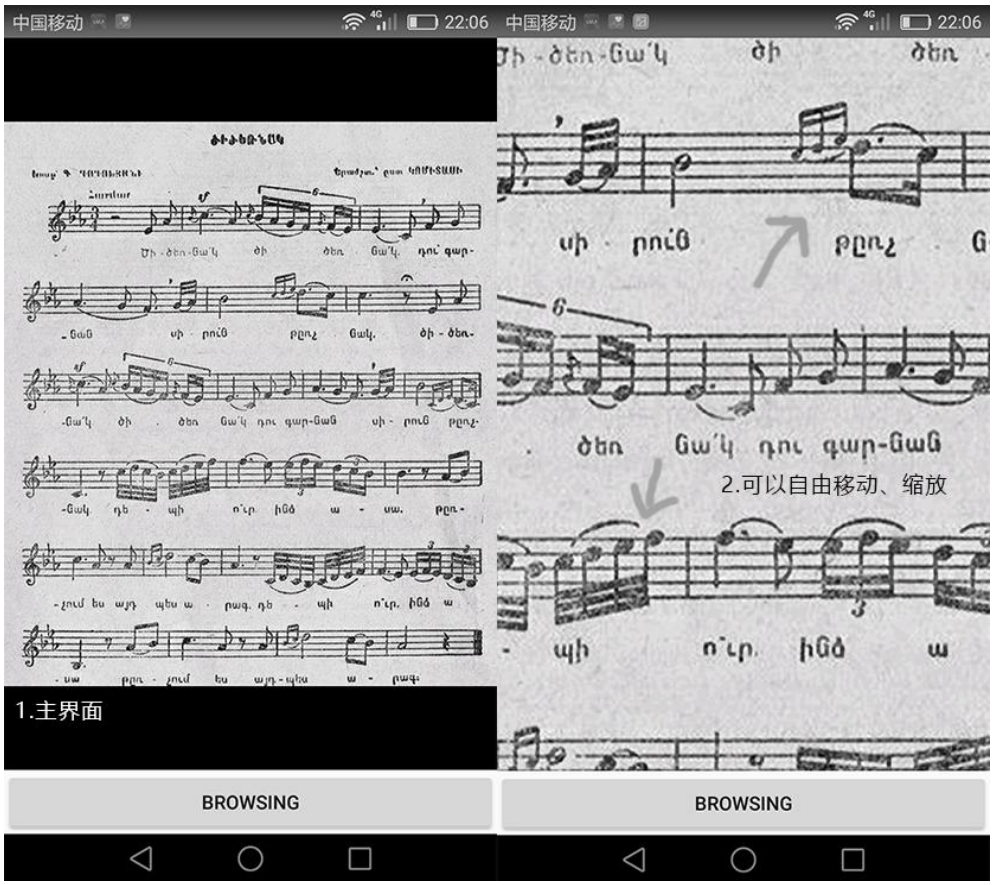
在 1000 次迭代之后，误差函数值逐步趋近于 0；8 轮完整训练以后，测试集正确率达到 99.98%。

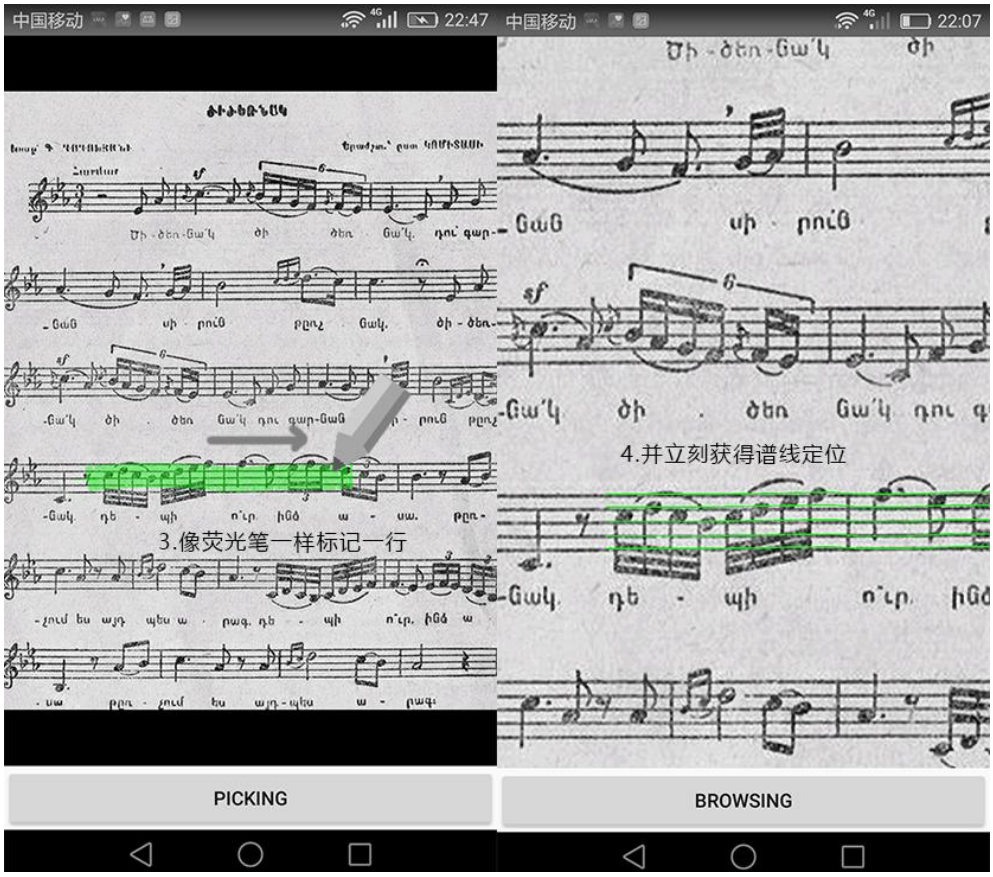
2.7.5 真实照片测试

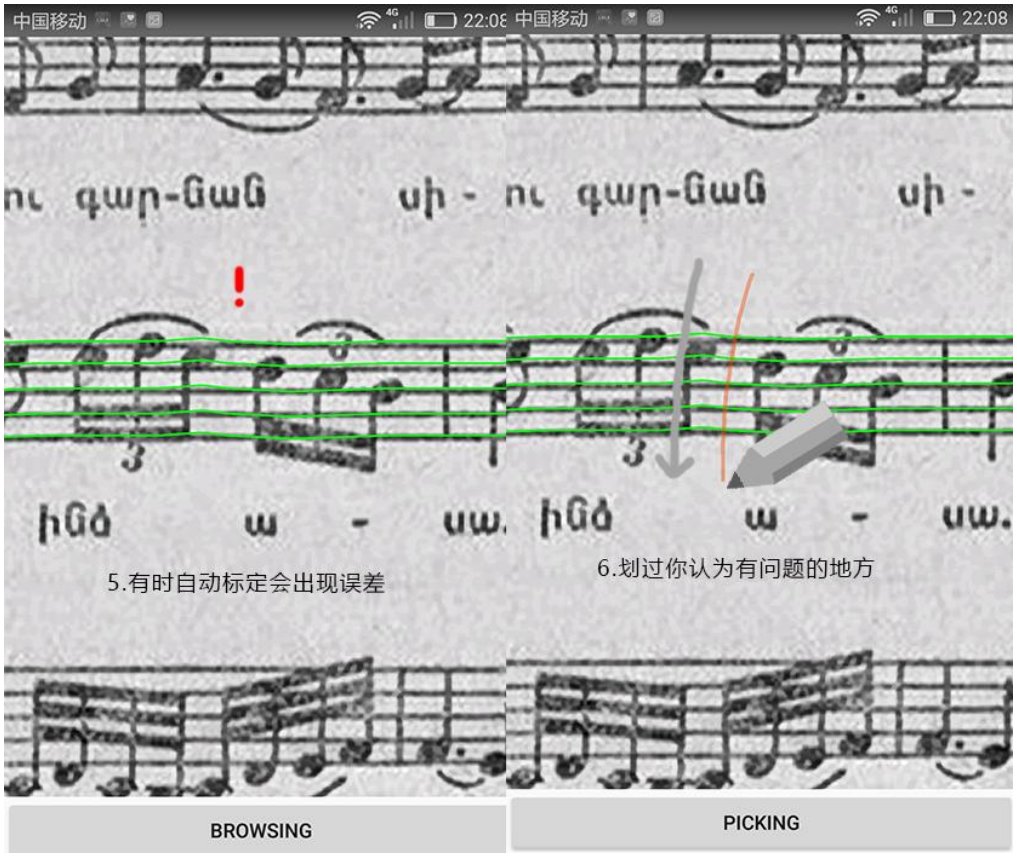


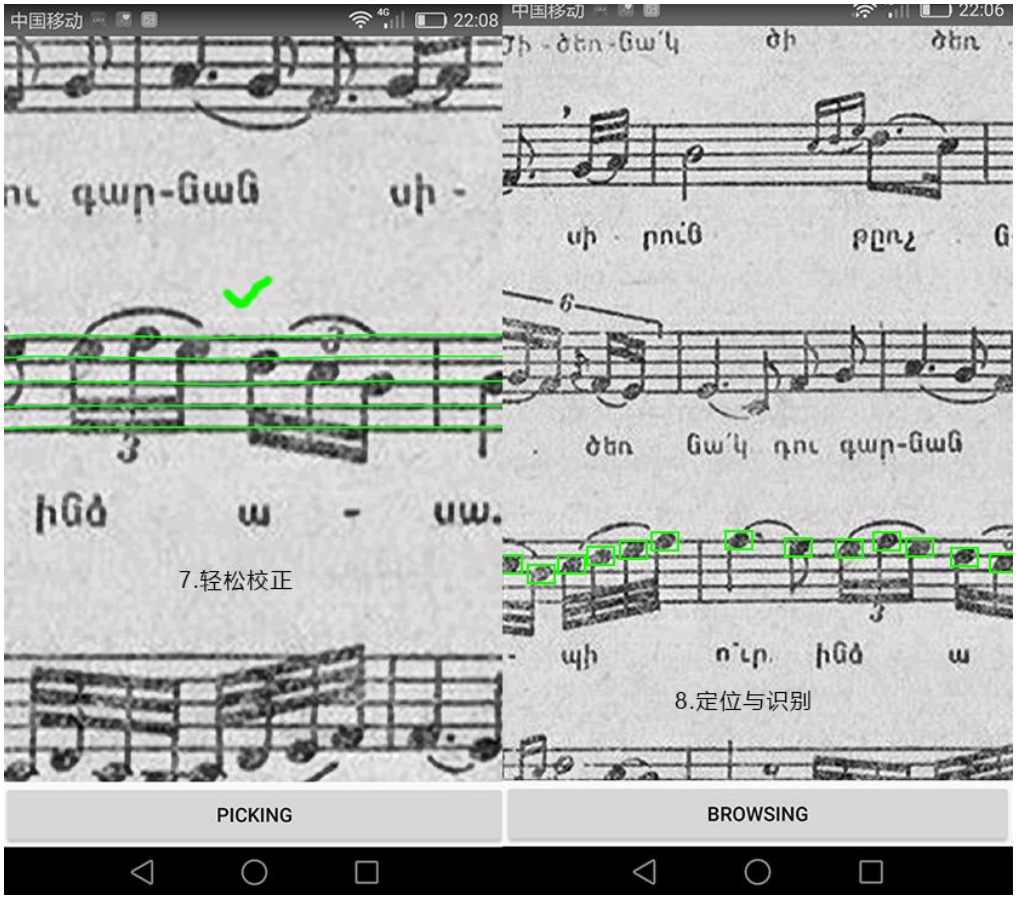
上图中的样例均能正确识别，注意到它们在清晰度、背景纹理、形状位置上的差异，说明 noteNet-1 已经具备一定的抗变能力。

第三章 应用界面展示




中国移动
22:47
3.像荧光笔一样标记一行
PICKING
中国移动
22:07
4.并立刻获得谱线定位
BROWSING


中国移动
22:08
5.有时自动标定会出现误差
BROWSING
中国移动
22:08
6.划过你认为有问题的地方
PICKING



第四章 总结与展望

这个项目远没有结束，我们计划把它长期进行下去。目前开发出的应用程序只是一个雏形，无论是算法还是交互，都还有很大改进空间。识别乐谱是一个有趣的问题，目的简单，而实现起来有许多实际的问题需要仔细考虑。目前，我们的应用程序暂且定位在“音乐学习辅助”这一栏。不过我们会继续研究——前段时间，Alpha Go 战胜了人类围棋高手；也许有一天，我们也可以找一个学音乐的人，来和我们的 noteNet 一较高下。

第五章 参考文献

[1] Lecun Y, Bottou L, Bengio Y. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[2] Yann LeCun, http://yann.lecun.com/exdb/mnist/.

[3] Ng A, Ngiam J, Foo CY, Mai Y, Suen C (2011). UFLDL tutorial. Available: http://deeplearning.stanford.edu/wiki/index.php/UFLDL Tutorial. Accessed September 5, 2012.