

Research Report: Chinese Character Recognition Database and Model

Due on August 22, 2017

07.2017

Yuwei Qiu

Contents

Chinese Character Database	3
Printed Chinese Handwritten Character Database	3
Handwritten Chinese Handwritten Character Database	3
Model	4
New Database	6

Results	6
Notes	7
Deep_ocr	7
Summary	8

Chinese Character Database 中文字符数据库

目前，找不到合适的印刷体中文字符数据库，故而暂定印刷体中文字符（Chinese printed character）数据库的建立和手写体中文字符（Chinese handwritten character）数据库的训练以及特征提取同时进行。

预期是：通过设计模型，训练网络，利用已有的庞大的手写体中文字符数据库得到手写体中文字符的features，以及相应训练好的kernels；后期在训练印刷体中文字符识别网络时，可以从由印刷体训练得到的kernels开始训练。

Printed Chinese Handwritten Character Database 印刷体中文文字数据库建库说明（摘要）

要求2500个字符，每个字符生成94种字体。

每个字符每种字体生成三幅灰度图：1）无噪声原图，2）加噪声，3）字符边缘虚化。

每幅图长宽比为128 x 128，字符居中。

图片顺序命名即可，按照2500个字符标注

Handwritten Chinese Handwritten Character Database 手写体中文文字数据库

手写体中文字符数据库选用Cheng-Lin Liu在2013年建立的*CASIA Online and Offline Chinese Handwriting Databases*。该数据库由中国科学院自动化所建立，选取的书写体样例取材于1020位作家的在2007到2010年之间的钢笔字手稿，这些样例中不仅包含独立单个字符，同时包含文段；标记和分割在2010年完成。

整个数据库包含六个小的子数据集，三个可以用于离线（offline），另外三个可用于在线（online）中文字符识别。在离线字符数据库中，独立单个字符数据库共包含来自390万样例中的7185个中文字符和171个标点符号；文段数据库包含5090整页样例和135万字符样本。

所有数据的样本均已完成符号级别的标注，同时给出各类别的NCGF（normalization-cooperated gradient feature）[1]特征模板。NCGF表征了字符在梯度变化上的特征，在日文字符和中文字符上作用显著，是目前用于日文、中文字符识别最广泛应用的一种特征。特征可视化如图

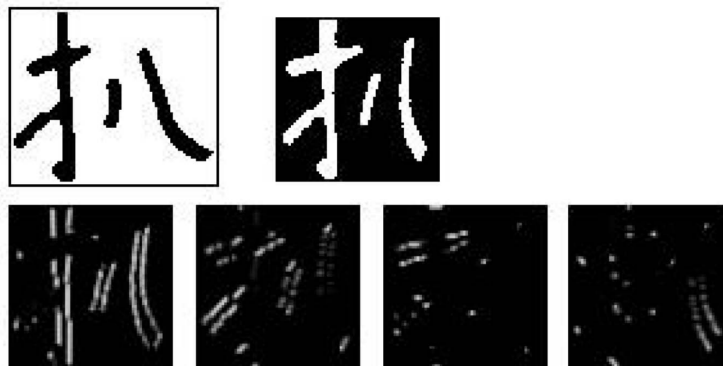


Figure 1: normalization-cooperated gradient 可视化。

相关的图片样例见图。

躲朵踪舵捌情墜蟻峨鵠
 角額沁蟻惡厄扼逼鄢餓
 恩而兀目京徑涓二貳發
 罰符伐乏閱法珥蕩帆番
 翻樊碩鉤繁凡煩反返范
 販犯飯泛坊芳方肪房防
 妨份訪紡放菲非啡飛肥
 匪啡味肺庖沸菹芬酉吩

独立单个字符

令一家落马、判刑、入狱，甚至犯死罪被执行死刑了，媒体关注的焦点往往不是法律问题，而更多的是企业家经营和管理上的问题。在媒体上发表各种意见的，不乏经济学家、管理专家，却很少有法律专家来参与讨论。这是一种不正常的现象。企业家不管在经营、管理上存在什么问题，最终的结局如果是走进监狱，最终的结论如果是经由法院判决有罪，那么，最重要的应该是法律问题！

文段

Figure 2: 离线中文字符样例：左独立单个字符样例，右文段样例。

植在我脑袋最深处可能是回味着刚
 心，忘记了要将车速放慢我的左

字符样例

植在我脑袋最深处可能是回味着刚
 心忘记了要将车速放慢我的左

标记内容

Figure 3: 标记样例：左字符样例，右识别内容。

收集数据时，研究人员提供三套完整的书写模板：每套中包含有10组文字样例，每组文字样例中含13-15页独立单个中文字符和5页文段。对于每一套书写模板，独立单个中文字符样例又分为三种：1) 标点符号，2) 常用中文字符，3) 非常用中文字符。

相关较为详细的介绍可见ICDAR2011相关文献 [2].

Model 中文字符识别模型

目标是中文字符识别，从给出的bounding box中识别出字符含义。单个字符的识别相当于图片分类，即根据字符特征判断字符所归属的类别。文段识别建立在单个字符识别的基础上，通过在提取特征的时候加入字符所出现的位置信息，找到特征与类别之间的联系，最终得到识别结果。

模型抽象见图。

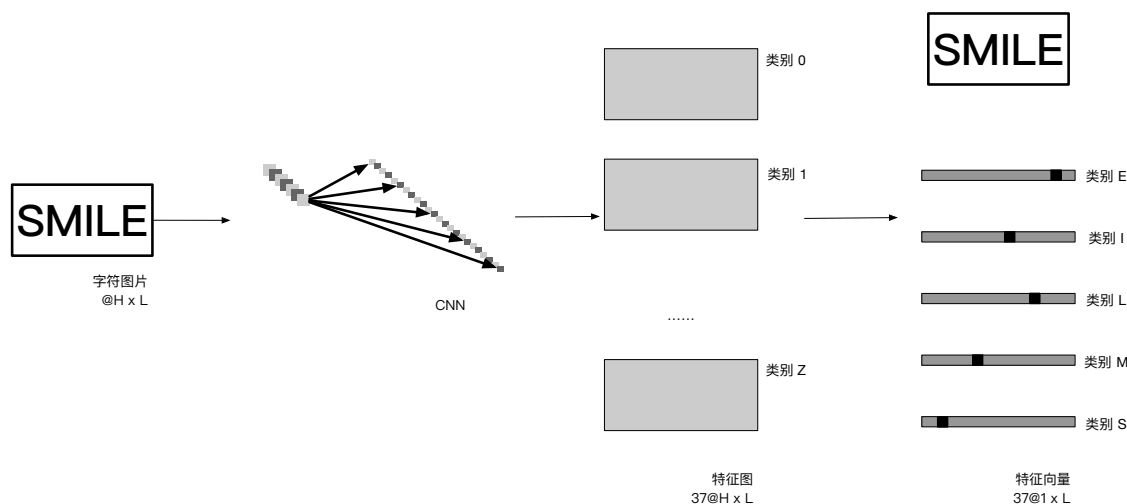


Figure 4: 模型的抽象。

建立字符识别模型时，参考文献 [3] 中字符识别的部分：该模型用于英文字符和数字字符识别，共有 37 个输出类别（Aa—Zz，0—9）。该模型采用卷积网络（CNN），输入为长 L 高 H 的 bounding box，其中的字符数目未知，字符大小也未知，具有单词级标注（即不存在带个字符的标注）。

CNN 输出为 37 个与 bounding box 等长等高的 heat map，每一个 heat map 都对应于 37 类别中的某一类，通过相关数学变换（变换具体过程略，详见参考文献 [3]），将 heat map 变换为一个与 bounding box 等长的一维行向量（共 L 列）。这个矢量在某些列达到局部最大值，则该类对应的位置即为该矢量对应字符出现的位置。通过这 37 个矢量，可以得到一个 bounding box 内的单词级别的识别结果。

相关样例见图。

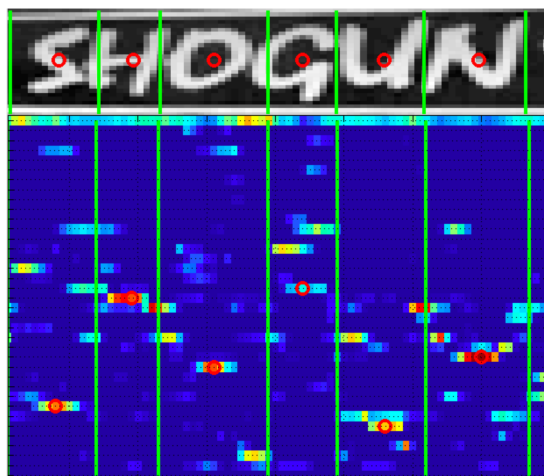


Figure 5: 单词SHOGUN的识别结果。

在相关文献的demo中，分别给出了需要实现该网络所需要的三层网络/分类器的结构和相关参数：

cat	0	1	2	3	4	5	6	7
old	1	2	3	4	5	6	7	8
new	1	2	3	4	5	6	7	8
cat	8	9	A	a	B	b	C	c
old	9	10	11	37	12	38	13	39
new	9	10	11	12	13	14	15	16
cat	D	d	E	e	F	f	G	g
old	14	40	15	41	16	42	17	43
new	17	18	19	20	21	22	23	24
cat	H	h	I	i	J	j	K	k
old	18	44	19	45	20	46	21	47
new	25	26	27	28	29	30	31	32
cat	L	l	M	m	N	n	O	o
old	22	48	23	49	24	50	25	51
new	33	34	35	36	37	38	39	40
cat	P	p	Q	q	R	r	S	s
old	26	52	27	53	28	54	29	55
new	41	42	43	44	45	46	47	48
cat	T	t	U	u	V	v	W	w
old	30	56	31	57	32	58	33	59
new	49	50	51	52	53	54	55	56
cat	X	x	Y	y	Z	z		
old	34	60	35	61	36	62		
new	57	58	59	60	61	62		

1) 文字/非文字分类层参数； 2) 字符行分类器参数； 3) 单个字符分类器参数；

以及各类model的相关数据集。

各类model含有gpu版本和cpu版本。

三类model在ICDAR2013上的测试结果分别为98.2%，91.0%，86.8%。

New Database 数据集准备

我们直接在已有的model上进行初始测试，主要采用了case-sensitive的model，也就是区分大小写共（10+26+26=62）类字符识别的模型。

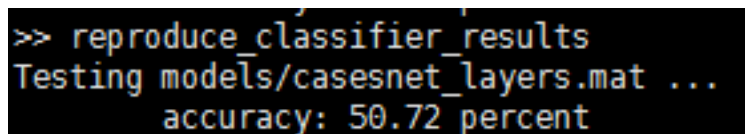
这里我们采用的数据集是人工合成的各类印刷体数字、大写小写字母的数据集，共31496张图片。同时，我们将新的数据集的label进行映射，将其映射到现有model的label方式上面。具体如下表

由于memory不足，所以将原有的128像素缩放为24像素，以降低运行所需要的时间和空间。

Results 结果

得到结果如图，在尚未进行调参之前，准确率大约50%，效果一般。

提取出第2类，也就是数字1的部分成功和失败案例，发现该model对花体印刷体字符识别效果并不显



```
>> reproduce_classifier_results
Testing models/casesnet_layers.mat ...
accuracy: 50.72 percent
```

Figure 6: 结果截图。

著，而且由于印刷体中字符I，L，1形状非常接近，所以导致部分识别结果出错。

Notes

针对以下问题进行补充说明：

以下提到的model均为Jaderberg在eccv2014中提出的基于自然场景字符图片的文字识别。而我在第一阶段想要解决的是非自然场景字符的识别，所以我决定先测试一下该model在非自然场景字符数据集中的准确率，再进行下一步工作。

表格共有三行——cat（字符类别），old（数据集原始label类号），new（model训练和测试所采用的label类号）。例如对于字符“0”，数据集中它的label号为0，在model训练时采用的label也为0；对于字符“Z（大写）”，数据集中它的label号为36，而在model训练采用的label为61。需要注意的是这两种label方式本质上并没有任何区别，只是为了方便测试所以我采用了model训练和测试采用的label方式。

将非自然场景字符图片输入到eccv相关文献中textspotting的model里进行测试得到的准确率。需要注意的是，该文献所提供的相关kernel参数均为基于自然场景文本字符训练得到的，故而直接在非自然场景字符数据集中测试得到的效果并不显著；但是，这个model却可以作为一个比较好的起点，我可以继续在这个model的基础上进行调参。

我发现，特别的对于数字1而言，花体字符在识别过程中错误率显著（高达65%），同时还有相当大比例的数字1被识别为字符L（大写字母）或字符l（小写字母）。以上只是特别针对数字1而言的结果分析，接下来的几天，我将逐类别分析每类字符的错误情况。

Deep_ocr

目前采用Deep_{orc}的已有模型，该模型采用投影切分法切分字符，采用深度学习方法进行识别，本周主要集中在中文词语、句子的切分。在华为所给的图片中只测试中文字符部分的准确度。

投影切分法简要概述 该模型采用的投影切分法的主要过程如下：1 输入一张图片，首先将一幅rgb图像转化为灰度图像，进行黑白反转之后二值化。2 采用自适应的边缘提取算子将文字部分取出。3 将文字部分进行转化，使得字符笔画为单像素宽度。4 在水平方向和竖直方向进行像素投影，即统计每一行和每一列内白色像素点的求和计算。5) 对于统计所得到的水平方向和竖直方向的总和数列，设置minimum_val和minimum_range：从第一行或第一列开始，当某一行或某一列的白色像素点总数小于minimum_val，设定该行为start，紧接着的白色像素点总数小于minimum_val的一行或一列，设定该行为end：当start和end之间距离大于minimum_range时，设定从start到end是一个文字。

主要误差分析 主要出现的几种错误情况包含1) 零碎像素级的非文字噪声处发生误判, 2) 两个文字被判定为一个文字, 3) 背景噪声过大导致边缘提取有极大的干扰, 在进行像素计数时产生较大问题。

误差规避与性能提升

对于错误情况1 和2, 可通过`minimum_val`和`minimum_range`的调整来规避。

错误1: 适当提高`minimum_range`, 降低该算法对于细小像素级噪声的敏感程度, 达到提升准确率的目的。具体解释即为, 由于细小噪声相对于文字块而言非常小, 所以提高`minimum_range`即将相对很小的噪声块判定为非文字。错误2: 由于中文字符大部分为方块字, 所以可近似认为文字块在水平和竖直方向的大小应该是几乎相同的, 所以切割的时候若出现长宽比或宽长比大于或小于1.8的情况, 即可大致判定出现了文字切分错误, 此时将该异常字符块平分即可。

Summary

一个学期的时间说长也短, 感觉自己没有做很多工作, 只是很喜欢把每个阶段遇到的问题, 得到的结果, 还有受到的启发记录下来。从最开始面对着一个黑屏的手足无措到后来能够比较熟练地操控linux系统和caffe, 我学到了很多。而且也很开心的能够将在这段实习中学到的相关算法和工具运用到课堂上, 并且取得了不错的成绩。

感谢王老师和宏马师兄认真负责的指导。

References

- [1] C.-L. Liu, "Normalization-cooperated gradient feature extraction for handwritten character recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 8, pp. 1465–1469, 2007.
- [2] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 37–41.
- [3] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *European conference on computer vision*. Springer, 2014, pp. 512–528.