**Tesseract-OCR使用**

### 1 准备

* 下载安装[tesseract](https://github.com/tesseract-ocr/tesseract/wiki)  （默认安装位置C:\Program Files (x86)\Tesseract-OCR）
* 下载安装训练工具[jTessBoxEditor](https://sourceforge.net/projects/vietocr/files/jTessBoxEditor/jTessBoxEditor-1.7.1.zip/download)。

简单测试：

tesseract test.jpg result.txt -l eng

* test.jpg 待识别图片地址
* result.txt 结果保存地址
* -l 指定要用的语言库

### 2 训练【制作自己的语言库】

**自定义一个语言库名字: testlang.testfont.exp0**

**命名方式：** [lang].[fontname].exp[num]

1. **制作样本图像。【通过截图，得到一行一行的文字】**



样本图像实例（图中有8个样本）

1. **生成 \*.tif格式的训练图**

点击jTessBoxEditor菜单栏中Tools--->Merge TIFF

1. 第一个对话框，选择所有的样本图像

，

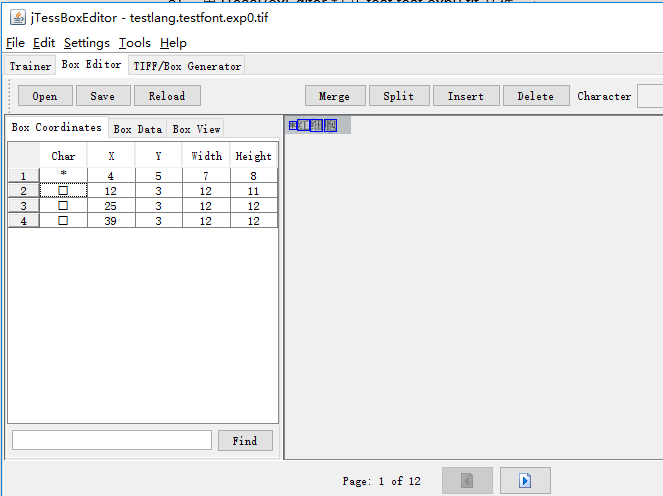
1. 第二个对话框，填入需生成的文件**testlang.testfont.exp0**.tif。



1. 合并得到**testlang.testfont.exp0**.tif
2. **生成box文件**

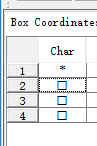
tesseract **testlang.testfont.exp0**.tif **testlang.testfont.exp0** -l eng -psm 7 batch.nochop makebox

1. **修改box文件。**
   1. 用jTessBoxEditor打开test.test.exp0.tif文件.



Box Editor -> Open

* 1. 手动逐个校正识别结果。



这几个框里面的内容，如果有错就校正（中文显示的都是口，不用管）

1. 新建文件批处理文件run.bat，内容从下面的附录中复制，修改前三行，执行run.bat，生成语言文件testlang.traineddata。
2. 由此，得到了我们自己的语言库testlang.traineddata。

### 三 识别

1. 将语言文件testlang.traineddata拷贝到\*\*\Tesseract-OCR\tessdata目录下。
2. 执行下面这行代码， -l 参数指定语言库，这里使用我们上面训练的结果。

tesseract test.jpg result.txt -l testlang

### 附录

run.bat 文件内容

|  |
| --- |
| rem [lang].[fontname].exp[num]  set lang=testlang  set fontname=testfont  set num=0  set name=%lang%.%fontname%.exp%num%  echo %fontname% 0 0 0 0 0 >font\_properties  tesseract %name%.tif %name% -l eng -psm 7 nobatch box.train  unicharset\_extractor %name%.box  shapeclustering -F font\_properties -U unicharset -O %lang%.unicharset %name%.tr  mftraining -F font\_properties -U unicharset -O %lang%.unicharset %name%.tr  cntraining %name%.tr    echo Rename Files..  rename normproto %lang%.normproto  rename inttemp %lang%.inttemp  rename pffmtable %lang%.pffmtable  rename unicharset %lang%.unicharset  rename shapetable %lang%.shapetable    echo Create Tessdata..  combine\_tessdata %lang%. |