里程数识别使用文档

一、安装tesseract-ocr

1、下载安装包并解压

2、直接执行下载好的tesseract-ocr-setup-4.0.0-alpha.20180109.exe，下一步、下一步安装。安装过程中，会让你安装额外的语言包，可根据选择下载。

3、配置环境变量。增加一个TESSDATA\_PREFIX变量名，（例如：变量值还是我的安装路径F:\Tesseract-OCR\tessdata）这是将语言字库文件夹添加到变量中;

4、重启，打开命令行，输入tesseract -v，可以看到版本信息则说明安装成功。

5、为了在python代码中使用tesseract功能，使用pip安装pytesseract：pip install pytesseract

二、样本训练

1、生成box文件

tesseract num1.sign.exp2.tif num1.sign.exp2 batch.nochop makebox

2、样本纠错

打开jTessBoxEditor，在Box Editor中进行样本纠错。完成之后保存。

3、生成font\_properties

echo sign 0 0 0 0 0 >font\_properties

4、生成字符集文件

unicharset\_extractor num1.sign.exp2.box

5、生成shape文件

shapeclustering -F font\_properties -U unicharset num1.sign.exp2.tr

6、生成聚集字符特征文件

mftraining -F font\_properties -U unicharset -O unicharset num1.unicharset num1.sign.exp2.tr

7、生成字符正常化特征文件

cntraining num1.sign.exp2.tr

8、生成训练结果

combine\_tessdata num1.

9、白名单：说明要识别的结果范围，0-9

三、使用

1、打开程序，选择需要识别的图片

2、可以看到里程数显示在右下角的文本框中