**血化单自动识别--技术路线**

2017.3.5

总目标：**将拍摄得到的血化单，识别出项目和结果两列数据。**

按照流程顺序，工作可分为3步：

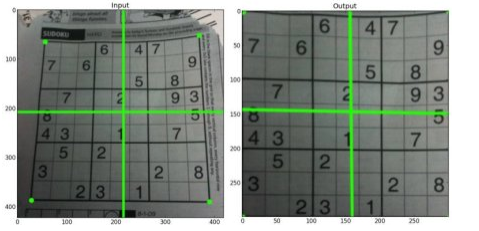
1. 图像预处理：将图像处理成单列的格式。【OpenCV】**【难点】**
2. 图像文字识别：对单列格式的图像进行识别。【Tesseract-OCR】
3. 识别结果的提取：针对2的结果（文本文件），进行解析、纠错。

## 一、图像预处理

1. 目的：将拍摄的血化单，进行处理得到单列数据。 如下图。

|  |  |  |
| --- | --- | --- |
| **输入** | **输出1** | **输出2** |
|  |  |  |

1. 工具： **OpenCV**，开放的语言接口支持大部分语言。为了直接集成，可以根据现有的项目语言来选择。
2. 预计的工作内容**：**
3. **透视变换warpPerspective**  🡪🡪🡪 根据四个角，将图片扶正，如下图。



1. **霍夫变换(Hough Transform)直线检测**  🡪🡪🡪 检测血化单上下两条线。
2. **锐化** 🡪🡪🡪 将图片变得黑白的，更清晰。
3. **canny 边缘检测**  🡪🡪🡪 检查文字轮廓。
4. **切割** 🡪🡪🡪 最终将图片切割成单列。

## 二、图像文字识别

1. 目标：单列文字图像 🡪 文本文字。
2. 工具： 开源的**Tesseract-OCR（**[**https://github.com/tesseract-ocr/tesseract**](https://github.com/tesseract-ocr/tesseract)**）。**使用方法参考附件：*Tesseract-OCR使用.doc.*
3. 预计的工作内容**：**
   1. 模型训练，通过对大量的血化单中的项目和结果两列的样本进行训练，得到自己的语言库。【耗时工作】
   2. 识别，用上面得到的语言库进行识别。

## 三、识别结果的提取

1. 目标： 得到最终的结果。
2. 工具： 建议使用Python等文本处理方便的语言。
3. 预计的工作内容：
   1. 数据纠错：因为化验项目有限，所以有字典可查。可以根据字典进行验证和纠错。