TechOcr算法分析及进度汇报

TechOcr的开发目标是：对特定版式的图像，进行结构化识别。其识别算法使用了目前开源ocr软件中，对中文识别较好的Tesseract。并在识别前，对图像进行 旋转、定位等预处理，以使得被识别的文字能够被结构化处理。

目前已经完成的v0.1版本，实现了对“营业执照”这一版式图像的识别，并对其内容进行结构化识别。

在开发中，遇到的问题是：Tesseract对图像中，哪些是文字的判定尚需提高。特别是对单独的一个左右结构、上下结构的汉字（关键点是这个汉字的笔画不连通），会将其识别为多个汉字。

下面就开发过程、过程中遇到的问题，以及预计的解决思路，向领导做一汇报。

一．Tesseract

Tesseract项目最初由惠普从1985年开始开发，并于2005年开源。自2006年起转由谷歌开发，从2007年5月的v1.04到2016年2月的v3.04.01，目前项目仍在进行中，最新版本为v3.05.dev。

Tesseract的可下载的[执行版本文件](https://sourceforge.net/projects/tesseract-ocr-alt/files/)，只有v3.02。国内网络上的资源也没有更新的。经测试，v3.02版本对中文的识别度约为50%，无法进行实际应用。

通过下载Tesseract的[源代码](https://github.com/tesseract-ocr/tesseract)，并按[说明](https://github.com/tesseract-ocr/tesseract/wiki/Compiling)编译后，得到最新的v3.05.dev。对于从拷屏得到的印刷体文字图像，字体大于48（这么大），可以做到近乎完美的识别。

二．Tesseract的识别模式

Tesseract可以对指定矩形范围内的图像进行识别，识别模式分为：字、词、行、块、页，五种。

如果矩形设定正确，字模式，识别汉字最好；词模式，识别标点最好。

另外Tesseract对倾斜、旋转的字体无法识别，大致能够容忍5度的倾斜。

三．TechOcr的工作任务

实际的识别工作，由Tesseract完成；TechOcr的工作就是找到矩形位置，并提交识别。具体分为以下步骤：

1．找到图像轮廓，或根据每一行字的倾斜程序，判定出图像的倾斜状态；并进行透视投影变换，得到矩形的待识别区域。

简言之，就是将图像转正。消除倾斜、旋转因素。

此步工作，需要较多的图像学、数学知识，在实现过程中，会放在最后完成。

在此之前，通过调整拍摄位置，以得到近似矩形的图像。

2．预识别，找到图像中单独的字，并根据这些字的内容与位置，判定出模板。

如“营业执照”中必定有“注册资本”、“营业期限”；“身份证”中必定有“性别”、“民族”。

通过找到的这些特征字，确定图像的版式。

经过测试，Tesseract在这样的识别中，有相当大的随机性：即同样一份原件，不同的拍摄照片中，可以识别出的字会不同。

3．实际识别

根据模板中，各个字段的位置定义，计算出图像中需要识别的矩阵位置，并提交Tesseract识别。

此步骤相对简单，但需要考虑“串行”。如套打发票中，印刷部分与打印部分的位置错位。

四．TechOcr V0.1完成部分

没有考虑旋转；

预识别中，选定的特征字较少，造成部分照片未找到特征字；

提交矩形识别时，未考虑串行。

五．其他技术点

Tesseract使用Leptonica作为图像处理库。Leptonica是一款很令人惊艳，但很小众的开源图像处理软件。换句话说，就是除了源代码，没什么资料。

而相对熟悉的是OpenCV图像处理库，至少网络资源丰富。所以在熟悉了两天Leptonica后，仍决定使用OpenCV。但这需要将IplImage（OpenCV格式）转换为Pix（Leptonica格式），目前已经完成。

六．业务应用场景

以上描述了TechOcr中的核心识别部分。

此部分可以作为可执行程序，单独运行；或提供动态库供其他程序调用。

另需要一个模板定义程序，定义“营业执照”、“身份证”等模板，并按XML格式存储，供核心程序使用。

根据模板定义，TechOcr可识别各类证件、发票、铭牌的图像，作为各类应用程序的辅助输入工具。