TechOcr程序进展情况汇报

近一周多来，在程序开发上进展不快，主要对图像预处理阶段的各个步骤进行了学习与验证。在原型验证程序中，完成了对汉字位置的判定；目前正在考虑如何智能的区分字段名与字段值。

**一．已经完成的工作内容**

**1．光线对图像的影响**

从图像学术语讲，就是“颜色的恒常性”问题，一个比较偏门的专业领域。即不同的光线，照在同一的物件上，得到的图像上会有不同的色彩值反应。在预处理时需要消除这种条件的影响。

在解决这一问题时，发现的另一个情况是：肉眼看上去清晰的图像，有时候并不一定适合计算机识别；而有些肉眼看上去不舒服的图像，反倒识别率高。

根据网络资源进行“颜色恒常性”处理后，其结果图像有不规则的色彩跳跃。所以根据算法自行写出函数后，得到了看上去更舒服的图像，但在进一步的处理中，效果不如前者。

**2．噪声对图像的影响**

针对每一张图像，都能找到一组适合的滤波。但找不到一组滤波，适合所有的图像。

目前选取了一组折中的滤波组合。考虑过智能识别滤波效果，但这必然会影响效率，暂时没有采用。

**3．DLL调用时的堆错误**

这是一个纯软件开发问题，与图像学无关。反映的现象是，在对std容器进行析构时，报内存不可访问错误。

其根本原因是调用程序（自己的DLL程序），与被调用的DLL（OpenCV的DLL程序）使用了不同的进程空间，而内存分配与释放，分别由不同部分的程序完成。

为此改变了整个工程的配置，将对OpenCV的调用，从动态DLL调用，改变为静态LIB调用，期间填了无数的坑。

综上所述，近一周多的时间内，完成了Tesseract OCR识别程序中的一个GetWords函数原本的工作。实现了对汉字位置的判定。

**二．正在进行的工作**

可以把对知识的认识程度分为四个阶段：不知道自己不知道、知道自己不知道、不知道自己知道、知道自己知道。

我对图像处理的知识，从一个月前的第一阶段，已经步入第二阶段了。

已经完成的部分，其策略是“不遗漏”，下一步是要在其中找到真正有效的部分。包括版心判定、字段名与字段值的区分。