# PaperTableRecognition项目开题报告

摘要

工业上很多场合需要纸质文档记录生产数据，然后再通过人工手动录入电子表格。虽然现在都向自动化生产，但是纸质记录还是必不可少的手段。纸质记录数据有很多优势，灵活性高，方便小量数据采集，可以针对个体用户，不需要系统集成。

PaperTableRecognition项目和其他项目有很多共同技术，比如车牌识别，电子抄表，OCR等。所以该项目不仅可以解决实际工业应用问题，还可以研究相关技术。

下面举例说明应用实例，某标准马达有限公司，研发人员需要对新设计的马达做很多实验，操作人员需要记录实验数据，然后再把记录的纸档实验数据交由研发工程师，然后由工程师助理将纸档实验数据手工录入电子档，然后再提交给工程师。有的时候，录入电子档的工作是由工程师自己完成的，需要耗费相当的时间成本。

关键技术：OpenCV,Tesseract-OCR,Machine learning,ImageProcess,ZBar;

## 方案设计

该系统的使用是针对于程式化的纸质表格，也就是说，待识别的表格是预先在系统中有记录的，系统预先知道该文档的存在，在表格的左上角有一个文档编号，文档编号的形式可以是条形码，二维码，数字序列等。使用的时候，将纸档表格扫描成图片，系统识别到文档编号后，系统就可以获知该表格各个单元格所在位置，数据形式等信息，系统可以针对性的对ROI进行识别。同时，系统在识别到文档编号后，会根据文档的类型，创建一个表格文档（EXCEL），在对ROI识别的数据会填入到EXCEL相应ROI所在位置。从而实现纸档表格的数据识别。系统对ROI的识别采用OpenCV，对字符的识别采用Tesseract-OCR，界面开发可采用MFC等。

这种方案重点是识别程式化的表格，针对ROI进行字符识别并填入相应的EXCEL中，可有限提高识别准确率和正确率。

### 程式化表格设计

表格的图像元素主要有文档编号（条形码或者二维码），单元格，边框，单元格内的说明文字。如下图。



其中文档编号最好以条形码或者二维码的方式存在，目的是为了在图像识别时，可以自动识别到原图中唯一的条形码，从而可以确认该文档的类型。如果仅仅是数字序列，系统会遇到2个问题：1.系统无法锁定哪一个数字序列才是文档编号；2.如果固定数字序列所在图片上的坐标位置，貌似可以解决1中的问题，但是系统如何精确定位数字序列的坐标位置是一个难题。

文档必须要一个边框，这个边框用来识别图片整体坐标，来匹配上面条形码指定的表格的坐标尺寸，这样才能确保图片中的表格可以完全映射到指定的表格中。我把这个边框命名为定位边框，定位边框越大越好，误差越小，定位越精确。所以我把文档最外围的边框作为定位边框。

空白文档中的文字信息是不需要识别的，因为通过文档编号条形码和定位边框，已经可以精确确定图片中需要识别的区域。并且知道识别的ROI内容将要填到哪一个单元格内。

### 识别文档编号

每一个程式化文档都有且仅有一个文档编号，文档编号用来识别该文档的样式，在系统中该编号是唯一的，相当于个人身份证。系统首先扫描图片中的条形码，可以使用ZBar，只要图片中存在条形码即可识别。识别后立即从系统数据库中查询该序列号的文档，然后生成一个.xlsx文件到指定目录下，作为识别生成的空白文档。

### 识别定位边框