# 武汉大学教学实验报告

电子信息学院 通信工程 专业 2021 年 10 月 5 日

实验名称 表单识别 指导教师 卜方玲

姓名 周轩洋 年级 2019 学号 2019302120083 成绩 ____

一、 预习部分

1. 实验目的

2. 实验基本原理

## 1、实验目的

1.学习使用 Python 进行基本的灰度、滤波、二值化等图像处理方法
2.了解图像文字识别

## 2、实验基本原理

1.cv2 库

OpenCV 是一个基于 BSD 许可（开源）发行的跨平台计算机视觉库，可以运行在 Linux、Windows、Android 和 Mac OS 操作系统上。它轻量级而且高效——由一系列 C 函数和少量 C++ 类构成，同时提供了 Python、Ruby、MATLAB 等语言的接口，实现了图像处理和计算机视觉方面的很多通用算法。 OpenCV 用 C++ 语言编写，它的主要接口也是 C++语言，但是依然保留了大量的 C 语言接口。

在计算机视觉项目的开发中，OpenCV 作为较大众的开源库，拥有了丰富的常用图像处理函数库，采用 C/C++语言编写，可以运行在 Linux/Windows/Mac 等操作系统上，能够快速的实现一些图像处理和识别的任务。此外，OpenCV 还提供了 Java、python、cuda 等的使用接口、机器学习的基础算法调用，从而使得图像处理和图像分析变得更加易于上手，让开发人员更多的精力花在算法的设计上。

本次实验在 Python3.7 环境下安装 cv2 库，实现对表单图片的灰度、二值化、膨胀腐蚀、分割等操作。

2.Tesseract-OCR

Tesseract 是惠普布里斯托实验室在 1985 到 1995 年间开发的一一个开源的 OCR 引擎，曾经在 1995 UNLV 精确度测试中名列前茅。但 1996 年后基本停止了开发。2005 年，惠普将其对外开源，2006 由 Google 对 Tesseract 进行改进、消除 Bug、优化工作。目前项目地址为:
https://github.com/tesseract-ocr/tesseract。

它与 Leptonica 图片处理库结合，可以读取各种格式的图像并将它们转化成超过 60 种语言的文本，我们还可以不断训练自己的库，使图像转换文本的能力不断增强。

本次实验使用 tessract-ocr 对分割后的表单图片进行文字识别。

## 二、 实验操作部分

### 1. 实验方法

### 2. 实验结果

### 1. 实验方法

1）实验大致流程

读取图片-->进行透视变换操作-->进行灰度、二值化处理-->识别出表单中的横线和竖线-->与操作找出交点-->按照单元格分割图片-->识别分割图片中的文字-->将识别结果写如 excel 中。

2）对表单图像的处理方法

使用 cv2 库中函数对原图片进行读取，同时进行灰度处理、找出表单的四角点以便透视变换操作；使用 imutils 中的工具进行透视变换；而后使用 cv2.adaptiveThreshold 进行二值化；对二值化处理后的图片用 cv2.getStructuringElement 和 cv2.erode 进行横线和竖线提取；对横线竖线进行与操作得到交点，由这些交点就可得出每个单元格的位置；根据交点将图片分割成为一个个的单元格；将分割后的图片通过 tessract-ocr 的接口传入进行文字识别；最后将识别出的文字写入 excel 表存储起来

### 2. 实验结果

1）读取表单文件

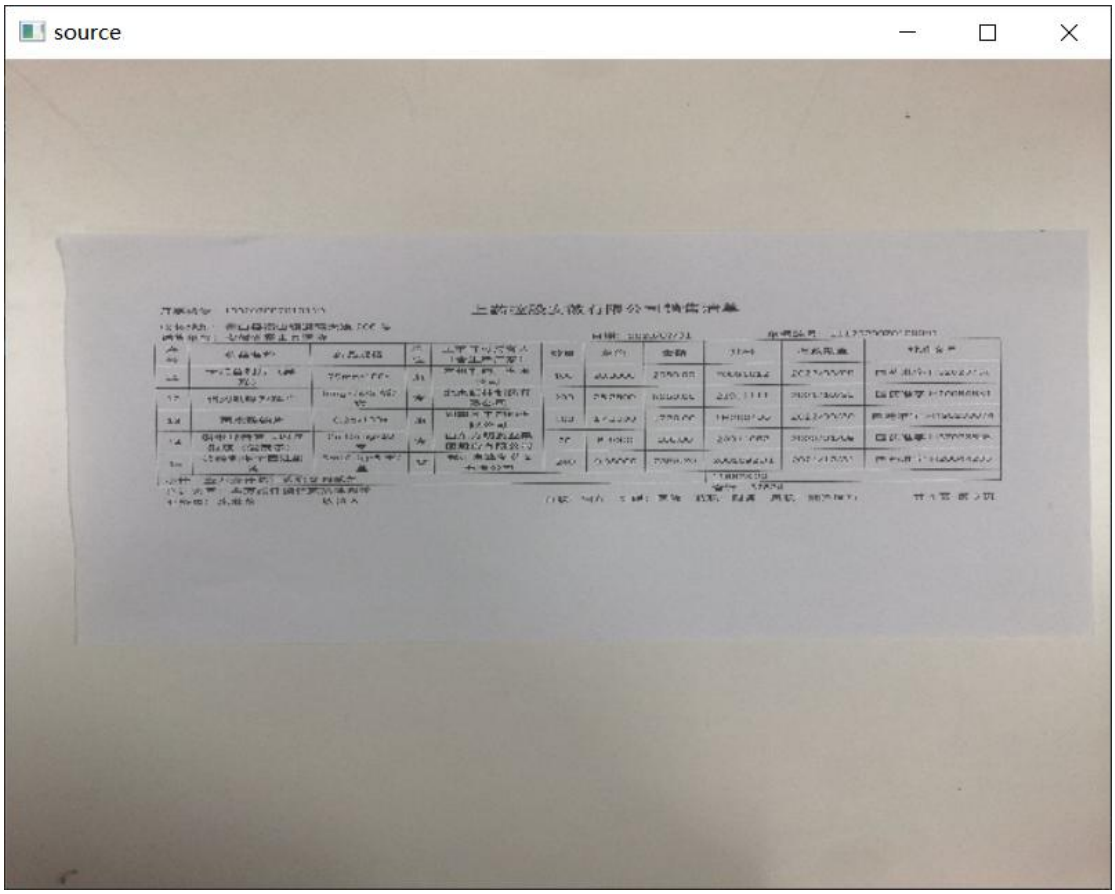


图 1 读取表单文件

2)透视处理、二值化并裁剪后

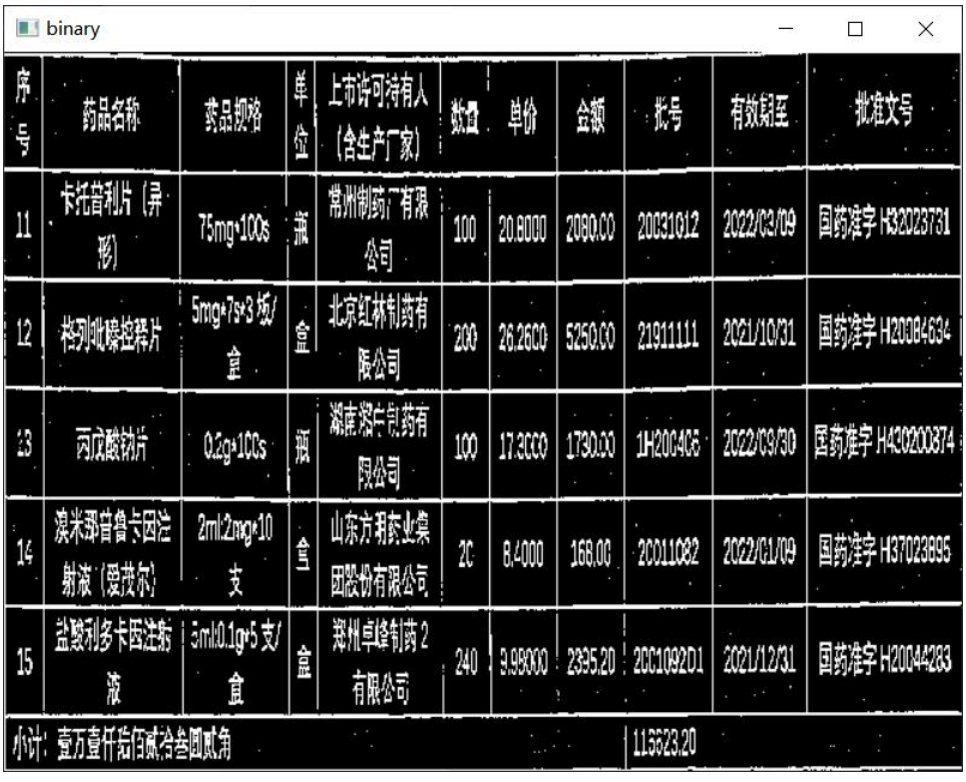


图 2

3)识别横线和竖线并求出交点

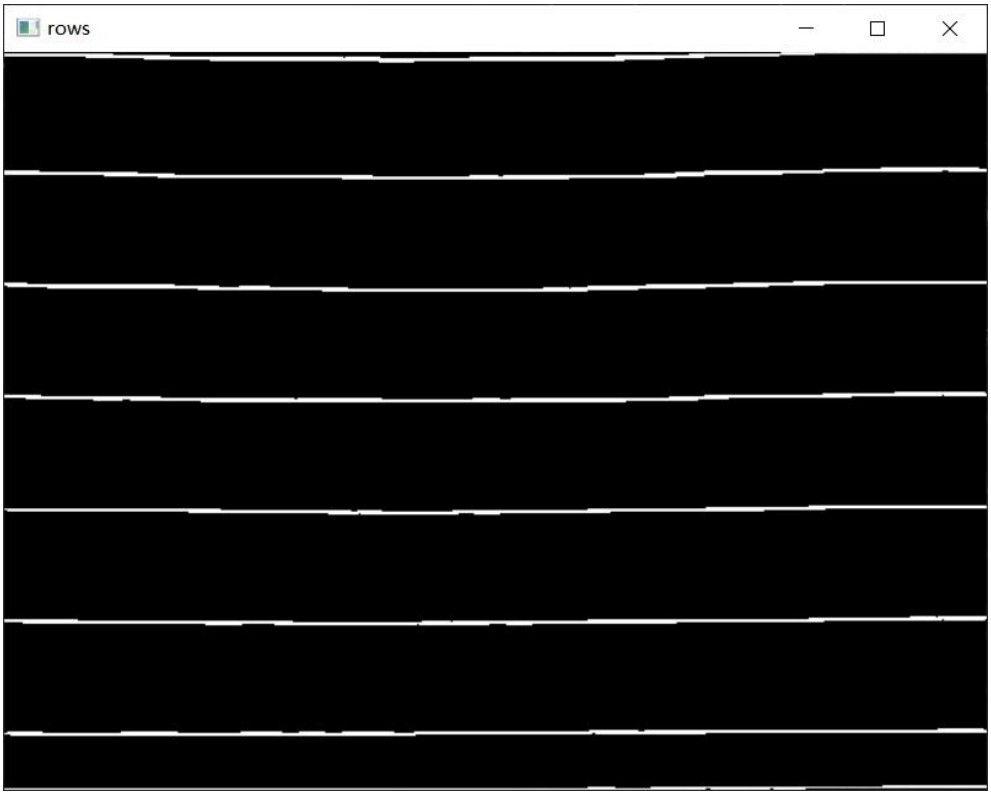


图 3 横线识别结果

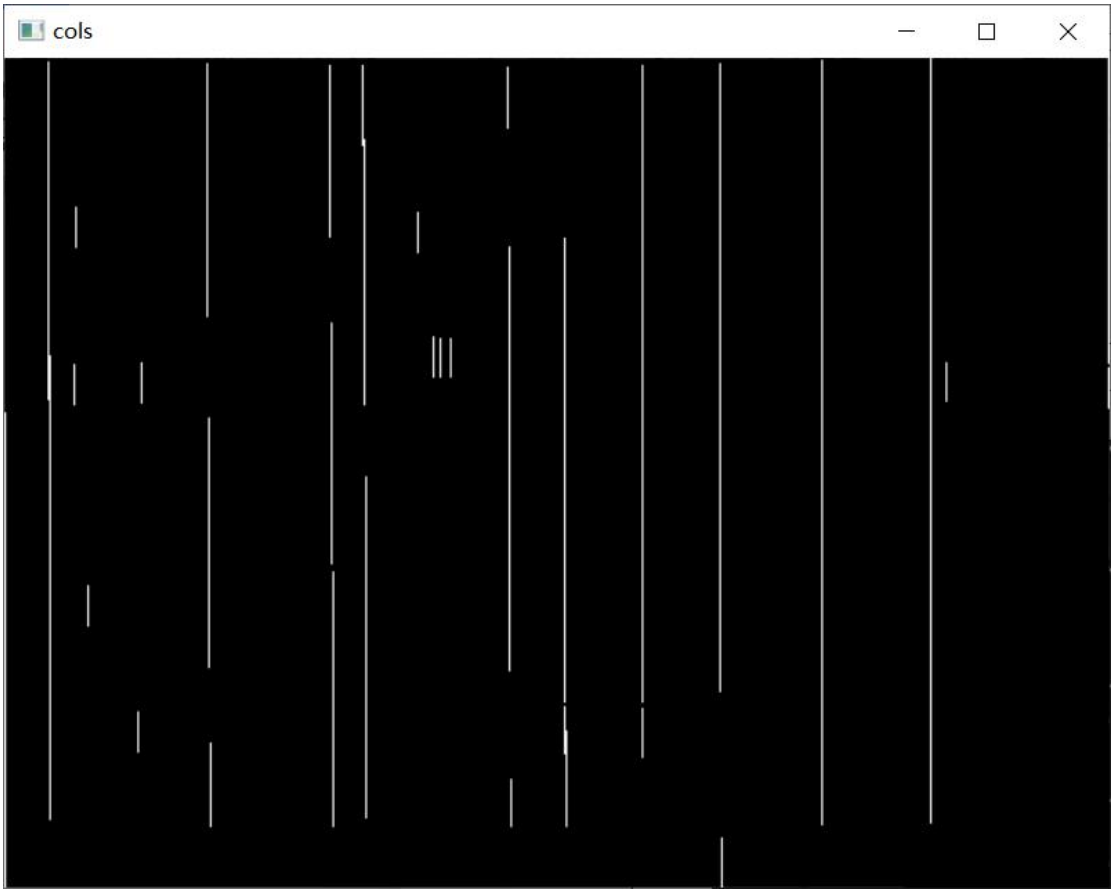


图 4 竖线识别结果

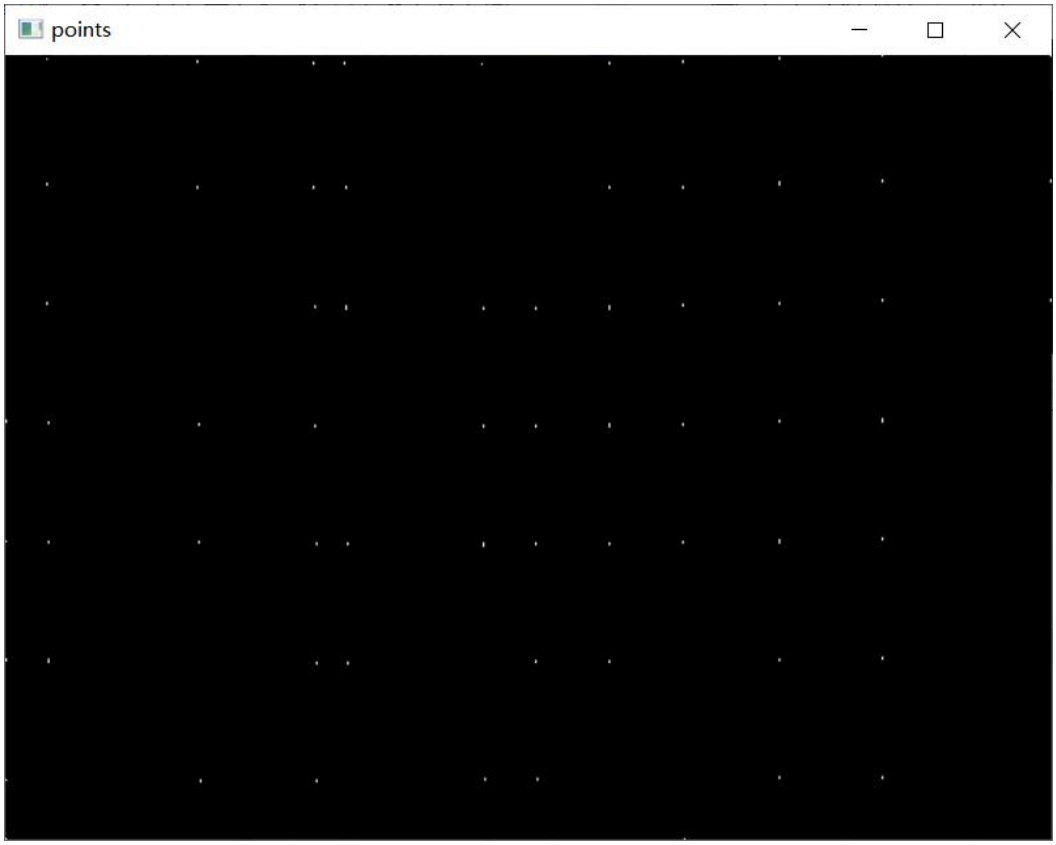


图 5 求出交点

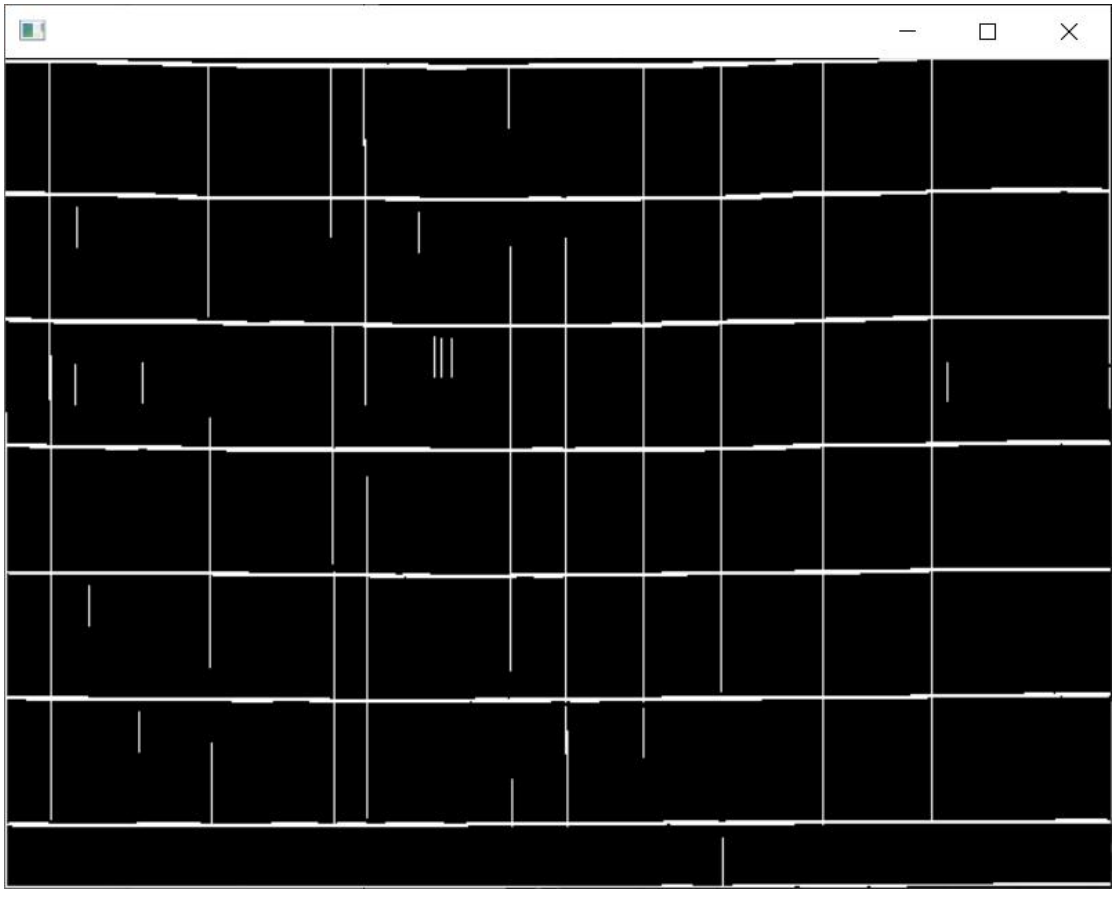

图 6 重绘表格

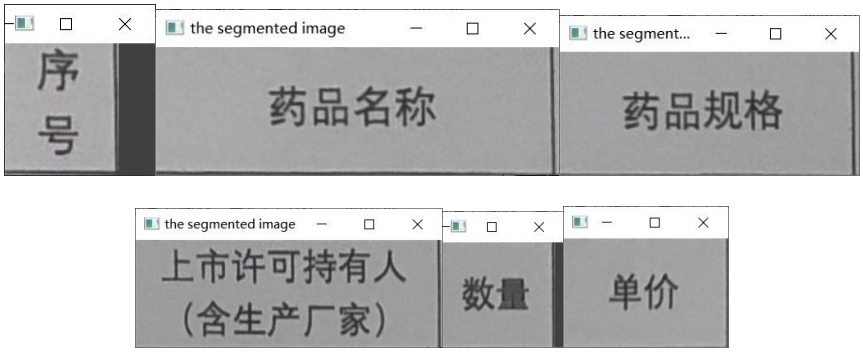


图 7 分割后的一些图片

4) 传入 tessract 进行识别

| | A | B | C | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 药品名称 | 药品规格 | 上市许可持有人(含生产厂家) | | 单价 | | | 有效期至 | 批准文号 | | |
| 2 | | 彩) | 75mgx*1005 | 常州制药厂有限 | 100 | 20.8000 | 2080.00 | 20031012 T | 2022/03/09 | "国药准字H32023731 | | |
| 3 | | 格列吡嗪控释片 | 5mg*7sS*3板/盒 | | 200 | 26.2600 | | 21911111 | 2021/10/31 | 国药准字H20084634 | | |
| 4 | | 永水丙戊酸钠片 | 0.2g*100s | 限公司 | 100 | 17.3000_ | 1730.00 | 1H200406 | 2022/09/30 | 国药准字H43020087 | | |
| 5 | | 14溴米那普鲁卡因注_"射液(爱茂尔) | 2ml2mg*10。支 | "团股份有限公司 | 20 | 8.4000 | 168.00善 | 20011082 | 2022/01/09 | 国药准字H37023895 | | |
| 6 | | 匹盐酸利多卡因注射液 | _5ml:0.1g*5支八盒 | 郑州卓峰制药2有限公司 | 240 | _9.98000 | "2395.20 | 2001092D1 | "2021/12/31zn | 国药准字H20044283 | | |
| 7 | | 刑口,壹万壹仟陆佰贰拾 | 圆毁角 | | | | | _116623.20 | | | | |

图 8 识别结果

## 三、 实验效果分析（包括仪器设备等使用效果）

1、通过二值化等操作准确地识别出了表单中的横线和竖线，由于横竖线是有一定宽度的，在用与运算求交点时，可能会出现多个交点，为了解决这个问题，本人采用了在一定范围内取第一个交点的算法，取得了较好的效果。

2、我们通过 excel 表可以看到，识别效果不是很好，在检查了单元格是否准确分割后，我发现识别错误的出现并不是因为没有准确分割单元格，而是 tessract-ocr 对中文的识别效果本身不好，为了更准确地识别中文，换成百度云-ocr 会更好。

3、由于时间仓促本人没有来得及直接将数据导入数据库，而是采用了写入 excel 表的方法，这一点可以加以改进。

## 四、 教师评语

指导教师　　年　　月　　日