图像信息识别技术方案简介

2020-07-14 星期二 12:38

1.制式文档

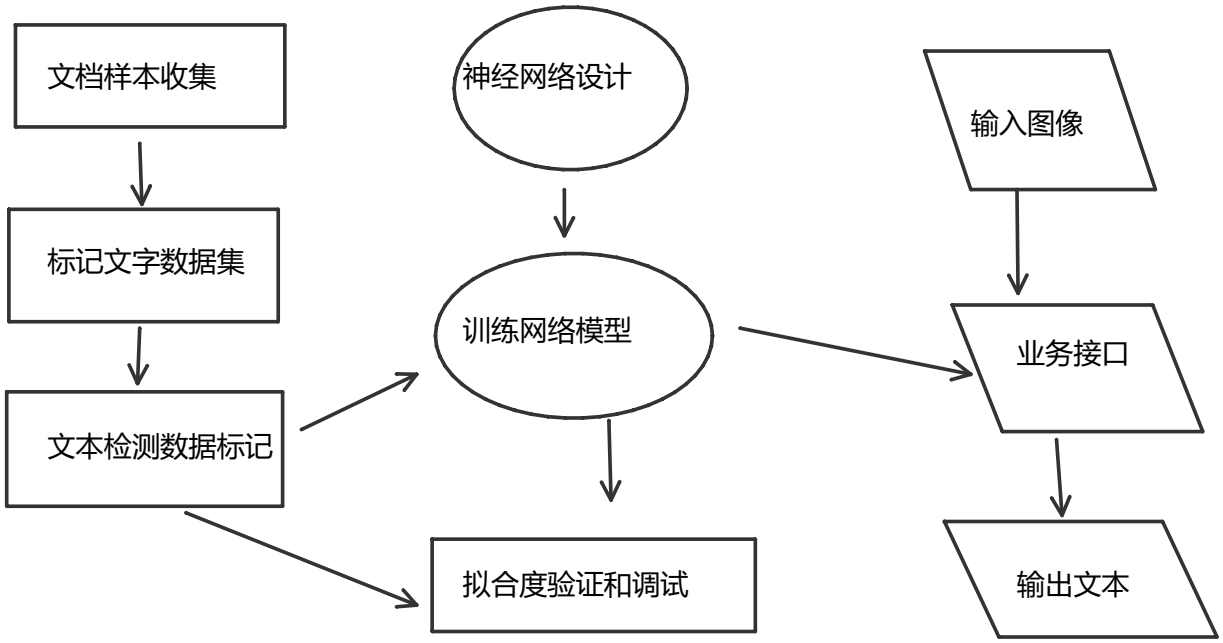
制式文档的识别，整个技术过程分为4部分。

1、首先，需要收集大量样本文档，并且每份文档的文字，需要人工标记出来每一行的每个字。

2、然后使用机器学习的一些算法（比如CNN，DNN，CTC等）进行训练模型，对训练结果验证拟合度，并且反复调整参数和验证结果，使准确率达到一个稳定的数值。

3、接下来，同上步骤，做文本目标检测算法模型。用于提取图形中的文字区域。此项工作目的在于把成页成段落的文字转换成单行文字，便于提高训练和识别效率，也能提高准确率。

4、最后是生产使用，编写数据输入输出接口，以及数据结构转换。这里主要是把生产环境中的数据，先转换成模型对应的格式，并且需要做一些降噪，压缩，归一化和过滤等预处理，保证输入的数据干净统一，防止结果出现脏数据。对于通过模型计算的结果，进行重新编码，保证输出结果和业务数据的格式统一。同时还要硬编码修复一些常规异常数据，提高系统整体的准确率。



2.表格文档

表格文档识别的核心过程是，先进行表格重建，然后调用制式文档的识别。

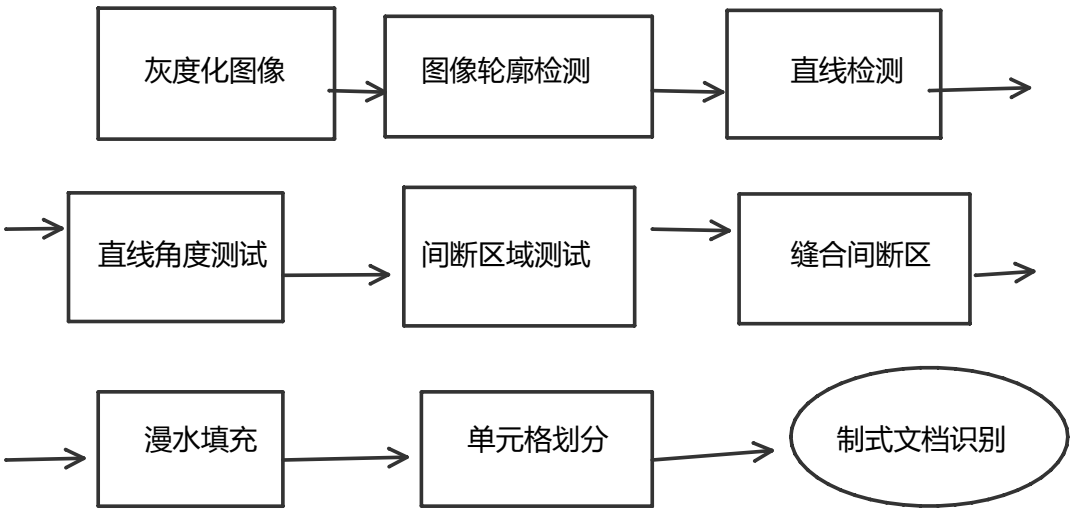
表格重建的过程，就是把图像中逻辑单元格区分开来，这样文本识别的单位就是表格单元格，而不是制式文档中的行单位。其中单元格的主要定义逻辑，就是单元格边框线段。

对于完整清晰的单元格，使用霍夫变换，就能做出直线段检测，消除文字边界的影响。这样剩下的直线拼接成的矩形，使用漫水填充识别闭包，就可以标记出每个单元格的边框。

由于扫描过程中，印刷和扫描质量的问题，导致一些表格的线条会模糊或者中断。使用霍夫变换做出的直线检测，可能会出现线段间隙中断。这样会导致单元格边框标记的过程出现错误。即水淹法会穿过线段间隙，连接到多个不同单元格。

对于以上情况，需要对间断线段进行修复。这里的思路是对霍夫变换的直线进行二次扫描。如果出现一定横坐标（纵坐标）差异(阈值)内的线段断点的纵坐标（横坐标）的间隙长度小于经验认定（一般为表格最小高度）的差值（阈值），边进行图形连接。这样就能够防止漫水填充识别的闭包连接到不同单元格。

完成以上步骤之后，对于每个单元格进行行识别，再进行行文字识别，即可完成表格文档的识别。



表格重建流水线

3.捺印识别

基于文书一般是白底黑色，红色捺印，对于彩色文档的捺印识别，最直接的办法是识别一定色差范围内的红色。

由于一些文书是复印扫描，整个文档图像会转换为灰度图。这样不能直接使用颜色区分。

根据常规经验，捺印的灰度会在白色背景和文字黑色之间。所以根据HSV的色彩模型做判断白色亮度即可。

4签章识别

如果是白底黑字，红色签章，可同捺印识别的方式，用一定范围内的红色来识别。

同理由于有些文件又有签章又有捺印，可能会出现误判。并且是黑白的灰度图，也会出现无法用红色区分的情况。

这里采用霍夫圆检测算法，直接检测文档图像中圆形的概率分布，并且通过最大最小半径，圆心最小距离等阈值过滤即可。