

**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目： 基于OCR的身份证要素提取设计与实现**

**专 业 计算机科学与技术**

**学 生 高尚**

**学 号 20161104492**

**指导教师 李艳玲教授**

**日 期 2019.11.16**

**计算机科学技术学院制**

1. **课题来源及研究的目的和意义**

课题来自Data Fountain 竞赛题目。

光学字符识别是属于图型识别的一门学科，其目的就是要让计算机识别图像中的文字并将其转换成机器可以存储的表示形式。OCR 技术由来已久，1929 年，德国科学家 Tausheck 首先提出了 OCR 的概念，并申请了专利。几年后，美国科学家 Handel 也提出了对 OCR 技术的想法。这些想法在计算机诞生后得以实现，OCR 就是利用光学技术对文字进行扫描，将其转化成计算机内码。1960 至 1970 年，各国相继开始对 OCR 的研究，最初的识别以数字为主即 0 至 9 的数字。1977年，东芝综合研究所研制的文字识别系统已经可以识别 2000 个单体印刷汉字。

80 年代初，日本武藏野电器研究所研制的文字识别系统可以识别 2300 个多体印刷汉字。 同国外相比，我国的 OCR 技术 70 年代才开始研究，当时主要识别的字符是数字、英文字母以及符号。80 年代左右，我国政府开始对汉字识别给予了充分的重视和支持，经过科研人员十余年的努力，OCR 技术取得长足的进步。从单体识别发展到多种字体混排的多体识别，从中文识别发展到中英混排的双语识别并支持简、繁体汉字识别。对于简单规范排版的图像，识别率已经达到了 98% 以上。

现在，OCR 已经逐步进入了人们生活、工作、学习等各个领域。如停车场门口的车牌自动识别系统，票据图像的数字化转化处理，电商货物照片的图像审核处理等。随着我国信息化水平的提高，OCR 技术的应用前景将更加地广阔。就目前从行业需求来看，金融、工商、电子商务等行业对信息识别的需求已经越来越广泛，促进了识别技术的大规模应用。而个人消费者对资料电子化、手写识别技术等各方面需求则拓展了 OCR 识别技术在这一领域的应用之路，另一方面，网络时代的高速发展使个人资料电子化、商务办公自动化等需求的呼声也变得越来越高。

基于以上需求，识别图像中的文字信息是一个具有广泛应用前途的研究方向，具有非常强的现实意义。

1. **国内外在该方向的研究现状及分析**

经过几十年的沉淀，OCR 技术取得了长足的进步，该技术已经从识别简单的 0 到 9 这十维数据，发展到识别几千甚至上万维汉字的水平，字符维度的量级、字符的语种、字符的字体差异都已经不再是问题。国外的 OCR 技术应用相对成熟，包括 IBM、Motorola、HP 和 Microsoft 等世界型大公司都陆续展开了这方面的研究，在他们的产品中绑定 OCR 技术。光符识别设备把报纸、杂志和其它印刷材料上的印刷字符读取到计算机内存。OCR 软件可以和任何流行的操作系统一起使用。除了识别印刷字符之外，OCR 也可能识别栏目布局， 这种布局会在报纸中出现。OCR 技术在学术界和工业界上也取得了很大的关注和发展，并不断向生活的各个方向延伸，拓展不同的业务场景。谷歌的 OCR 开源产品Tesseract 可以识别单行或多行的文本图像，其操作简单、效果稳定的特点使其成为开源界很好的字符识别软件之一。除此以外，腾讯、百度、阿里巴巴等公司也都推出了自己的 OCR 产品，可以做到在云端在线识别不同场景图像中的字符。

目前，OCR 技术所处理的图像场景主要分两类即文本图像和自然场景图像。本课题处理的图像均为文本图像，且本课题处理的文字绝大多数是中文，查阅文献后，

国外研究专家提出先将句子分割成单个字符，之后将分割好的字符进行逐个与字符库识别。

1. **主要研究内容**

本课题的目标是将经OCR扫描的身份证图片识别出来，即识别身份证扫描图像中的各要素文字。 需要学习各种图像检测、识别的算法模型，总结他们的特点，根据研究的实际业务场景来选择最具有针对性，算法能力最强， 运行效率最高的算法。

在保证准确率的前提下，提升整套算法的运行效率。由于 OCR 的重点在于文字的识别，本研究会更倾向于保证文字识别的准确率，因此会选择研究深度较大的模型进行训练预测。

1. **研究方案**

OCR 算法一般包括三个步骤，第一步是特征提取即提取图像中可能是文字的候选区域：1）采用Open Source Computer Vision Library开源库对图像进行灰度、二值化处理；2）在字符区域图像呈现出白色，间隙区域为黑色。第二步是对图像中的文字区域进行定位检测：1）通过水平投影的方式将字符区域与间隙其余分割开，确定字符所在行；2）使用垂直投影分割法，将字符从二值化图像的每一行中分割出来。第三步是对上一步检测出来的区域进行文字识别：1）采用Lenet模型，并根据测试结果改进；2）采用Caffe或TensorFlow深度学习工具训练模型；其中前两步是文字识别前的准备工作，而第三步是整个 OCR研究的核心内容。

1. **进度安排**

1．10月～12月：充分准备课题所需基础知识（包括但不限于：Python语言处理图像，Open Source Computer Vision Library开源库，机器学习.吴恩达，图像存储格式，读写像素，深度学习，深层卷积神经网络，掌握一种开源框架，模型调参及精度优化）

2．12月～1月：完成模型测试与改进，撰写论文

3．1月～2月：完成项目整体代码，应用软件开发流程编写演示软件，撰写论文。

4．2月～3月：完成模型最优解，演示软件编写及测试完毕，毕业论文校正内容、格式错误。

1. **研究过程中可能遇到的困难和问题，解决的措施**

1.研究过程中可能遇到1）图像数据中的文字识别效果不理想；2）模型的精度值未达到预定目标；3）样本数据数量的局限性；4）编写演示软件时的开发流程问题。

2.解决措施：调研各种文献，学习各种图像检测、识别的算法模型，总结他们的特点，选择最具有针对性，算法能力最强，运行效率最高的算法。

1. **主要参考文献**

[1] T.Wang, D. J.Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional

neural networks. In ICPR, 2012. 1, 6, 7

[2] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with

convolutional neural networks. IJCV (Accepted), 2015. 1, 2, 3, 6, 7

[3]汪惠权. 用于发票处理的数字图像处理系统的研究与开发[D].东华大学,2018.

[4]苏东出,孙萍.基于阿里云的身份信息采集系统的实现[J].内蒙古科技与经济,2018(17):71-73.

[5]李阳,桂强,张文秀.光学识别技术在病案数字图像归档工作中的应用[J].中国病案,2019,20(06):6-8.

[6]韩志超. 基于智能手机的身份证信息识别系统的研究与实现[D].广西师范大学,2018.

[7]汪一文. 深度卷积神经网络在OCR问题中的应用研究[D].电子科技大学,2018.

[8]刘泳文. 基于图像识别的搜题系统的研究与实现[D].西华师范大学,2016.

[9]饶凌云. 证件图像识别技术研究与应用[D].中南大学,2013.

[10]田大增. 视觉文档图像识别预处理[D].河北大学,2007.

[11] R. B. Girshick, “Fast r-cnn,” international conference on computer vision, pp. 1440–1448, 2015.