

**计算机科学技术学院**

**本科生毕业论文**

**基于OCR的身份证要素提取设计与实现**

**班 级： 云计算数据中心运营**

**学 号： 20161104492**

**姓 名： 高尚**

**指导教师： 李艳玲**

**2020 年 3 月 1 日**

毕业论文目录

[1 绪论 1](#_Toc13518)

[1.1 研究背景 1](#_Toc12950)

[1.2 研究难点 2](#_Toc7764)

[1.3 研究内容 2](#_Toc11715)

[2 系统需求分析 2](#_Toc16813)

[2.1 系统后端处理流程 2](#_Toc20670)

[2.2 系统前端处理流程 3](#_Toc20174)

[2.3 开发语言及工具 3](#_Toc28613)

[3 系统后端设计 3](#_Toc9368)

[3.1 图像处理 3](#_Toc17259)

[3.2 文本处理 7](#_Toc25019)

[3.3 文字识别 8](#_Toc15969)

[3.4 结果纠正 11](#_Toc18913)

[4 系统前端实现 12](#_Toc10870)

[4.1 函数设计 12](#_Toc2340)

[4.2 网页设计 13](#_Toc3783)

[5 系统测试 13](#_Toc28681)

[5.1 主界面功能 13](#_Toc12337)

[5.2 结果页功能 14](#_Toc18046)

[6 总结 14](#_Toc5101)

[致 谢 15](#_Toc8658)

[参考文献 15](#_Toc14851)

全文共 18 页 7264 字

基于OCR的身份证要素提取设计与实现

计算机科学技术学院 2016级云运营 高尚 20161104492

指导老师 李艳玲 副教授

摘要 本文针对OCR系统中身份证复印件设计一套身份证要素信息提取系统，系统采用深度卷积神经网络技术进行文字识别，系统共分为前端和后端两部分，前端系统负责上传图片和展示识别结果，后端系统负责图像处理和文字识别。提取系统具备低耦合特点，各个子模块可以独立运行，提取系统经数据集测试，识别准确率为90%。

关键词 OCR；文字识别；深度卷积神经网络

1 绪论

* 1. 研究背景

文字是人们日常交流使用最多的形式之一，随着近些年互联网和智能手机的蓬勃发展，人们的沟通方式逐渐变得多样化，从单一的文字形式拓展为语音和图像等多种数据形式，面对复杂的语音数据和图像数据，计算机固有的处理方式逐渐变得低效，如何让计算机看懂和理解图像是提高信息处理效率的关键。

光学字符识别(Optical Character Recognition，简称OCR)是指利用扫描仪等电子设备拍摄图像，通过检测图像中形状和亮度的变化识别出字符，图像中的字符提取为文本格式后，可供文字处理软件进行编辑加工。传统OCR识别系统使用模板匹配和几何特征抽取的方法进行识别文字，其主要应用方向为从扫描文档中提取文字信息，系统的处理流程包括:图像预处理、图像二值化、单字符切割、连接识别结果等。

中文具有类别多、结构复杂的特点，传统OCR识别系统中单字符切割模块难以正确识别出汉字，例如，“从”误分为“人　人”，“双”误分为“又 又”等。深度学习是基于机器学习的一项人工智能技术，它可以让计算机在得到数据时具备和人脑一样的学习能力，通过学习海量数据的不同特征，进而掌握不同的能力。

得益于深度学习在图像领域的快速发力，卷积神经网络技术(Convolutional Neural Networks，简称CNN)和连接时序分类技术(Connectionist Temporal Classification，简称CTC)的组合可以实现端到端的不定长文本识别。

* 1. 研究难点

身份证信息的OCR识别作为成熟的人工智能应用之一，广泛应用于用户注册、银行开户、交通出行等多种场景，大幅提升了信息的处理效率。但在某些场景下的身份证OCR识别仍然存在着一些问题，例如，复印件中的身份证信息提取，除光照因素外，还受到复印质量、水印盖章等多种外界因素影响，使得通用身份证识别系统很难达到满意的结果。

CCF大数据与计算智能大赛（CCF Big Data & Computing Intelligence Contest，简称CCF BDCI）是由中国计算机学会大数据专家委员会于2013年创办的国际化智能算法、创新应用和大数据系统大型挑战赛事。此次比赛提出了基于OCR的身份证要素提取[[[1]](#footnote-1)]算法赛题，赛题目标是设计出一个具备抗噪声干扰能力强的OCR模型，并且能够准确识别身份证中姓名、地址、身份证号码和身份证有效日期等信息。赛题提供的数据集相比普通的身份证信息，具有以下难点：（1）图像清晰度参差不齐；（2）图像中的部分要素信息被加盖的水印遮挡；（3）复印件中身份证的位置、方向较为随意；（4）训练集数据未提供各要素的位置信息，只提供所有要素的文本内容。

* 1. 研究内容

针对赛题数据的特点，本文提出一套针对身份证复印件信息提取的技术方案，整体思路如下：（1）定位身份证正反面顶点信息提取身份证区域；（2）根据各要素相对位置提取各要素区域；（3）识别文字并对结果进行校正，输出识别结果。

本方案主要研究内容包括：（1）对复印件图片进行图像切割、图像旋转校正和去除水印等图像处理方法；（2）使用模板匹配技术定位身份证各要素信息位置；（3）使用中文合成数据集预训练文字识别模型，在比赛数据集上训练模型微调得到最终文字识别模型；（4）使用集束搜索（Beam Search）优化各要素识别结果。

1. 系统需求分析
   1. 系统后端处理流程

后端系统包括三个模块：（1）图像处理：检测身份证顶点模块；（2）文本处理：检测文本行模块；（3）文字识别：识别文字信息模块；处理流程如图2-1：

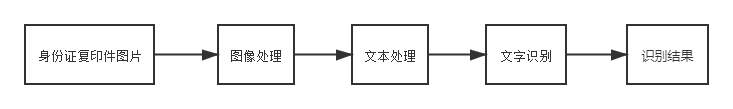


图2-1 后端处理流程

* 1. 系统前端处理流程

前端系统包括两个界面：（1）WEB端——首页：上传身份证复印件图像；（2）WEB端——结果展示页：显示文字识别结果；处理流程如图2-2：



图2-2 前端处理流程

* 1. 开发语言及工具

后端系统采用Python编程语言搭建，Python语言具有简单、速度快和丰富的第三方库等优点，可快速的构建深度学习应用。

前端系统采用Flask轻量级WEB应用框架，相比同类应用框架，Flask更轻便、更灵活，因此，它又被称为“微框架”：在保持代码简洁的同时快速实现网站的搭建。Flask核心函数包括Werkzeug和Jinja2，Werkzeug支持URL路由请求集成；可以处理HTTP基本事务，快速响应客户端推送过来的访问请求；Jinja2支持自动HTML转移功能，页面加载过程会将源码进行编译形成python字节码，从而实现模板的高效运行。Flask工作流程如图2-3：



图2-3 Flask工作流程

1. 系统后端设计
   1. 图像处理

图文识别的准确率一定程度上依赖图片处理结果的好坏，图片处理的目的在于剔除图片中噪声数据，凸显待识别区域；好的处理结果可以有效提高识别准确率。目前，常用的图像处理技术包括:图像滤波、图像二值化、图像目标检测、图像形态学操作。

* + 1. 图像锐化

图像滤波是图片预处理中常用的方法，它需要在不丢失图像细节特征的条件下对目标图像的噪声进行抑制。滤波操作的关键在于选择一个可以消除噪声的滤波器，在选择时需要考虑两点：能有效地去除背景中的噪声；能很好地保护目标图像的形状、大小和独特的几何结构。通过对数据集的分析，图片有不同程度的模糊现象，直接输入模型会难以提取身份证轮廓信息，本文使用锐化内核[3]解决模糊现象，锐化处理是使图像灰度反差增强，从而使模糊图像变得清晰。锐化对比如图3-1：

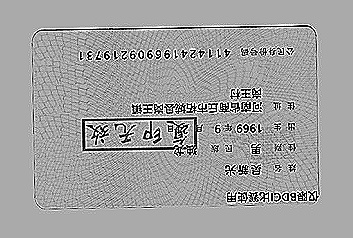
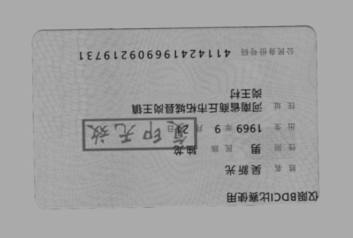


图3-1 锐化前后对比图

* + 1. 图像二值化

图像二值化是将灰度图中的像素值设置为0或者255，使整个图片呈现出黑白效果，实现目标区域与背景区域的分离，经过二值化的图像，其数据量减小，图像变得简单，有利于后续的图像处理。本文二值化方法使用openCV开源库中的adaptive Threshold函数，它是一种局部性阈值，可以动态自适应地调节属于自己像素点的阈值。它通过人为设定一个区域，在此区域内比较各像素点数值与该区域像素平均值的大小关系确定此像素点是属于黑或者白。

* + 1. 图像目标检测和切割

定位身份证顶点的前提是检测到身份证的轮廓信息，本文轮廓检测方法采用openCV开源库中的findContours函数，因为身份证轮廓为矩形且只需要保存矩形的四顶点坐标，在参数上选择cv2.CHAIN\_APPROX\_SIMPLE压缩水平、垂直方向的元素，只保留该方向的起点和终点坐标。目标轮廓检测时输入二值图，对二值图进行腐蚀膨胀处理去除因复印质量导致的背景噪声，去除效果如图。对检测到的轮廓进行面积筛选，轮廓面积大小占复印件图像的0.05-0.25为身份证轮廓信息，否则为无效信息，筛选后得到身份证正反面区域，然后求取该区域的最小外接矩形得到身份证顶点坐标。腐蚀膨胀对比如图3-2：



图3-2 腐蚀膨胀操作

设置身份证的尺寸为445\*280，根据上述得到的身份证顶点坐标和预设尺寸计算出透视变换矩阵，透视变换用于解决图像倾斜和不平行的问题，使用透视变化后得到身份证正反面图像。切割过程如图3-3：

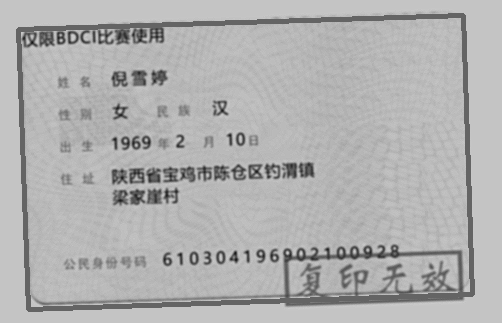


图3-3 切割过程

* + 1. 图像方向矫正

图像匹配的目的是寻找图像中的特定元素，进而确定图像中的水印位置、字符位置。主要技术包括:相似性测度匹配方法、模板匹配方法、尺度不变特征变换，本文采用模板匹配方法。模板匹配需要提前制作特定元素的模板图像，进行模板匹配时使用模板图像从图像左上角（从左到右，从上到下）开始移动，计算图像中每个像素与模板图像的相似度，相似度最大值的位置即为匹配到模板图像的准确位置。

数据集中的身份证图片左上角标有＂仅限BDCI比赛使用＂的特殊logo，这可以作为判断身份证放置方向的标志。针对此特殊标志，首先提取logo区域的图像，将其制作为156\*26的模板图像，模板匹配的全图区域更改限定为图片的左上角和右下角区域，分别计算模板图像与不同区域的相似度，如果图片右下角区域的最大相似度值大于左上角，说明匹配到身份证图片右下角有特殊logo，而这特殊logo本应在左上角，进行图像翻转为正确方向。方向矫正如图3-4：



图3-4 方向校正

身份证正面与反面拥有不同的要素信息，在识别身份证正反面部分仍采用模板匹配方法。首先在数据集中随机选择两张身份证正反面图片，然后使用拉普拉斯函数计算得到身份证正反面的模板图像。拉普拉斯函数是图像锐化操作的一种[4]，它可以增强图像的细节和图像中灰度突变的区域，提高图像的对比度，使用拉普拉斯函数处理后的图像可以保留丰富的边缘信息。接着将模板图像分别与待匹配图像进行全图计算得到相似度总和，根据相似度大小关系识别出身份证的正反面。拉普拉斯模板图像如图3-5：



图3-5 拉普拉斯模板图像

* + 1. 图像去除水印

图片中的印章水印是此次图文识别任务的一个难点，水印的位置不固定且有一定几率遮挡到文字部分，影响后续的识别效果；本文采用灰度值线性变换方法去除水印。首先利用模板匹配精确定位到水印位置，水印灰度值为a,不透明度为alpha，transparency为水印的透明通道信息，无水印的图片灰度值为b，叠加水印后的图片灰度值为y，叠加公式为（3-1），通过公式（3-2）计算得到无水印图片的灰度值b，最后将灰度值还原到水印图片中去除水印。去水印效果如图3-6：

 （3-1）

 （3-2）

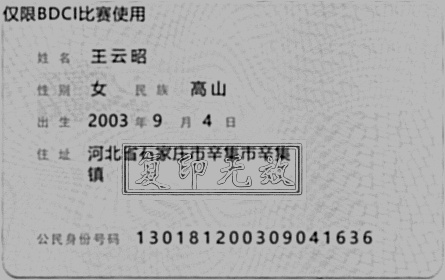


图3-6 去水印效果对比

* + 1. 图像处理流程图

综上所述，图片处理模块需要用到图像锐化、图像二值化、图像目标检测和切割、图像方向校正、图像去除水印等技术，各个图像处理技术使用顺序如图3-7：

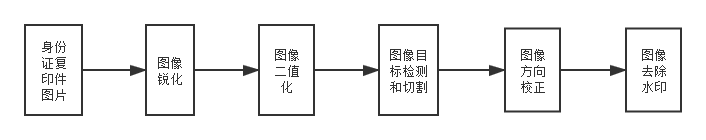


图3-7 图像处理流程

* 1. 文本处理

通过上述处理，得到干净的身份证正反面图像。身份证页面格式固定，根据身份证正反面模板图像确定各要素区域的相对位置和区域长度信息，这些位置信息有助于确定图像中各要素的位置范围，提高切分文本行图像的准确性。

将身份证正反面图像与模板图像进行匹配，在匹配过程中，会得到两个值，一个是最高响应值代表两者的匹配程度，另一个是最高响应值的坐标代表目标图像相对于模板的位置偏差，根据位置偏差值修正各要素的坐标信息，实现精准切分文本行图像。切分效果如图3-8：

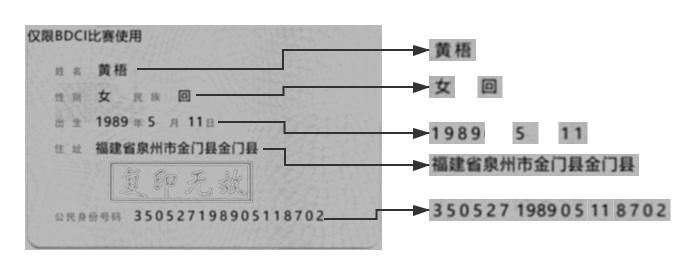




图3-8 文本行图像

* 1. 文字识别
     1. 深度学习

深度学习是机器学习的分支，由多层感知器搭建的深度学习网络可以组合低层特征形成抽象的高层特征，识别出数据的分布式特征，实现既定的任务要求。深度学习的原理在于模拟人脑进行分析学习和解释数据，它在2012年的ImageNet比赛中大放光彩，一举将图像分类的正确率提高了约10%[1],此后，深度学习被广泛应用在图像和视觉领域，大大超越了传统机器学习所取得成绩。

TensorFlow是由Google主导开发的深度学习框架，是目前主流的深度学习框架之一，它拥有良好的延展性和支持分布式系统，在多GPU和多主机上拥有极强的性能，随着Google不断更新，现在使用TensorFlow框架的应用可快速部署在各类服务器、PC终端和网页并且支持GPU和TPU高性能数据运算。

Torch是由Facebook主导开发的深度学习框架，在Python平台上名为PyTorch，它拥有GPU加速的张量计算和包含自动求导系统的深度神经网络，同时它还有入门简单和快速高效的特性受到广大研究人员的喜爱[2]，本文采用PyTorch深度学习框架。

* + 1. 卷积神经网络结构

卷积神经网络(Convolutional Neural Network,简称CNN)，是一种前馈型神经网络结构，主要结构包括卷积层、池化层[1]。卷积层用于提取输入数据的特征图，不同的卷积核可以提取到不同的特征图，通过多种特征图准确学习输入数据的特征，它的基本计算流程是对输入数据的局部矩阵和卷积核各个位置上的元素相乘，然后相加所有结果得到卷积层结果。计算过程如图3-9：

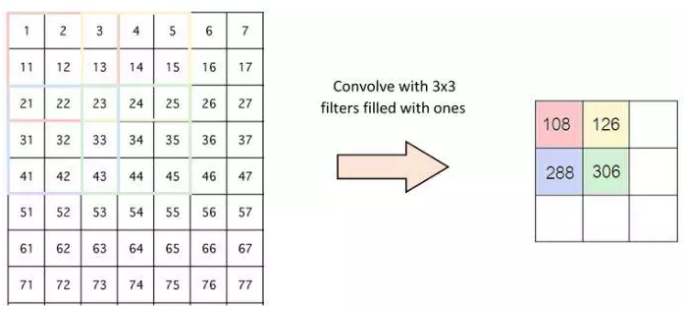


图3-9 卷积层运算过程

池化层又称采样层，用于对特征图进行降维，减少卷积层的参数数量和压缩数据，提高模型的容错性，主要分为最大池化和平均池化，本文采用最大池化。计算过程如图3-10：

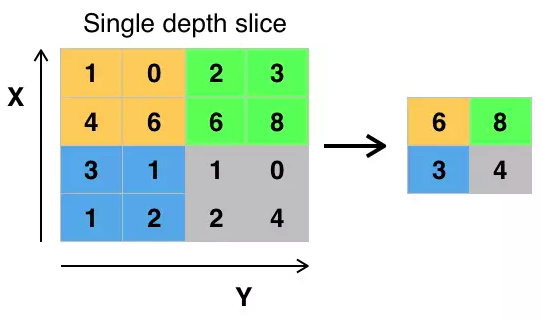


图3-10 池化层运算过程

* + 1. 识别模型

输入模型的图像尺寸为1\*22\*220，因为身份证信息为单行文本且文本长度不固定，所以图像的宽度一定大于高度。Dropout是用于神经网络训练的随机化技术，它会随机地将部分神经网络单元激活设置为零，在第一层使用dropout是为了增强模型的泛化能力；中间各层使用的dropout均为随机屏蔽部分神经单元，解决模型过拟合问题。模型经过卷积层和池化层提取特征图，使用1\*1的卷积核对特征图进行降维处理，缩短模型训练时间，使用BatchNorm函数对模型进行正则化，避免模型训练过程中的梯度消失问题。最后将特征图转换为类别数\*batch size\*预测字符数的特征向量并输出。模型结构如图3-11：

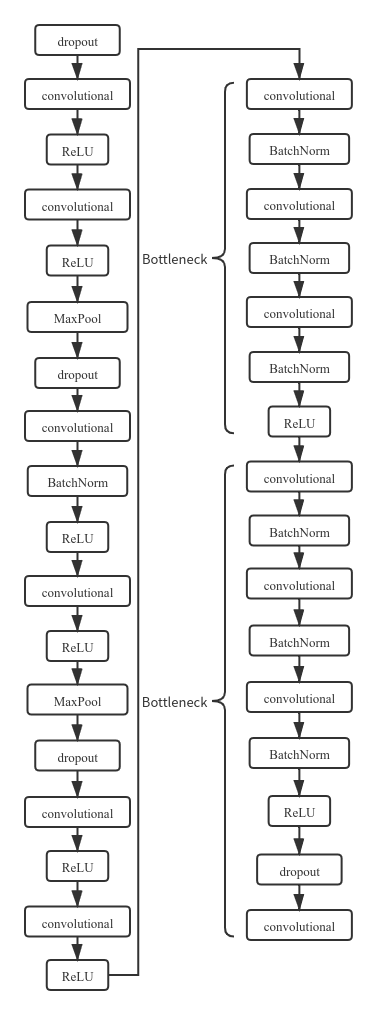


图3-11 中文识别模型

* + 1. 模型训练

模型训练时，从数据集中读取图像和标签值，标签值为中文，需要先对中文进行字典编码处理。首先建立字符字典，例如字典{王：1，张：2，高：3，……}，获取国标一二级字库作为基础字符集（6763），遍历训练集结果将基础字符集中不存在的汉字增加至字符集中，最终字符集个数为6853，覆盖了常用汉字和部分生僻字；使用已建立的字典将中文转换为数值参与模型计算。训练模型的损失函数选择CTCLoss，此损失函数用于解决模型预测值与标签值不对齐的问题，例如：标签值为“你好，世界”，而预测值为“你好好好，世界”。优化器选择均方根平均数（Root Mean Square Prop，简称RMSProp），RMSProp会根据全局学习率得到自适应学习率，在梯度较大的方向学习率会收缩，加快得到最优解的过程，学习率设置为0.1，输入模型的batch size参数为32，每50步进行一次验证，每300步进行一次模型存储。

模型训练分为两部分：（1）在中文合成数据集上预训练得到模型数据；（2）在赛题数据集上训练得到最终模型数据，训练集部分数据如图3-12。训练模型共使用约100万张中文合成数据和44955张赛题数据，最终模型损失函数稳定在10附近，模型性能经验证集测试后准确率为90%。

72728531_7792318770050a3ef3eb44cf1848bc936f6f02e200022d46a459f4364ad37943dce07c8fb

图3-12 部分中文合成数据和赛题数据

* 1. 结果纠正

针对身份证信息中性别、民族、日期等识别结果都在有限范围的类别内，采用基于词典的Beam Search方法优化识别结果。

生日信息校正：生日本身具备一定的规律，如年份必定是19或20开头，月份和日期有一定的数值范围，身份证号码中的第7-14位与出生日期的年月日是一一对应的，基于这些特点对生日信息进行校正。

身份证号码的前六位为地址码，根据公开的全国省市区级行政编码制作字典，例如：{150000 内蒙古自治区，150100 呼和浩特市，……}，因为数字相比中文识别简单，将识别出的地址码在字典中查找得到签证机关的识别结果。此方法相比识别模型直接得到的结果信息更为准确且时间更快。

有效期校正：有效期和出生日期一样，具备日期约束，例如身份证的有效期根据办理时的年龄可以分为四类：五年、十年、二十年、长期。根据这些特点，推算出对应的有效期截至日期，最终实现有效期校正。

1. 系统前端实现
   1. 函数设计
      1. 预加载模型

识别模型是系统中最重要和最耗时的模块，在WEB服务初始化的过程中进行预加载模型数据，不仅可以在前端网页中提供更快的识别速度，而且可以避免重复加载模型数据浪费服务器内存等硬件资源。加载模型流程如图4-1：

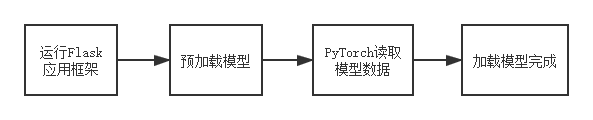


图4-1 预加载模型流程

* + 1. 文件上传

前端系统通过文件上传控件接收图像，当客户端的请求变为POST时，后端系统读取文件对象并判断上传文件是否为图片格式，若是则进行后续图像识别流程。文件上传流程如图4-2：

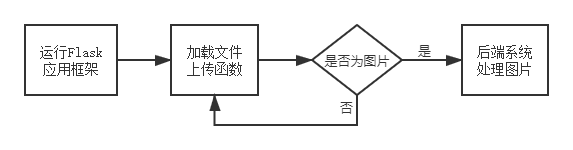


图4-2 文件上传流程

* + 1. Flask运行流程图

综上所述，Flask应用框架内有预加载模型和文件上传函数，其运行流程如图4-3：

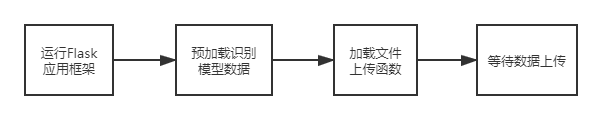


图4-3 Flask运行流程图

* 1. 网页设计
     1. 首页

首页的背景选自计算机学院网站图片，网页标题采用灰底黑字，文件上传按钮居中显示，使访问者快速获悉网站的主要功能。

* + 1. 结果页

结果页展示身份证要素信息识别结果和身份证正反面图片，便于使用者核对识别结果是否准确。

* + 1. 页面切换流程图

总上所述，前端有首页和结果页，首页收集身份证复印件图片，结果页展示文字识别结果，各页面切换流程如图4-4：

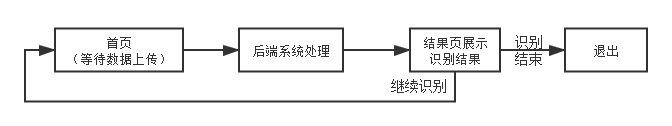


图4-4 页面切换流程图

1. 系统测试
   1. 主界面功能

首页效果如图5-1，图中左上角为网站标题，标题下方为技术和准确率介绍，在图片的中央，有两个按钮，点击第一个按钮选择图片文件，点击第二个按钮提交图片并切换到结果展示页。

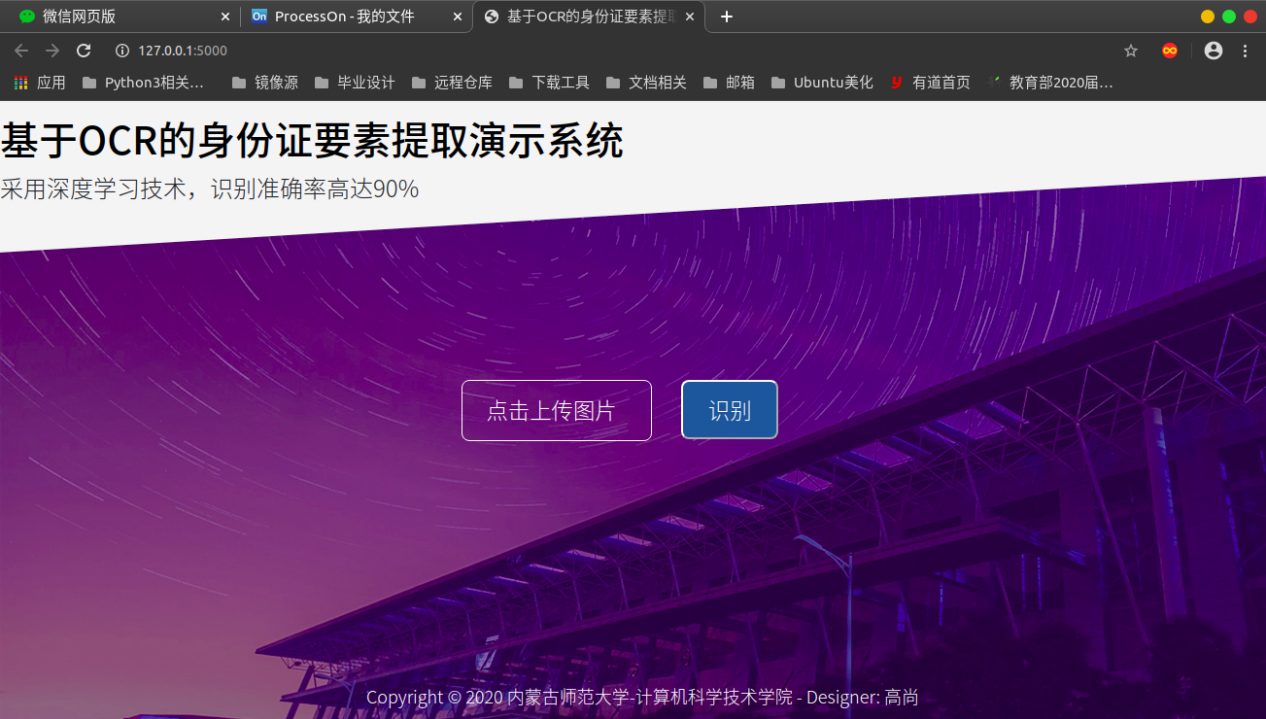


图5-1首页展示

* 1. 结果页功能

识别结果页如图5-2，在图片的右侧，展示有文字识别结果和身份证正反面图片，点击图片左侧的文件上传按钮，可继续识别身份证图片。



图5-2结果页展示

1. 总结

本文采用CNN和CTC组合的技术方案，提供了一种端到端的不定长字符识别方案，此方案可准确识别身份证复印件中的各元素信息，经数据集测试，识别准确率为90%。竞赛提供的数据集为生成数据，与真实样本存在差异，因计算机性能有限，复印件图片中去除水印未达到最好效果，若采用神经网络模型训练去除水印，可进一步提高准确率；此方案针对竞赛数据设计实现，故依赖许多强假设，当应用场景发生变化时，需要对方案进行相应的调整。

致 谢

在四年的大学时光中，我既收获了丰富的专业知识，也得到了同学们和老师们的热心帮助，在此，我向他们表示衷心的感谢。

首先，真诚地感谢我的论文指导老师李艳玲教授和王维庆老师，感谢她们在

忙碌中抽出时间为我的毕业设计做细心和耐心的指导，指引我在人工智能方向一步步成长。

最后，我要感谢家人在毕业设计期间对我生活上的帮助和精神上的鼓励，使我以积极的心态面对作品中的种种困难。

参考文献

1. Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.
2. 本刊讯.Facebook发布深度学习框架PyTorch 1.3[J].数据分析与知识发现,2019,3(10):65.
3. 王宇轩.数字图像处理的锐化算法[J].通讯世界,2019,26(02):275-276.
4. 陈浩.图像经典边缘检测算子的研究与比较[J].电脑编程技巧与维护,2019(12):150-152.
5. 沈臻,韩震宇.基于机器视觉的OCR自动识别系统的研发[J].科技与创新,2019(08):144-145+147.
6. 戈嘉宇,刘为嵩.基于深度学习的身份证识别系统的设计与实现[J].电子世界,2020(02):109.
7. 夏昌新,莫浩泓,王成鑫等.基于深度学习的图像文字识别技术研究与应用[J/OL].软件导刊:1-5[2020-04-12].
8. 王雪冰,姜道义,张海洋.基于卷积神经网络的文字识别优化方法研究[J].中国石油大学胜利学院学报,2019,33(04):39-42.
9. 田瑶琳.基于RGB分割的含数字水印图像的文字识别[J].福建电脑,2019,35(04):62-64.
10. 杨晨红. 基于机器学习的光学字符识别与应用研究[D].西安电子科技大学,2018.
11. 李宇霞,孙永奇,闫茹等.基于CNN图像识别与语义的可靠性优先路径搜索算法[J/OL].计算机工程:1-12[2020-04-17].https://doi.org/10.19678/j.issn.1000-3428.0056844.

**OCR-based ID Element Extraction Design and Implementation**

College of Computer Science and Technology 2016 Cloud Data Centre Operation and

Maintenance Gao Shang 20161104492

Directed by Li Yanling Associate Professor

**Abstract** In this paper, we design a set of information extraction system for ID card elements in OCR system, which uses deep convolutional neural network technology for text recognition, and the system is divided into two parts, the front-end system is responsible for uploading pictures and displaying recognition results, and the back-end system is responsible for image processing and text recognition. The extraction system has a low coupling feature, each sub-module can be run independently, and the extraction system has been tested by the data set, with an identification accuracy of 90%.

**Keywords** OCR; deep convolutional neural network; text recognition

1. [] 赛题链接：https://www.datafountain.cn/competitions/346 [↑](#footnote-ref-1)