基于 OCR 的身份证要素提取

系统之神与我同在

易显维
架构与解决方案处/高级算法专家
建信金科武汉事业群
中国-武汉
necther@qq.com

李欢
IT 开发部/项目经理
大连海科信息技术有限公司
中国-大连
ehuman@hotmail.com

李虎
架构与解决方案处/算法工程师
建信金科武汉事业群
中国-武汉
623859912@qq.com

李安
算法部门/算法工程师
UGCEA
中国-广州
lianEnAndy@163.com

团队简介

团队成员 4 名，分别来自武汉、大连、广州。因为热爱探索新技术，通过比赛联系在一起，热爱工作、热爱生活。

易显维：建信金科有限责任公司武汉事业群高级算法专家，擅长数学建模及计算机视觉相关技术。主要负责给武汉事业群各项目组提供各类算法解决方案。近两年数次获得算法比赛优异成绩。

李欢：大连海科信息技术有限公司，IT 项目经理，擅长项目管理、问题分析、近两年数次获得比赛前三名优异成绩。

李虎：硕士毕业于厦门大学数据挖掘实验室，在公司负责目标检测和 OCR 项目的模型调优和项目实施工作。

李安：热爱学习，多次取得数据科学竞赛 top3 的成绩。

摘要

身份证影像文件在商业银行中被广泛应用于认证、信息采集等领域，具有极高的商业价值。但是在实际应用中存在以下两大挑战问题：1. 图像质量不佳；2. 印章水印干扰。

针对上述问题，本文结合当下流行的深度学习方法，提出了一整套端到端的解决方案。该方案有效解决了目标检测，图像倾斜校正，内部要素定位，要素识别，结果校正这五大问题。1. 在目标检测中，采用 yolo-V3 算法，并对位置靠近的正反面样本图片进行了特殊处理；2. 在图像倾斜校正中，提出了一种“两步法”的图像倾斜校正方法，即先正倒粗调再精细微调的方法；3. 在内部要素定位中，提出了一种先 U-Net 处理再投影法处理的方法，同时构建了样本生成器，生成 U-Net 需要的有无印章的训练样本，有效提高了定位准确度；4. 在要素识别中，提出在部分栏位中用“分类法”代替识别法对要素进行识别，提高了准确率；5. 在结果校正中，针对身份证的特点，对不同要素分别设计了校正算法，实验证明，校正有效提高了识别准确率。

关键词

“两步法”，样本生成器，U-Net，“分类法”，栏位校正，

1 方案概述

本方案主要包含如下八个部分：

1. 正反面定位：使用目标检测的 yolo-V3 算法在原图像中找到身份证的正反面位置。2. 正倒分类：分别使用分类网络，对截取出的身份证正反面图片，判断其图像的正倒结果。3. 图像倾斜校正：使用 Canny 边缘检测和霍夫变换算法对正倒分类的结果进行进一步方向校正。4. 身份证要素定位：采用投影法对身份证中的要素文本进行定位。5. 多行文本行二次定位：针对住址和发证机关的多行文本行的情况，采用 U-Net 和投影法进行二次切分。6. 要素识别：分别对每个要素训练 CRNN 网络或者分类网络用于识别。7. 结果校正：包括住址的校正、发证机关的校正以及出生日期和身份证号码要素的互相校正。8. 展望：本文之后还可以再改进的地方。

2 正反面定位

由于目标并不复杂，所以本方案采用 yolo-V3[1]这种速度较快的方法。在主干网络的选用上，为了追求更快的速度，尝试过使用 Mobile-Net[2]用于定位的主干网络，但发现效果较差。所以本方案中使用 VGG16[3]为主干网络的 Yolo-V3 作为目标检测网络。另外由于样本中存在正反面图像距离很近的情况，所以单独挑选出了该种情况进行样本标注，提升了定位准确率。

3 正倒分类

由于正反面的正倒方向特征差别很大，所以对网络深度要求不高，本方案中使用 Inception-V3 进行分类。该网络具有计算量小的特点，完全满足使用。

获得高质量模型最保险的做法就是增加模型的深度（层数）或者是其宽度（层核或者神经元数），但是这里可能会出现如下的缺陷：1. 参数太多，若训练数据集有限，容易过拟合；2. 网络越大计算复杂度越大，难以应用；3. 网络越深，梯度越往后穿越容易消失，难以优化模型。

4 图像倾斜校正

4.1 算法步骤

上一章节对身份证的正反面进行了正倒的判断粗调，但是仍然需要对角度进行微调，即本文摘要中提出的“两步法”。微调包括以下两个步骤：

1. Canny 边缘检测[4]，得到二值化的边缘分布图片；

2. 霍夫变换[5]检测直线，根据直线计算校正角度；

4.2 倾斜校正结果展示

图 1 校正结果图

5 身份证要素定位

本文首先尝试使用了 ctpn 方法，思路是识别出所有的文字框，然后用相对定位的方法，识别效果如图 2 所示。这里的相对定位又有两个思路，一是用文本框纵坐标的排序定位，但存在类似第二张图片的情况，多余的文字也会被识别出来，多或少识别出文本框都会使方法失效；二是利用左上角的“仅限 BDCI 比赛使用”作为标志位置，人为设置与该标志位的距离来确定要素文本框的位置，该方法要求文本框左上角顶点的坐标不变，但是比较下面两张图片，左上角顶点坐标漂移很大。综上，ctpn 结合相对位置失效。

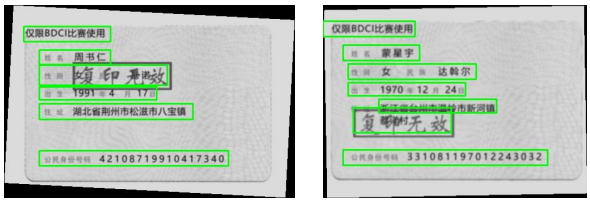


图 2 ctpn 定位效果图

继续上述思路，如果我们能将身份证左上角的位置固定住，那么相对位置就一定是准确的（身份证图片没有折叠的情况）。所以我们选用了投影法，由于在下一章节的二次定位中也使用了投影法，将在下一章节中进行详细介绍。

投影法的识别效果图如图 3 所示，可以看出每个要素的位置都定位的很好，不会受到干扰。

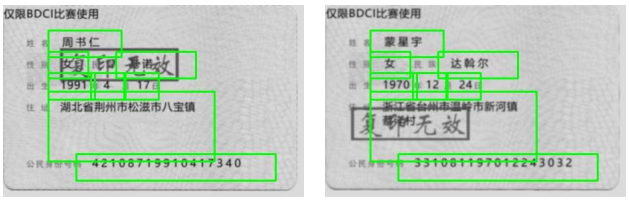


图 3 投影法定位效果图

各个要素的定位相对坐标如下表 1 所示：

要素名称	y_min	Height	x_min	Width
姓名	38	40	68	110
性别	70	40	68	60
民族	70	40	68	60
年	102	40	68	70
月	102	40	132	52
日	102	40	182	52
住址	130	105	68	250
发证机关	195	45	165	240
有效期限	230	35	165	240

表 1 各要素的定位相对坐标表

6 多行文本行二次定位

由于需要用到投影法，而印章大部分是在身份证的内容区域，即要识别的区域，印章会对定位结果造成很大的干扰。所以本章节的处理包括两个步骤，去印章和定位。

6.1 去印章

去印章采用 U-Net[6]算法。除了本节需要用到去印章，在后面的校正章节也需要用到去印章。

训练需要准备有印章和无印章的图片，因此本文提出了一种“样本生成器”方法，有印章的样本的制作步骤如下：

1. 统计各个要素的语料文字分布，用于生成样本中的文字内容；2. 随机从语料中抽取内容，生成图片样本的图片和标记内容；3. 在上一步生成的图片上，通过合成技术，将印章绘制在样本上；

上述步骤的 1，2 即是无印章的构造方法。

有印章和无印章的图片样本如图 4 所示：

图 4 印章身份证样本图片，左为有，右为无

训练时，从模型的输入端输入有印章的样本，将无印章的样本作为目标既可训练 U-Net 模型。

6.2 二次定位

采用投影法进行二次定位，定位后将多行本文行切分成单行文本。具体处理步骤如图 5 所示：

图 5 处理步骤图

原始图如图 5-1 所示，通过腐蚀变成色块，腐蚀后的图片如图 5-2 所示，通过均值化加二值化处理，处理后的图片如图 5-3 所示。

然后可以计算出投影后的文字区域，再利用连通域将临近的区域联通起来，最后得出最终的文字区域。

7 要素识别

即在某些栏位的识别中，本文提出用“分类法”代替识别法。在身份证识别问题中，性别、民族和出生日期等栏位的输出结果是有固定类别的，例如性别只有“男”和“女”两种，民族只有 56 种。采用分类法代替识别法可以有效提高准确率。其余要素的识别使用 CRNN[7]识别法，各要素使用的识别方法如表 2 所示：

要素名称	识别方式	备注
姓名	CRNN	
性别	分类	男、女
民族	分类	56 个民族
年	分类	1958 至 2009
月	分类	1 至 12
日	分类	1 至 31
住址	CRNN	
发证机关	CRNN	
有效期限	CRNN	

表 2 各要素使用的识别方法表

8 结果校正

8.1 住址校正

算法亮点：1. 利用国家统计局“全国行政区数据集”；2. 增加 jieba 专用词语；3. 采用 gensim 库的 TF-IDF 模型进行校正；4. 剔除无用字；5. 采用编辑距离；6. 合并不同年份的地址变化数据。对于地址中严重“破损”数据可以修复。

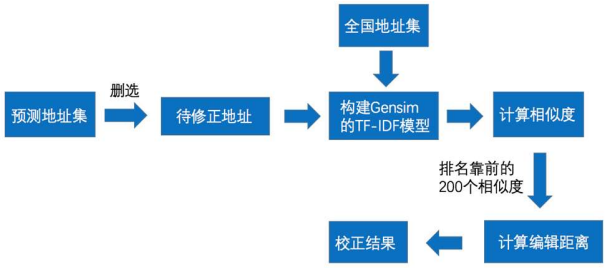


图 6 地址要素校正算法框图

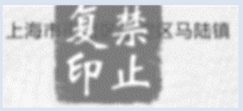
要识别的图片	识别文字	校正后文字
	上海市街园区 新理区马陆镇	上海市市辖区 嘉定区马陆镇

图 7 地址要素的校正结果

8.2 发证机关校正

算法亮点：1. 采用已校正的住址；2. 采用编辑距离计算；对于无法利用自身数据校验的数据，再次利用地址和发证机关的数据关系进行校正。

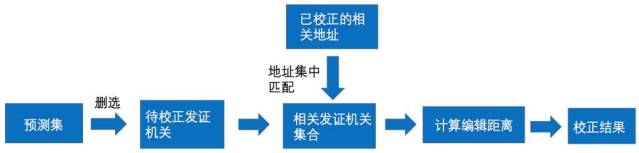


图 8 发证机关要素校正算法框图

要识别的图片	识别文字	校正后文字
	昌吉回族自治港 徽昌鲁安局	昌吉回族自治州 奇台县公安局

图 9 发证机关要素的校正结果

8.3 出生日期和 ID 栏位的互相校正

算法亮点：1. 身份证号码中的第 7-14 位与出生日期的年月日栏位是对应一致的，可以根据这个进行相互校正，取两者中置信度高的作为结果。2. 用 U-net 算法判断置信度：即用 U-net 算法判断栏位的受污染程度，污染程度小的认为预测的置信度高。

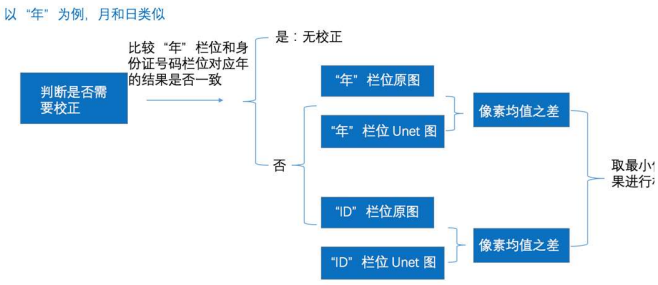


图 10 出生日期和 ID 要素相互校正算法框图

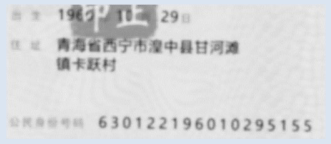
要识别的图片	识别年	识别月	校正年	校正月
	1962	11	1960	10

图 11 出生日期和 ID 要素的校正结果

9 展望

本文提出的一整套解决方案，在一定程度上能改善 OCR 识别中图像质量差和印章干扰等问题，但是仍然存在一些不足和可以进一步探究的地方。

1. 模型压缩：在不大量牺牲准确率的基础上，设计更加快速的网络结构，这类方法主要有模型裁剪，稀疏化等。2. 提炼出 OCR 框架：我们试图将已有工作提炼出产品化的框架，并加入模板匹配算法，能快速响应业务需求。3. 图像还原：我们试图在本文已有 U-Net 还原算法的基础上，设计出还原效果更好的图像还原算法。

致谢

4 个月，转瞬即逝。

首先感谢主办方提供的平台和学习机会，以及在将近 4 个月的时间里的辛勤筹备和组织。然后要感谢我的队友们，不管遇到什么困难和难题，大家都想办法去解决，去克服，感谢大家的付出，感谢大家没有放弃。

最后，我们今后还会继续支持 CCF 大赛，继续参加 CC 主办的其他人工智能比赛，同时希望 CCF 大赛越办越好。

参考

[1] Redmon, Joseph, Farhadi, Ali. YOLOv3: An Incremental Improvement[J].

[2] Howard A G , Zhu M , Chen B , et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.

[3] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. CoRR, abs/1412.0623, 2014.

[4] Canny J . A Computational Approach To Edge Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6):679-698.

[5] Ballard D H . Generalizing the Hough transform to detect arbitrary shapes[J]. Pattern Recognition, 1981, 13(2):111-122.

[6] Ronneberger O , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015.

[7] Shi B , Bai X , Yao C . An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(11):2298-2304.