



OCR

OCR

概述

预处理

识别

后处理

发展前景

小组成员：倪紫薇 冯硕 刘聪聪 李昭 刘玉洁

1

什么是OCR？



OPTICAL CHARACTER RECOGNITION

光学字符识别

用电脑的外部设备——光电扫描仪/照相机，先将文本扫描成图像

ocr系统对图像进行文字的识别（分析和处理），将含有数字、英文字母、汉字的图像转成文本形式

最后可用“文件发送”或“导出”输出到文档编辑软件中

1

OCR的广泛应用



印刷体识别

手写识别

基于特定场景识别

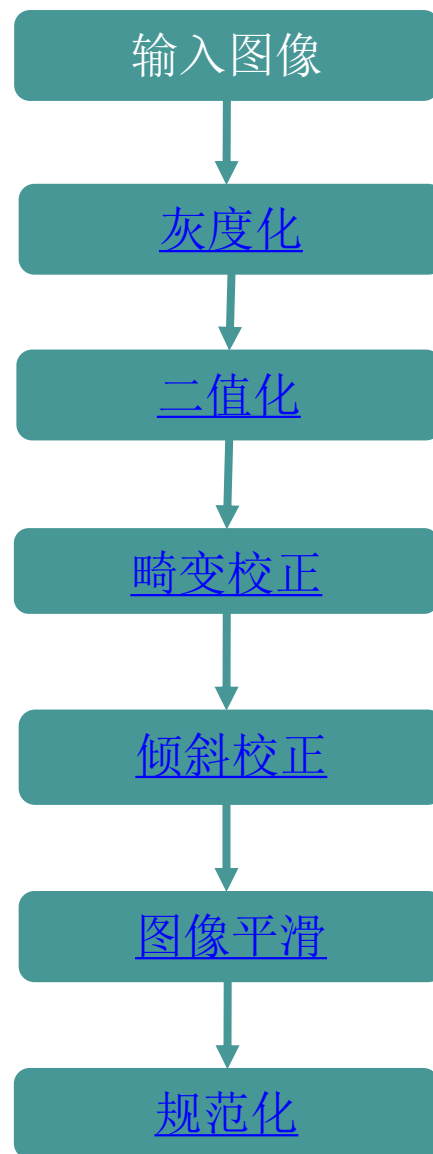
基于通用场景识别



2

图像预处理

对包含文字的数字图像进行处理以便后续进行特征提取、学习。
这个过程的主要目的是减少图像中的无用信息，来提高文字的识别率



3

OCR识别类型

传统**OCR** – 基于图像处理+统计机器学习

深度学习**OCR** – 基于CNN的神经网络

3

基于字符特征的数字字符识别

场景：车牌号、名片、身份证号、
银行卡号、快递号、
支票号、邮编.....

BP或SVM算法实现复杂，计算量大，大材小用。

只需能将 10个字符分别区分开的特征组合。



识别流程图

3

基于字符特征的数字字符识别

图像投影-判定图像中字符位置

投影->射线积分

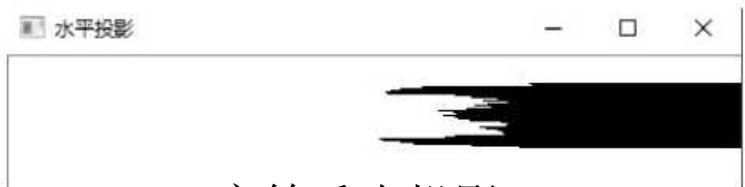
垂直投影：旋转角 $-\pi/2$

水平投影：旋转角 π

通过颜色变化判断字符位置



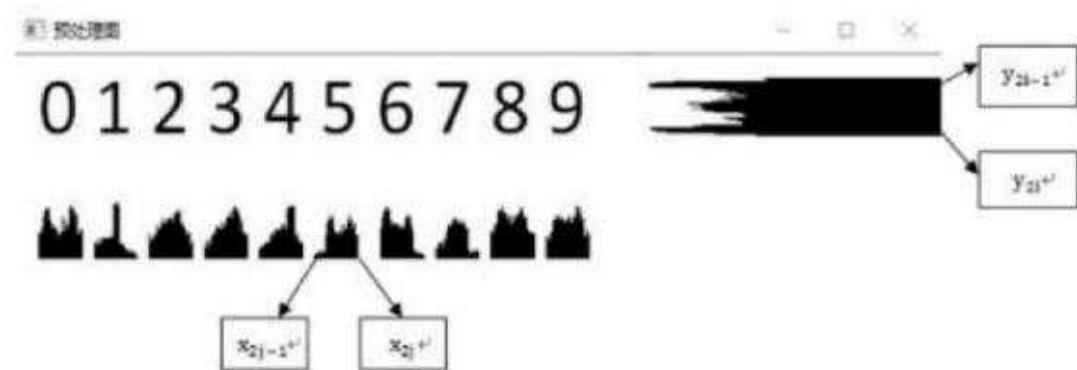
预处理效果图



字符垂直投影



字符水平投影



3

基于字符特征的数字字符识别

“区分向量”识别-
统计特定向量与字符相交次数

第 i 行、第 j 个字符的识别需要
三个区分向量

$$\gamma_{1(i,j)} = \frac{y_{2i} + 3y_{2i-1}}{4}$$

$$\gamma_{2(i,j)} = \frac{3y_{2i} + y_{2i-1}}{4}$$

$$\gamma_{3(i,j)} = \frac{x_{2i} + x_{2i-1}}{2}$$

区分向量与字符相交次数 分类				左右侧		辨别结果
2	2	2	0			0
1	1	1	1、7	0	0	1
				1	1	7
1	1	3	2、3、5	1	0	2
				1	1	3
				0	1	5
2	2	3	8			8
1	2	3	9			9

3

基于字符特征的数字字符识别

“区分向量”识别-
相交位置位于竖直区分向量 y_3 左右侧

相交位置位于 y_3 左侧，记为0

相交位置位于 y_3 右侧，记为1

区分向量与字符相交次数				分类	左右侧		辨别结果
2	2	2		0			0
1	1	1		1、7	0	0	1
					1	1	7
1	1	3		2、3、5	1	0	2
					1	1	3
					0	1	5
2	2	3		8			8
1	2	3		9			9

3

欧式距离-KNN临近模型

欧式距离：计算待求字符与每种字符每一项特征平均值的欧式距离。

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\dots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^n(x_i-y_i)^2}$$

KNN思想： 如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别。

3

欧式距离-KNN临近模型

模型求解-识别字符'A':

1. 准备数据，预处理。
2. 在 'A'到'Z'中随机选400组数据作为优先级队列，求和待求字符的欧式距离。
3. 求训练集的20000组数据与待求数据的欧式距离，加入队列或舍弃。

结论：所求字符相似度最大的是'A'。
算法精度高。

优先级数列字符统计图

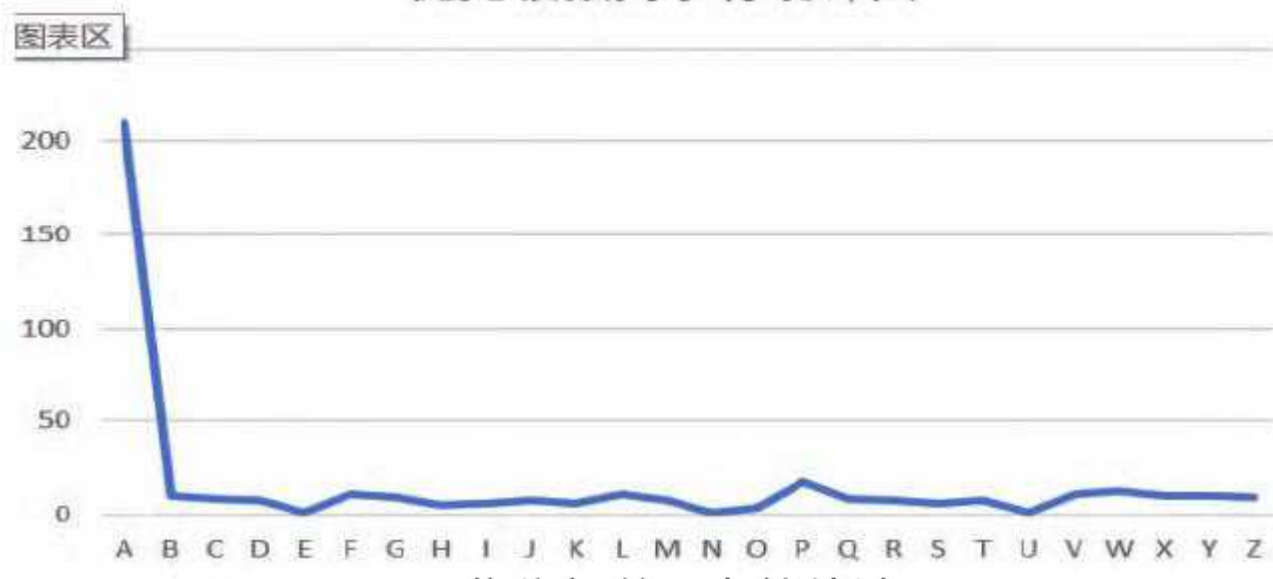


表 3 欧氏距离—KNN 算法验证表

平均执行时间（毫秒）	精确度（%）	误判率（%）
180.24	93.121	6.879

3

传统的OCR的不足

- 要求版面结构有较高的规则性且前背景可分性强。
- 通过人工设计特征来训练字符识别模型，在字体变化、模糊或背景干扰时，此类单一特征的泛化能力迅速下降。
- 人们需要花费大量时间做特征的设计。

.....

3

卷积神经网络(CNN)

CNN是传统神经网络的一个改进。

特点：把特征提取完全交给机器，自动完成。

3

卷积神经网络层次结构

数据输入层

原始图像预处理

卷积计算层

对一个块的像素进行内积运算。

输出提取的特征。

1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved
Feature

3

卷积神经网络层次结构

激励层

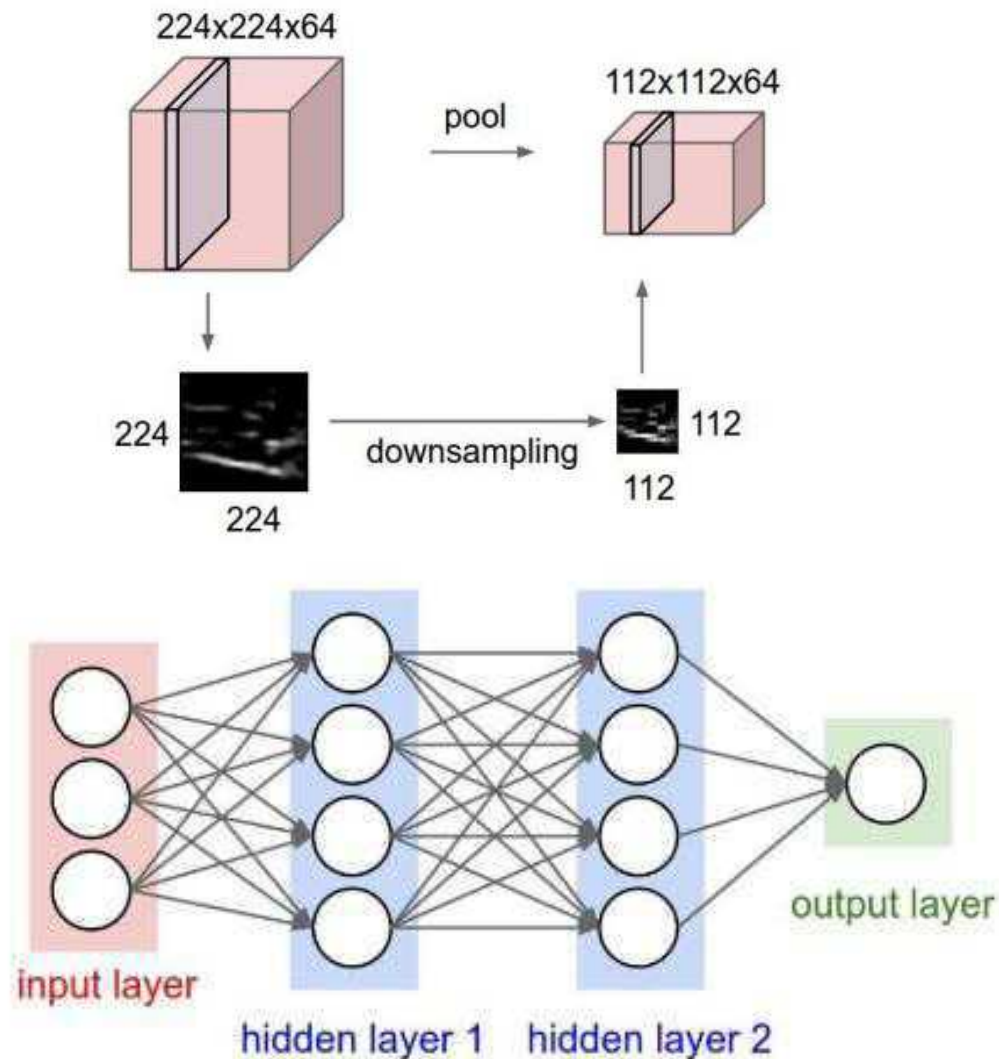
把卷积层输出结果做非线性映射。

池化层

压缩图像。

全连接层

两层之间所有神经元都有权重连接。



3

卷积神经网络的常用框架

Caffe

- 源于Berkeley的主流CV工具包，支持C++,python,matlab
- Model Zoo中有大量预训练好的模型供使用

Torch

- Facebook用的卷积神经网络工具包
- 通过时域卷积的本地接口，使用非常直观
- 定义新网络层简单

TensorFlow

- Google的深度学习框架
- TensorBoard可视化很方便
- 数据和模型并行化好，速度快

3

卷积神经网络优势与劣势

优势:

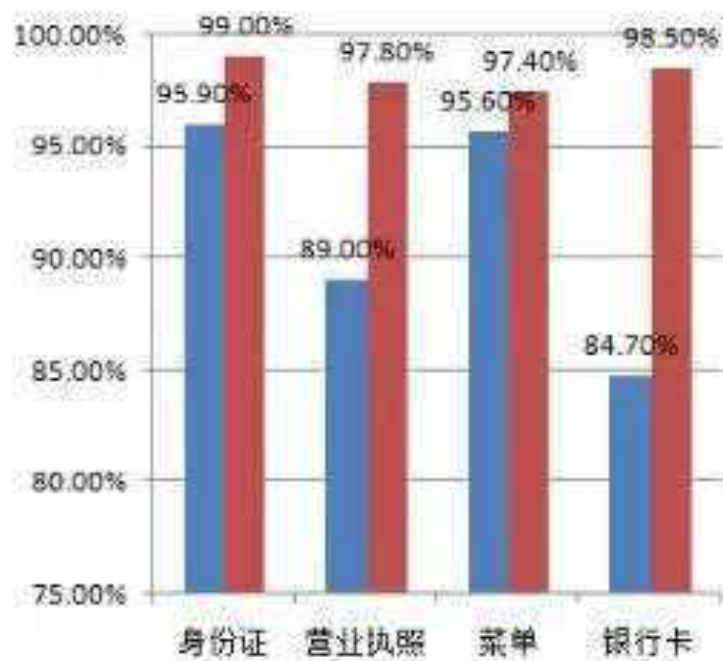
- 训练相对容易，无需复杂的特征提取过程。
- 卷积层通过共享权重参数，相比传统的神经网络，减少参数个数，对内存要求降低
- 对图像的扭曲、变形、像素迁移保持稳定、具有一定不变性特征。

劣势:

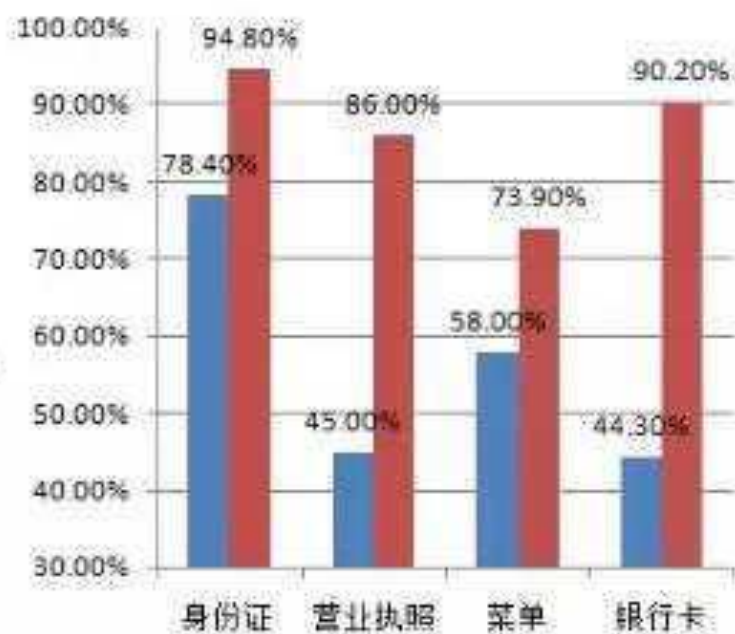
- 训练需要大量数据、时间和硬件资源。

3

传统OCR和深度学习OCR识别率对比



■ 字符识别率 (传统OCR)
■ 字符识别率 (深度学习OCR)



■ 条目准确率 (传统OCR)
■ 条目准确率 (深度学习OCR)

4

OCR后处理

为什么要进行后处理？

后处理：采用某种方法(包括人工的或者计算机的)在OCR 的单字识别结果(一般为首选字和候选字, 候选字一般会给出该字作为正确字输出的概率)中挑选出最符合的汉字进行输出,包括**纠正错误的汉字,给出拒识汉字,确定模糊的汉字**。

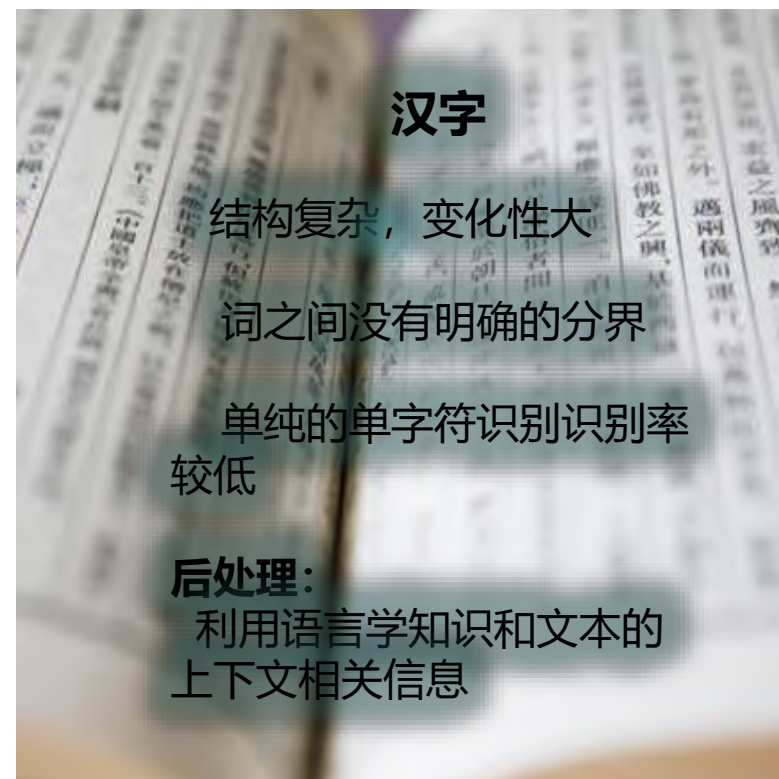
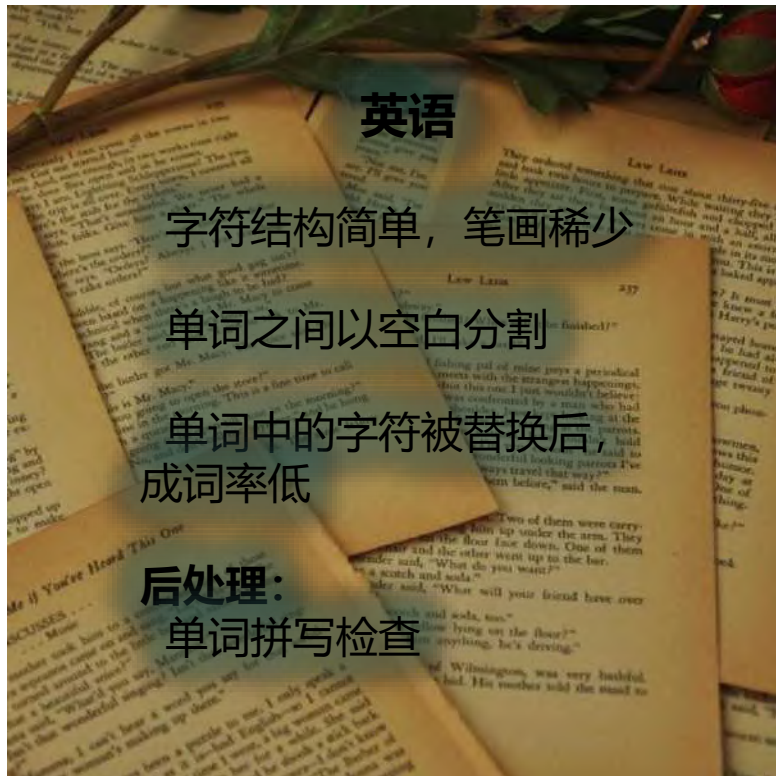
主要靠人工完成
工作量大，长时间工作
易出错



计算机自动识别
提高系统的整体性能
和字符的识别率

4

OCR后处理



4

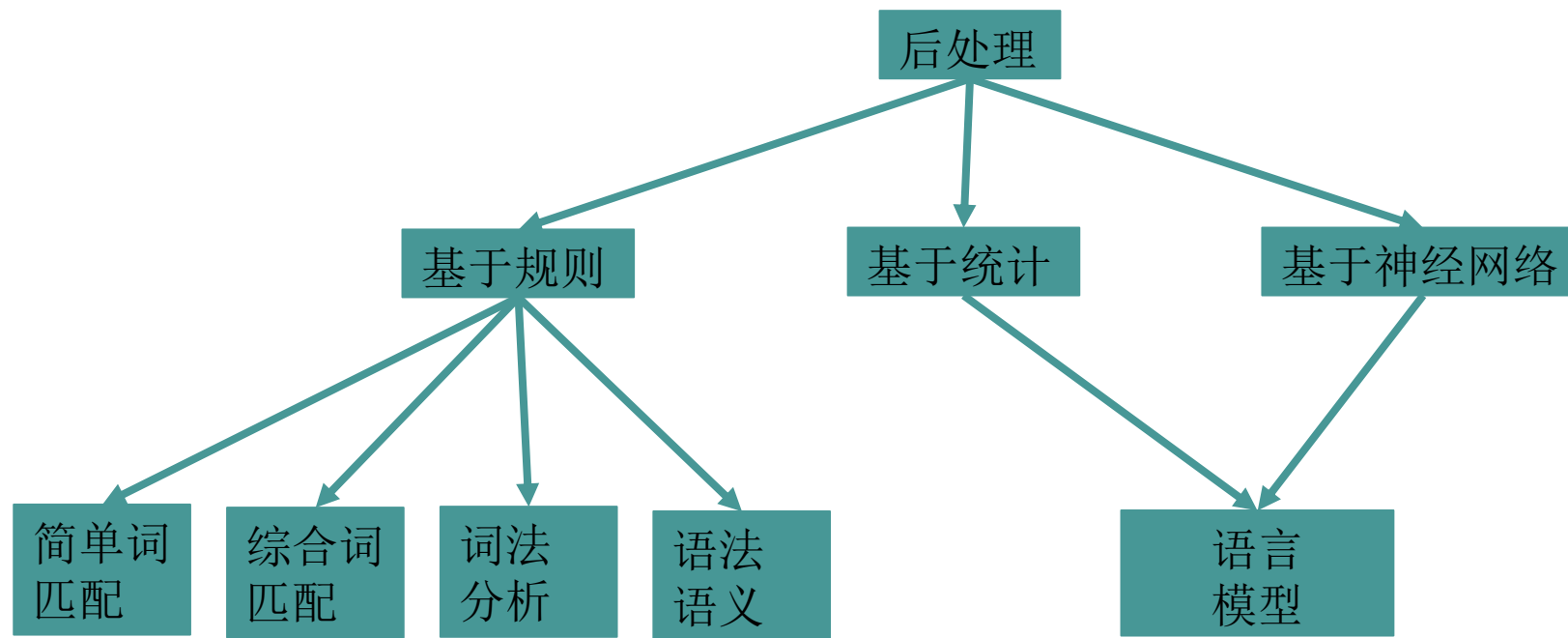
OCR后处理

方 法

基于规则的后处理

基于统计的后处理

基于神经网络的后处理

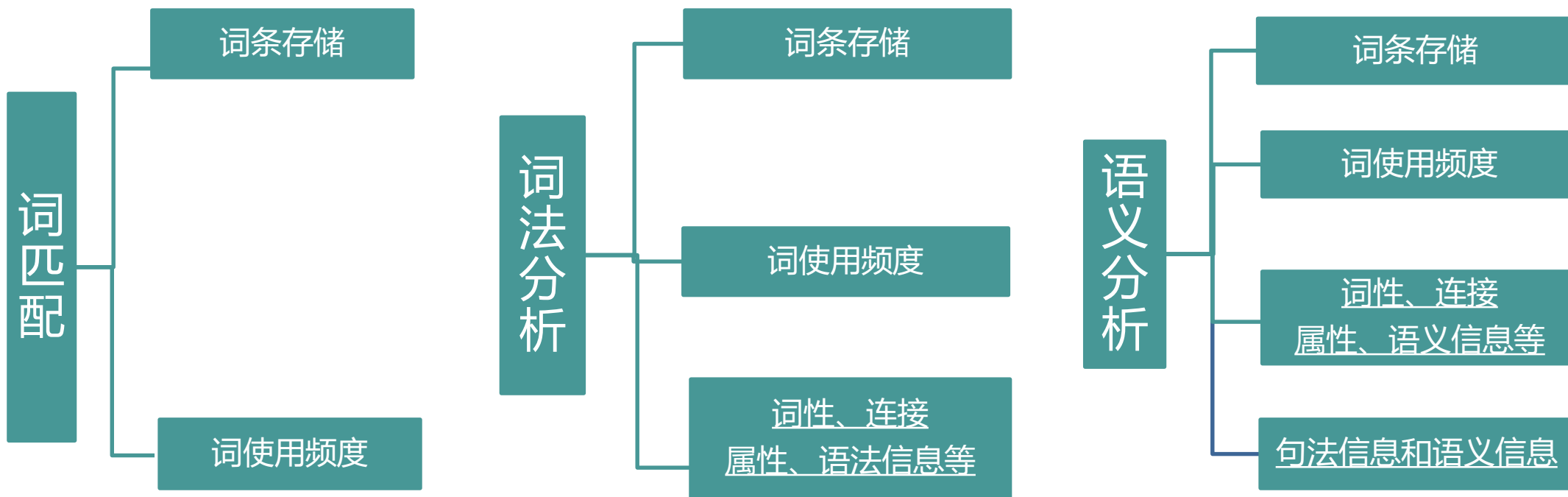


4

OCR后处理

一、基于规则的后处理

关键：建立汉语词条数据库



4

OCR后处理

二、基于统计的后处理

N元文法 (N-gram) 模型:

假定在一个语句中,第*i*个词出现的概率条件的依赖于它前面的*N*-1个词,即将一个词的上下文定义为该词前面出现的*N*-1个词,这样的语言模型就叫做*N*元文法模型(*N*-gram)。

词 ω 出现的概率: $P(\omega_i = \omega | c) = P(\omega_i = \omega | \omega_{i-N+1}^{i-1})$

N=2时, 为Bigram语言模型:

$$P(\omega_i = \omega | c) = P(\omega_i = \omega | \omega_{i-1})$$

N=3时, 为Trigram语言模型:

$$P(\omega_i = \omega | c) = P(\omega_i = \omega | \omega_{i-2}^{i-1})$$

中国科学。
院?



OCR后处理

[illegible]

关键：文本语料库统计

4

OCR后处理

三、基于神经网络的后处理

N-gram模型:

在离散空间下表示语言的分布

缺乏对词的相似性的有效表示, 存在严重的**数据稀疏**问题

神经网络语言模型:

将词映射到连续的空间内, 有效地解决了数据稀疏问题

计算复杂度远高于N-gram模型

人工智能

$P(\text{能}|\text{智}) > P(\text{智}|\text{工})$

$P(\text{智}|\text{工}) = 0?$

4

OCR后处理

三、基于神经网络的后处理

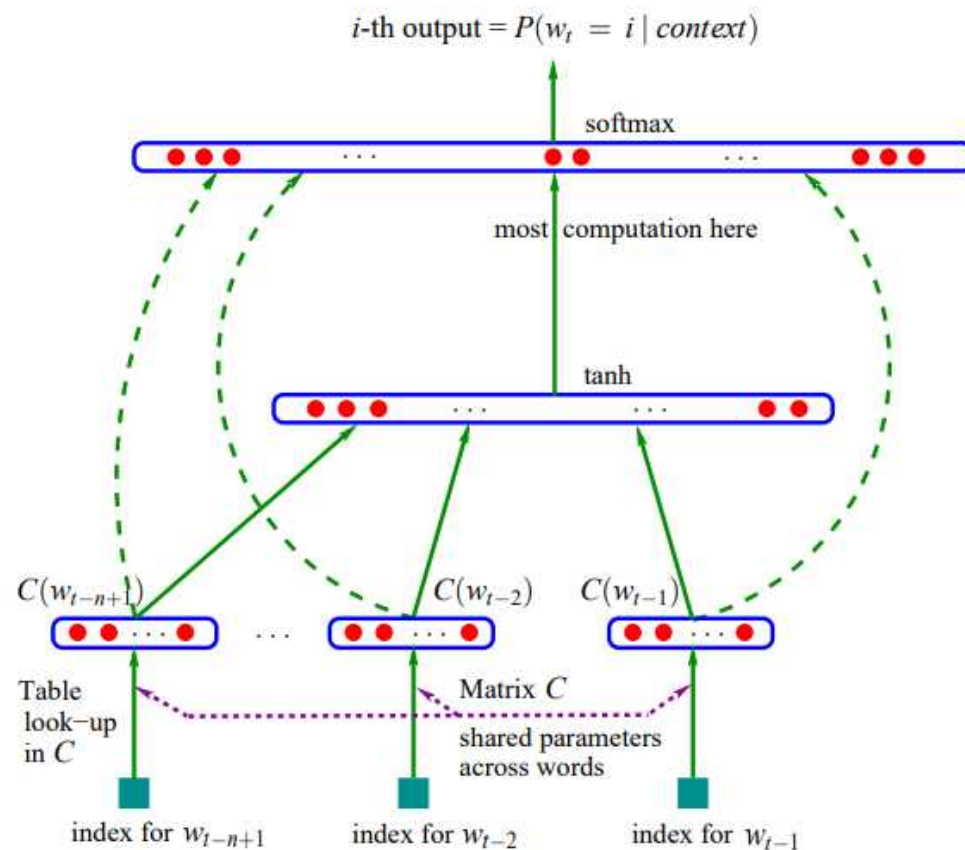
神经网络语言模型 (NNLM)

输入层：长度为 $n-1$ 的文本序列

映射层：输入的每个词映射为特征向量

隐层：通过函数将向量转化为一个概率分布

输出层：保证所有结果和为1



存在的问题

虽然OCR技术取得了很大进展，但仍有许多研究空间。

- 场景文字识别面临诸多挑战

- 场景文字的多样性
- 场景文本背景复杂
- 其它干扰因素

- 实时性问题

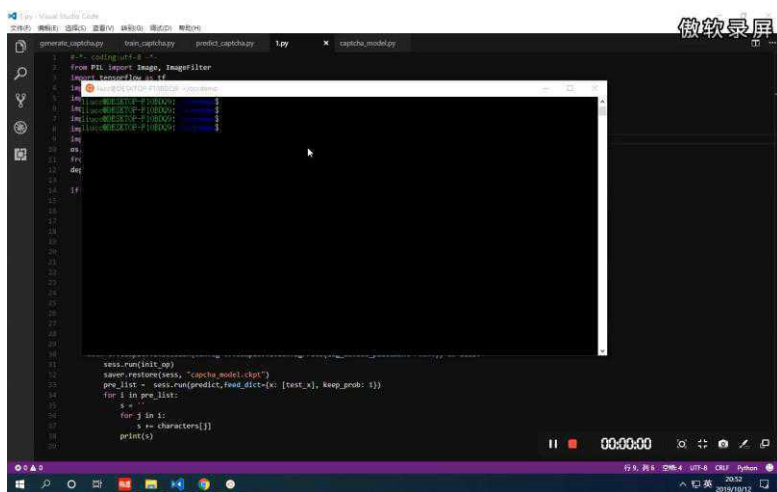
- 文本检测和处理还不能达到完全地实时，目前许多OCR应用仅适用于具有离线处理模式的捕获图像。

- **更高准确率的通用场景文字识别**
随着计算机视觉处理和自然语言处理领域的发展，场景文本识别的准确率还有进一步提高的空间。
- **更加通用的OCR应用范围**
随着复杂场景下文本识别准确率的提高，可以使得OCR应用在更多领域。
- **实时性能**
OCR识别性能的提升使得文本识别更加迅速

7

演示

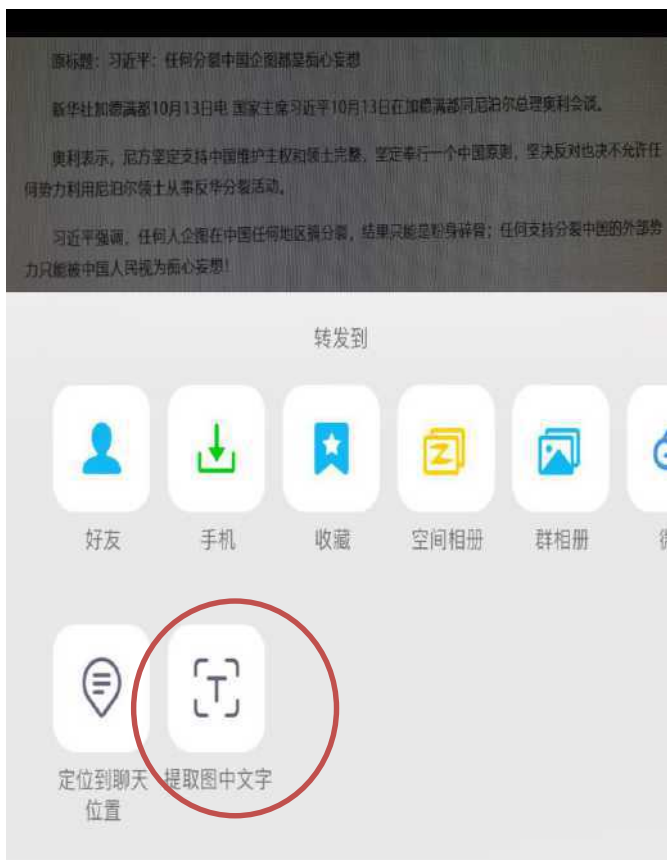
- 验证码识别



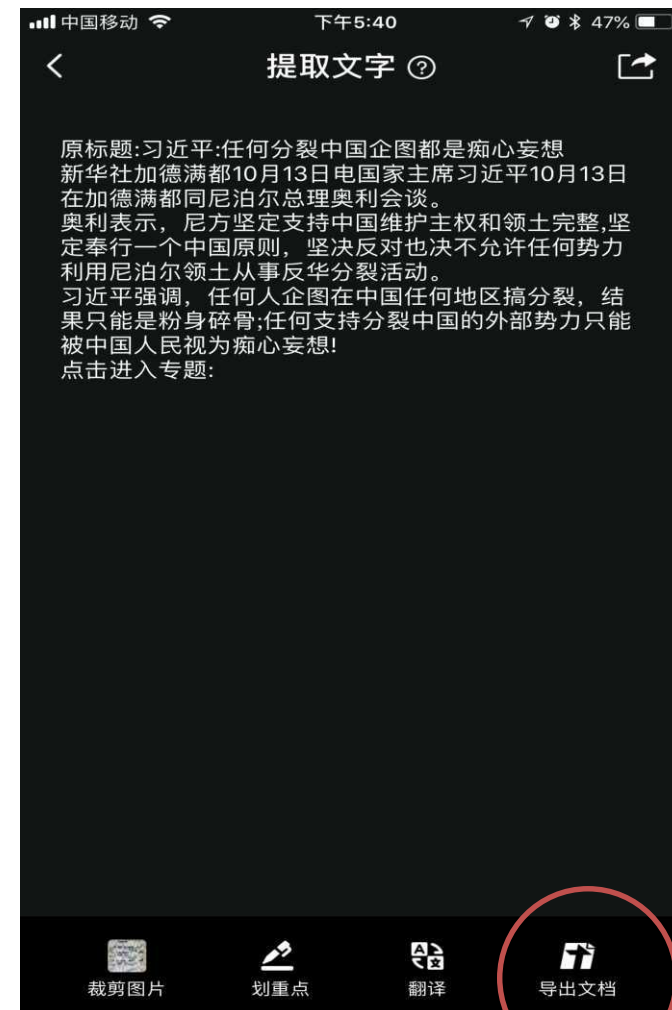
谢谢观看！

1

印刷体文字识别



取消



1

手写文字识别



各类手写汉字或手写数字拍摄照片或扫描件

识别结果

序号	内容
1	2019年10月7日
	宽度: 462 高度: 94
	左间距: 129 上间距: 227
2	大数据分析与应用
	宽度: 669 高度: 97
	左间距: 139 上间距: 375
3	petmcountexceeded; starting again.
	宽度: 1292 高度: 171
	左间距: 146 上间距: 475

请输入网络图片URL

检测 或 本地上传

图片文件类型支持PNG、JPG、JPEG、BMP，图片大小不超过2M。

1

票据文字识别



3400161130 安徽增值税专用发票 No 00666375 3400161130 00666375 开票日期: 2016年11月11日

名称: 上海优扬新媒信息技术有限公司
纳税人识别号: 91310114585239729M
地址、电话: 上海市嘉定区工业区汇源路55号H幢4层 021-83460220
开户行及账号: 中国工商银行上海市嘉定支行 1001700619300415149

货物或应税劳务、服务名称	规格型号	单位	数量	单价	金额	税率	税额
信息费					2524.75	6%	151.49
合计					¥2524.75		¥151.49
价税合计(大写)	贰仟陆佰柒拾陆圆贰角玖分			(小写)	¥2676.24		
销货方	合肥兴嘉动地信息技术有限公司 纳税人识别号: 91340100686877076E 地址、电话: 合肥市金寨路71号科技大厦5层 0551-65411799 开户行及账号: 招商银行合肥南七支行581903160110132			备注			
收款人: 李平	复核: 李平	开票人: 秦丽萍					

识别结果

| 发票基本信息

发票类型 专用发票
发票名称 安徽增值税专用发票
发票代码 3400161130
发票号码 00666375
开票日期 2016年11月11日
收款人 李平
复核 李平
开票人 秦丽萍

| 购买方

名称 上海优扬新媒信息技术有限公司
纳税人识别号 91310114585239729M
地址、电话 上海市嘉定区工业区汇源路55

请输入网络图片URL

检测

或

本地上传

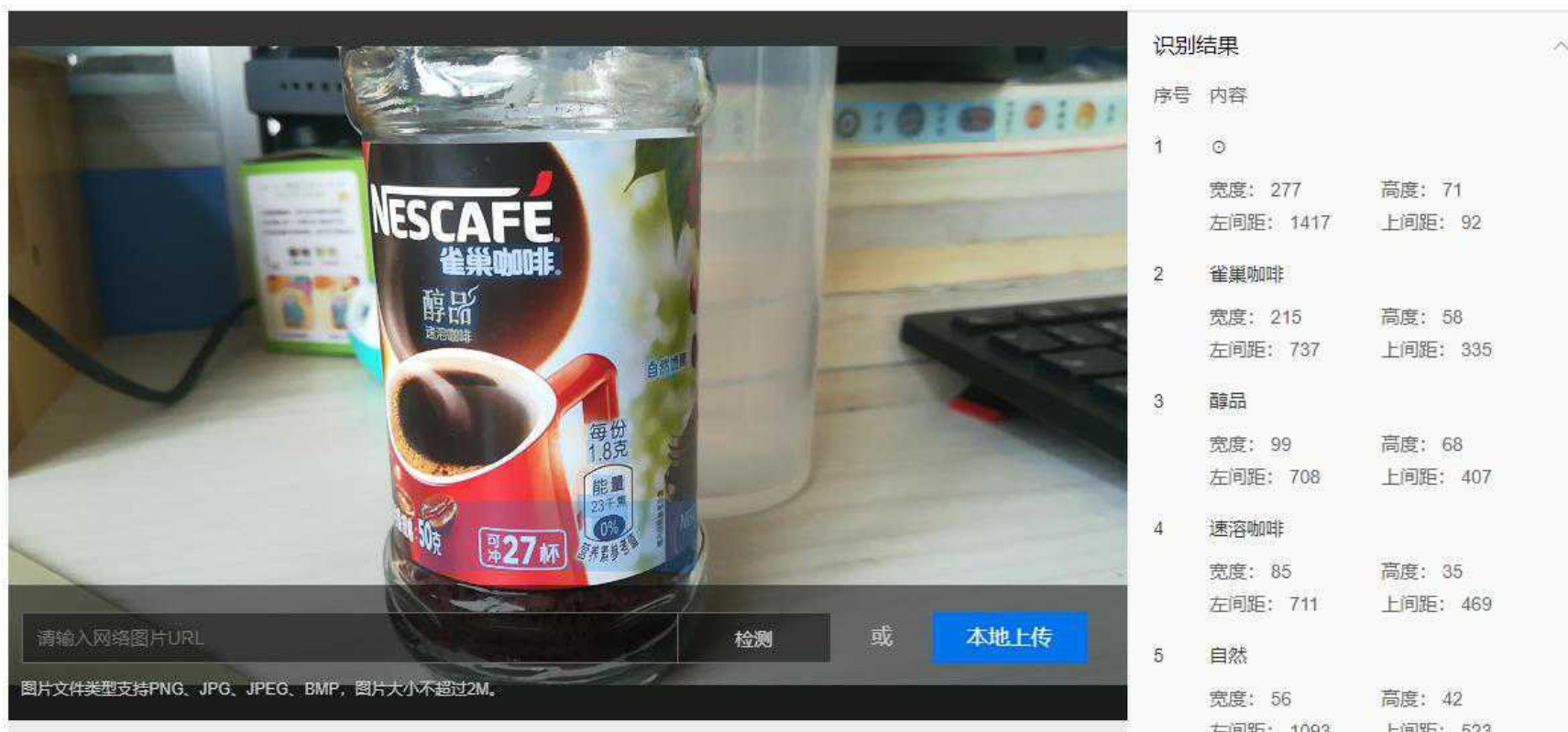
图片文件类型支持PNG、JPG、JPEG、BMP, 图片大小不超过2M。

1

通用场景文字识别

识别任意的自然场景图片中的文字信息

通用文字识别 高精度版 含位置信息版 高精度含位置版



The screenshot displays the Baidu AI Open Platform's OCR interface. On the left, a photograph of a Nescafé jar is shown. The jar's label includes the text "NESCAFÉ", "雀巢咖啡", "醇品", "速溶咖啡", "每份 1.8克", "能量 23千焦", "0%", and "可冲 27 杯". Below the image is a text input field with the placeholder "请输入网络图片URL", a "检测" (Detect) button, and a "本地上传" (Upload from local) button. A note at the bottom states: "图片文件类型支持PNG、JPG、JPEG、BMP, 图片大小不超过2M." On the right, the "识别结果" (Recognition Results) panel is open, showing a table with 5 items. Each item includes a serial number, content, width, height, left offset, and top offset.

序号	内容	宽度	高度	左间距	上间距
1	◎	277	71	1417	92
2	雀巢咖啡	215	58	737	335
3	醇品	99	68	708	407
4	速溶咖啡	85	35	711	469
5	自然	56	42	1093	523

2

真彩色图像→灰度图像



平均值法 $\text{Gray} = (R+G+B) / 3$

加权值法 $\text{Gray} = 0.3R + 0.59G + 0.11B$

公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市慈善总会 2013年5月12日 No 1130034711
Donor: Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects(Currency)	数量 Amount	金额 Total amount
图书	图书		14555
金额合计(小写)			14555
金额合计(大写)			壹万肆仟伍佰伍拾元

接受单位(盖章): 温州市慈善总会 复核人: 开票人: 120
Receiver's Seal Verified by Handling Person

注: 本票据限于2013年12月31日前填开使用方为有效。

第二联 收据



公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市慈善总会 2013年5月12日 No 1130034711
Donor: Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects(Currency)	数量 Amount	金额 Total amount
图书	图书		14555
金额合计(小写)			14555
金额合计(大写)			壹万肆仟伍佰伍拾元

接受单位(盖章): 温州市慈善总会 复核人: 开票人: 120
Receiver's Seal Verified by Handling Person

注: 本票据限于2013年12月31日前填开使用方为有效。

第二联 收据

2

灰度图像→二值图像



公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市慈善总会 2019年5月12日 No. 1130036721

Date: Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects (Currency)	数量 Amount	金额 Total amount
书籍	书籍		1000
金额合计(小写): 1000			
金额合计(大写): 壹仟元整			

温州市慈善总会 财务专用章

接收单位(盖章): 温州市慈善总会 复核人: 开票人: 12

Receiver's Seal Verified by Handling Person

感谢您对公益事业的捐赠! Thank you for support of public welfare!



公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市慈善总会 2019年5月12日 No. 1130036721

Date: Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects (Currency)	数量 Amount	金额 Total amount
书籍	书籍		1000
金额合计(小写): 1000			
金额合计(大写): 壹仟元整			

温州市慈善总会 财务专用章

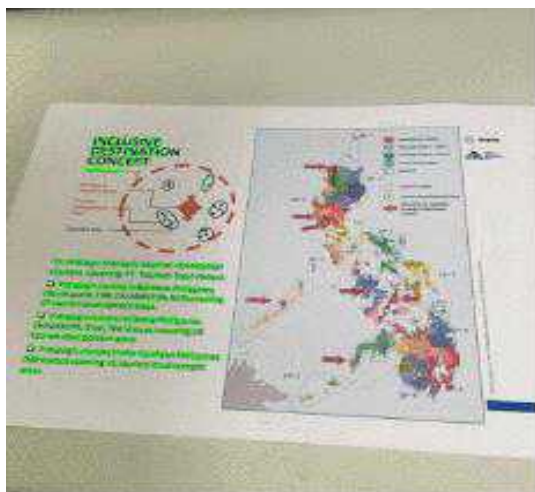
接收单位(盖章): 温州市慈善总会 复核人: 开票人: 12

Receiver's Seal Verified by Handling Person

感谢您对公益事业的捐赠! Thank you for support of public welfare!

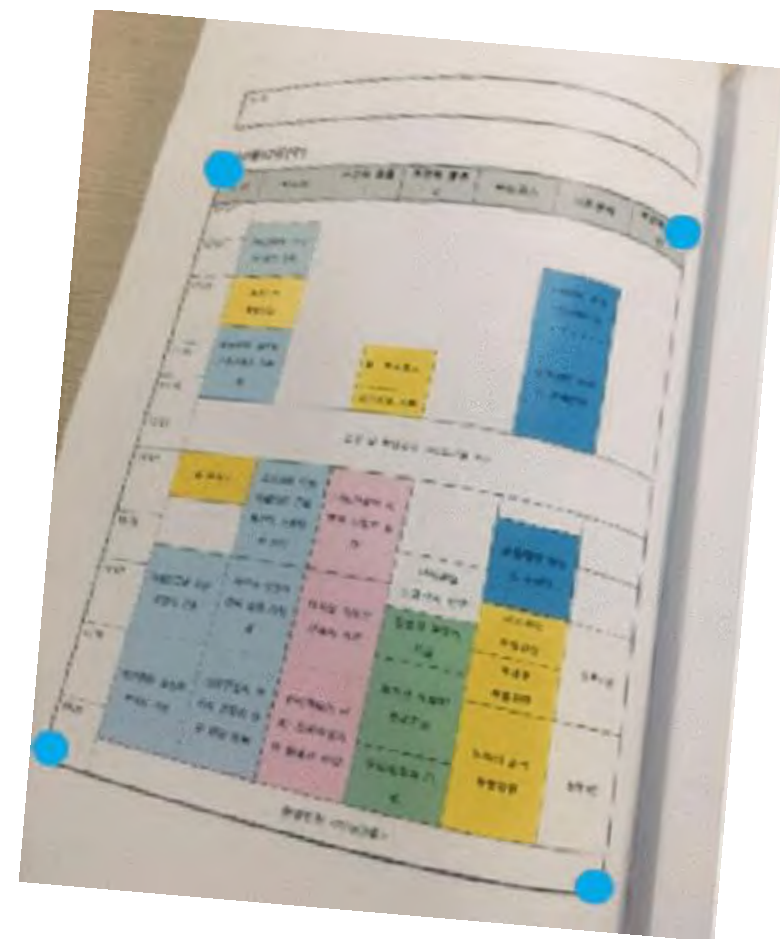
2

图像的畸变校正



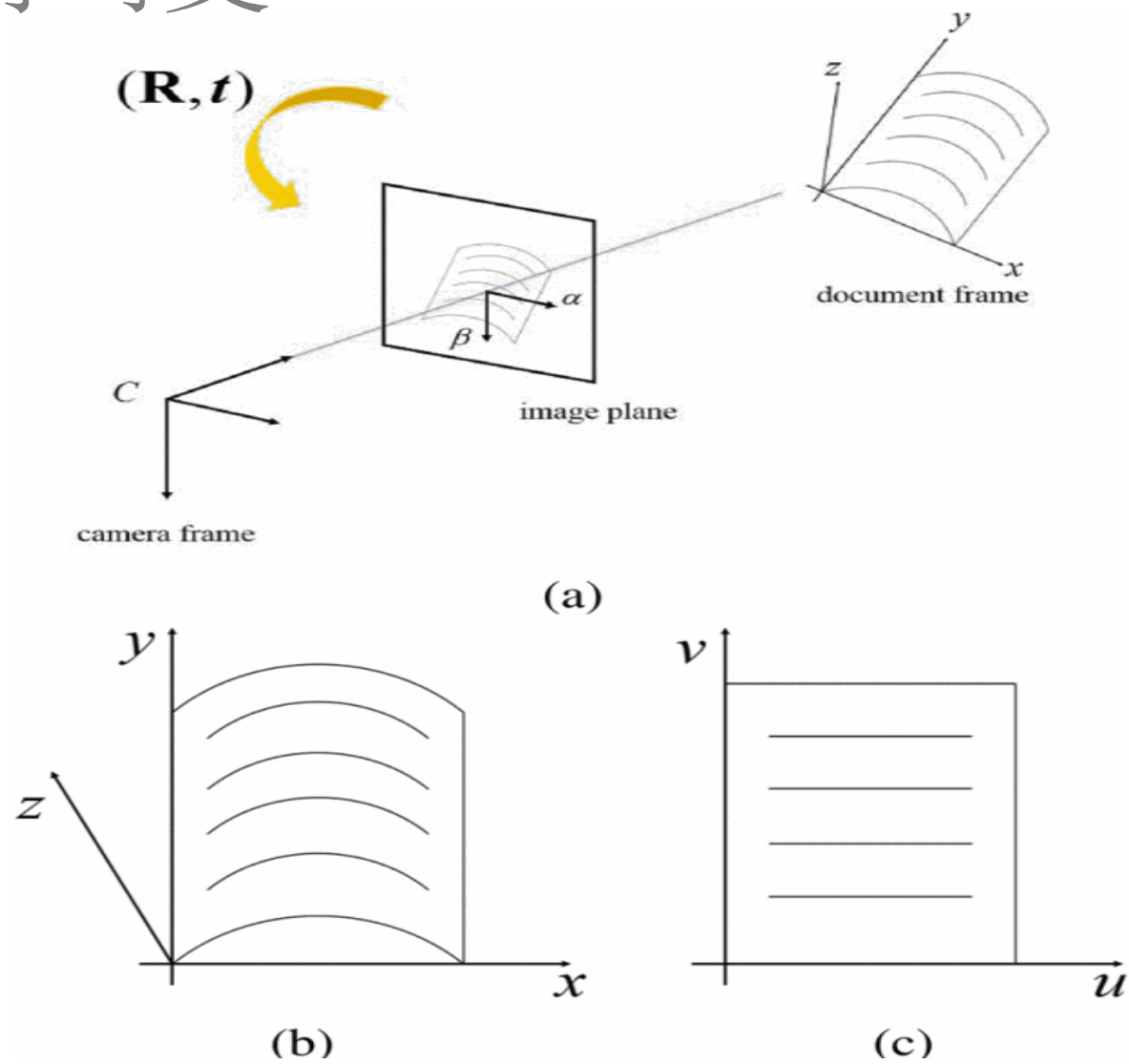
...ein ungeheuer von Prinz Kaltbluth in seine Schranken
verwiesen, eine rothaarige Prinzessin aus den Klauen desselben befreit
und auf die folgenden Abenteuer von Prinz Kaltbluth verwiesen wurde,
in denen es ebenfalls garantiert nicht ungeheuer- oder prinzeßinnenfrei
zugehen würde. Solche Lektüre vermehrt vielleicht nicht den Sprach-

destens dreiköpfiges Ungeheuer von Prinz Kaltbluth in seine Schranken
verwiesen, eine rothaarige Prinzessin aus den Klauen desselben befreit
und auf die folgenden Abenteuer von Prinz Kaltbluth verwiesen wurde
in denen es ebenfalls garantiert nicht ungeheuer- oder prinzeßinnenfrei
zugehen würde. Solche Lektüre vermehrt vielleicht nicht den Sprach-



2

几何畸变



2

几何畸变校正



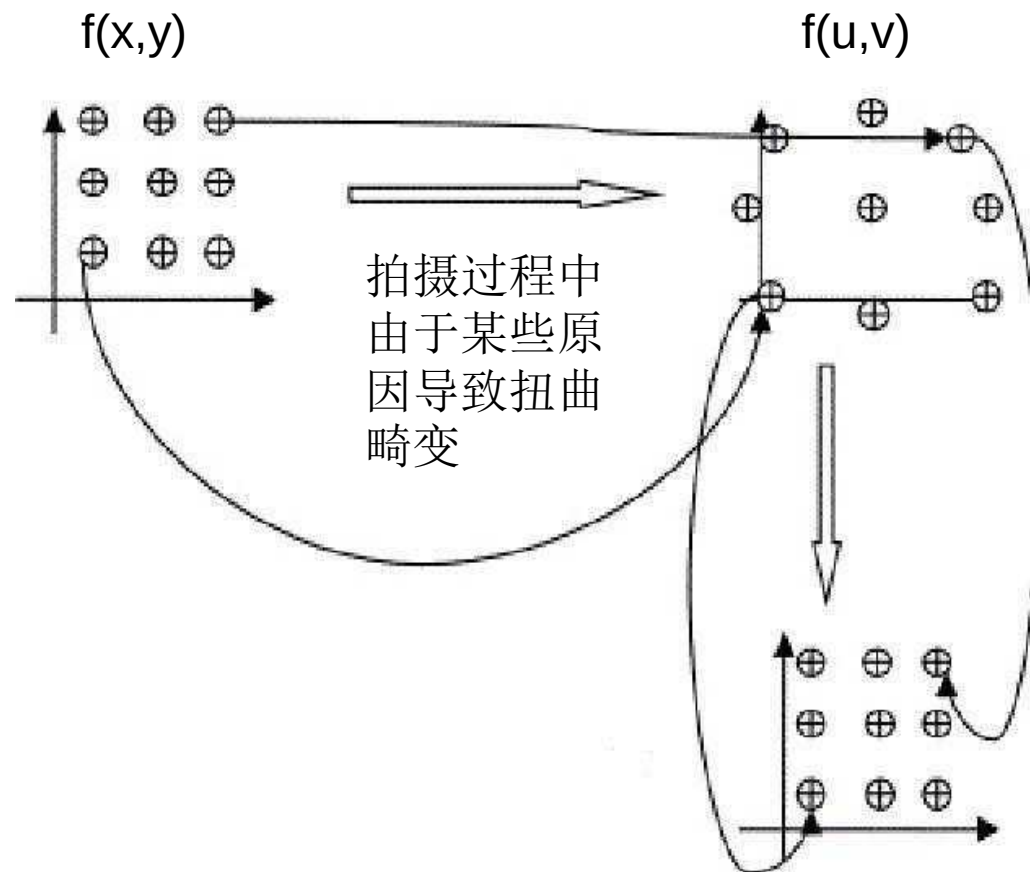
畸变校正步骤:

空间变换

$f(x,y)$ 非畸变图像的像素 $f(u,v)$ 畸变后的图像像素
径向畸变模型 切向畸变模型

确定找出由 $(u,v) \rightarrow (x,y)$ 的坐标变换(地址映射)
对畸变图上像素重新排列以恢复原空间关系

灰度插值



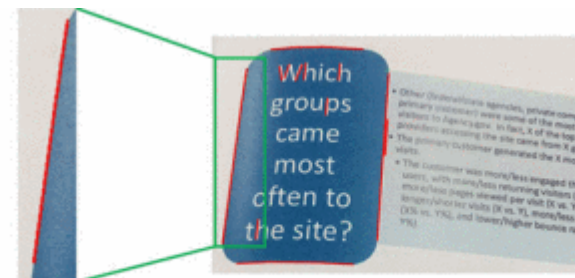
2

图像的倾斜矫正

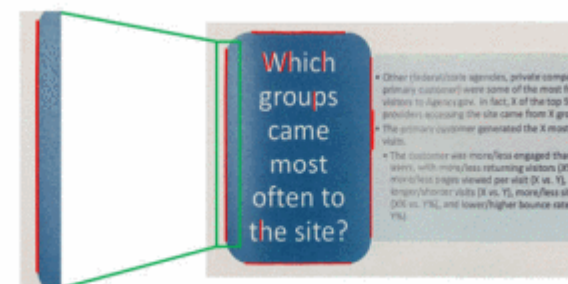
文本图像的倾斜检测方法:

1. 基于投影图的方法

2. 基于Hough变换的方法



Input image



Rectified image

Dewarping process
 Θ



2

图像平滑

中值滤波去除椒盐噪声

10	9	4	7	2
4	31	12	23	30
5	43	1	32	8
54	7	3	43	4
23	76	56	45	9



10	9	4	7	2
4	31	12	23	30
5	43	23	32	8
54	7	3	43	4
23	76	56	45	9

公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市龙湾区 2007年5月12日 No. 00000000000000000000
Donor: Wenzhou City Longwan District Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects(Currency)	数量 Amount	金额 Total amount
书籍	纸张	10000	
金额合计(小写) in Figures: 10000			
金额合计(大写) in Words: 壹万圆整			
接受单位(盖章):	复核人:	开票人:	
Receiver's Seal	Verified by	Handling Person	

感谢您对公益事业的支持! Thank you for support of public welfare!

第二联 收据



公益事业捐赠统一票据(141)
UNIFIED INVOICE OF DONATION FOR PUBLIC WELFARE

捐赠人: 温州市龙湾区 2007年5月12日 No. 00000000000000000000
Donor: Wenzhou City Longwan District Y M D

捐赠项目 For purpose	实物(外币)种类 Material objects(Currency)	数量 Amount	金额 Total amount
书籍	纸张	10000	
金额合计(小写) in Figures: 10000			
金额合计(大写) in Words: 壹万圆整			
接受单位(盖章):	复核人:	开票人:	
Receiver's Seal	Verified by	Handling Person	

感谢您对公益事业的支持! Thank you for support of public welfare!

第二联 收据

2

规范化

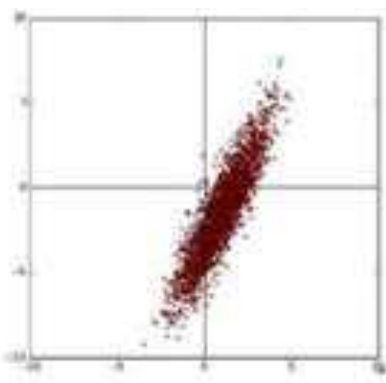
1.位置规范化

消除文字点阵位置上的偏差

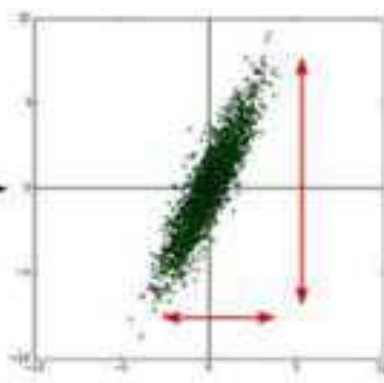
2.大小规范化（归一化）

变换为同一尺寸大小

原始图像



位置规范化



归一化

