



# 上海银行 OCR 分析报告

上海萃舟智能科技有限公司

2018-01-28

## 1. 技术分析

OCR 的基本原理就是通过扫描仪将一份文稿的图像输入给计算机,然后由计算机取出每个文字的图像,并将其转换成汉字的编码。其具体工作过程是,扫描仪将汉字文稿通过电荷耦合器件 CCD 将文稿的光信号转换为电信号,经过模拟/数字转换器转化为数字信号传输给计算机。计算机接受的是文稿的数字图像,其图像上的汉字可能是印刷汉字,也可能是手写汉字,然后对这些图像中的汉字进行识别。对于印刷体字符,首先采用光学的方式将文档资料转换成原始黑白点阵的图像文件,再通过识别软件将图像中的文字转换成文本格式,以便文字处理软件的进一步加工。

目前市场上较为成熟的 OCR 产品有：证件识别 SDK、车牌识别 SDK、文档识别 SDK、银行卡识别 SDK、表格识别 SDK、票据识别 SDK、名片识别 SDK、护照识别 SDK、身份证识别 SDK。目前，银行、保险、金融、税务、海关、公安、边检、物流、电信工商管理、图书馆、户籍管理、审计等很多行业都已经应用了 OCR 技术。OCR 技术让大家减少了设备配置，降低了人力成本，提高了工作效率。

在 OCR 领域，中文 OCR 一直是其中的痛点和难点。随着近年来深度学习的不断发展，中文字符识别精度得到大幅提高。本公司多年从事图像处理和计算机视觉方面的工作，在中文 OCR 方面有着丰富的技术和经验。

结合本公司的技术经验和上海银行的业务需求,本公司为上海银行开发如下票据识别系统。具体识别内容为以下四大要素:流水号、收款人姓名、收款人账号以及收款金额。票据示例如图 1 所示。针对该四大要素,本公司从应用的技术手段给出相应的可行性分析。

图 1

上海银行  
Bank of Shanghai

OCX流水号: 201801303304805002538

个人业务凭证(填单)

附件

张

|  |  |                    |   |   |   |   |      |  |   |   |   |
|--|--|--------------------|---|---|---|---|------|--|---|---|---|
| 业务类型   | <input type="checkbox"/> 活期存款 (活期存款) <input type="checkbox"/> 无卡(折)存款<br><input type="checkbox"/> 汇款 <input type="checkbox"/> 申请支票 <input type="checkbox"/> 兑付支票 <input type="checkbox"/> 其他 |                    |   |   |   |   |      |  |   |   |   |
|  | 户名   | 陆东升                |   |   |   |   | 扣款方式 | <input type="checkbox"/> 现金 <input checked="" type="checkbox"/> 转账 |   |   |   |
|  | 账号   | 620522001004338226 |   |   |   |   | 扣款方式 | <input type="checkbox"/> 普通 <input type="checkbox"/> 加急            |   |   |   |
|  | 开户行  |                    |   |   |   |   | 用途   |  |   |   |   |
| 币种: <input type="checkbox"/> 人民币 <input type="checkbox"/> 美元 <input type="checkbox"/> 欧元 <input type="checkbox"/> 港币 <input type="checkbox"/> 日元 <input type="checkbox"/> 其他 <input type="checkbox"/> 英镑 <input type="checkbox"/> 港币 |  |                    |   |   |   |   |      |  |   |   |   |
| 金额   | 亿  | 千                  | 百 | 十 | 万 | 千 | 百    | 十  | 元 | 角 | 分 |
|  |  |                    |   |   | 3 | 7 | 7    | 1  | 9 | 0 | 0 |

客户确认

本人已阅  
户须知” 兹  
资料真实,有  
位要素填写正  
银行照此办理

客户签名: 27

银行填写

交易时间: 2018/01/30 12:34:44 交易类型: 活期转账  
 付款人姓名: 沈如菊  
 付款人账号/卡号: 622468001018405260  
 收款人户名: 陆东升  
 收款人账号/卡号: 620522001004338226  
 交易流水号: 201801303304805002538 核心流水号: TT18030081507017  
 转账金额: 37,719.00 CNY 手续费: 0.00 汇划费: 0.00  
 备注:  
 网点号: C00015107 操作员号: 330480

识别的四大要素：流水号、收款人姓名、收款人账号和收款人金额，在图2

和图 3 中分别由①②③④所标识。

图 2

上海银行 Bank of Shanghai

个人业务凭证(填单) 附件 张

OCX流水号: 201801303304805002538 ①

|           |  |  |
|-----------|--|--|
| 业务类型      | <input type="checkbox"/> 人民币 (活期/定期) <input type="checkbox"/> 外币 (折) 存款  |  |
|           | <input type="checkbox"/> 汇款 <input type="checkbox"/> 申请本票 <input type="checkbox"/> 兑付本票 <input type="checkbox"/> 其他  |  |
| 收款人       | 户名: 陆东升  | 扣款方式: <input type="checkbox"/> 现金 <input checked="" type="checkbox"/> 转账 |
|           | 账号/卡号: 620522001004338226  | 收款方式: <input type="checkbox"/> 普通 <input type="checkbox"/> 加急            |
|           | 开户行:   | 用途:  |
| 币种:       | <input type="checkbox"/> 人民币 <input type="checkbox"/> 美元 <input type="checkbox"/> 欧元 <input type="checkbox"/> 港币 <input type="checkbox"/> 日元 <input type="checkbox"/> 其他 <input type="checkbox"/> 英镑 <input type="checkbox"/> 澳元 |  |
| 金额        | 亿 千 百 十 万 千 百 十 元 角 分  |  |
|           | ¥ 3 7 7 1 9 0 0  |  |
| 交易时间:     | 2018/01/30 12:34:44 交易类型: 活期转账   |  |
| 付款人姓名:    | 沈如菊  |  |
| 付款人账号/卡号: | 622468001018405260   |  |
| 收款人户名:    | 陆东升 ②  |  |
| 收款人账号/卡号: | 620522001004338226 ①   |  |
| 交易流水号:    | 201801303304805002538 ①  |  |
| 转账金额:     | 37,719.00 CNY  |  |
| 备注:       | 手续费: 0.00 汇划费: 0.00  |  |
| 网点号:      | CN0015107 操作员号: 330480   |  |

客户签名: ②

图 3

上海银行 Bank of Shanghai

个人业务凭证(填单) 附件 张

OCX流水号: 201801303304805002538

|           |  |  |
|-----------|--|--|
| 业务类型      | <input type="checkbox"/> 人民币 (活期/定期) <input type="checkbox"/> 外币 (折) 存款  |  |
|           | <input type="checkbox"/> 汇款 <input type="checkbox"/> 申请本票 <input type="checkbox"/> 兑付本票 <input type="checkbox"/> 其他  |  |
| 收款人       | 户名: 陆东升  | 扣款方式: <input type="checkbox"/> 现金 <input checked="" type="checkbox"/> 转账 |
|           | 账号/卡号: 620522001004338226  | 收款方式: <input type="checkbox"/> 普通 <input type="checkbox"/> 加急            |
|           | 开户行:   | 用途:  |
| 币种:       | <input type="checkbox"/> 人民币 <input type="checkbox"/> 美元 <input type="checkbox"/> 欧元 <input type="checkbox"/> 港币 <input type="checkbox"/> 日元 <input type="checkbox"/> 其他 <input type="checkbox"/> 英镑 <input type="checkbox"/> 澳元 |  |
| 金额        | 亿 千 百 十 万 千 百 十 元 角 分  |  |
|           | ¥ 3 7 7 1 9 0 0  |  |
| 交易时间:     | 2018/01/30 12:34:44 交易类型: 活期转账   |  |
| 付款人姓名:    | 沈如菊  |  |
| 付款人账号/卡号: | 622468001018405260   |  |
| 收款人户名:    | 陆东升  |  |
| 收款人账号/卡号: | 620522001004338226 ③   |  |
| 交易流水号:    | 201801303304805002538  |  |
| 转账金额:     | 37,719.00 CNY ④  |  |
| 备注:       | 手续费: 0.00 汇划费: 0.00  |  |
| 网点号:      | CN0015107 操作员号: 330480   |  |

客户签名: ②

票据识别系统的技术流程为: (1) 图像预处理; (2) 文字行提取; (3) 文字行字符识别; (4) OCR 后处理。

在票据扫描过程中, 由于光照、拍摄角度、字体印刷以及印章等的影响, 扫描后的字体质量并不能达到最优。为了避免噪声、角度倾斜造成的干扰, 在字符识别之前, 我们需要对票据进行旋转、仿射变换以及二值化等相关预处理操作, 使得票据中的字符呈现最佳效果。此外, 为了利用票据中的直线信息, 并消除直线对之后字符识别的影响, 在预处理过程中, 我们还需确定票据中直线的位置信息, 在预处理完成后, 首先我们利用依据直线信息对票据中的文字信息进行划分, 其初步结果如图 4 所示。



图 4 依据直线进行文字区域提取

| 业务类型   |       | <input type="checkbox"/> 账户开户（活期除外） <input type="checkbox"/> 无卡（折）存款  |  |
|--|-------|---|--|
|  |       | <input type="checkbox"/> 汇款 <input type="checkbox"/> 申请本票 <input type="checkbox"/> 兑付本票 <input type="checkbox"/> 其他 |  |
| 收款人  | 户名    | 陆东升   |  |
|  | 账号/卡号 | 620522001004338226  |  |
|  | 开户行   |   |  |
| 币种： <input type="checkbox"/> 人民币 <input type="checkbox"/> 美元 <input type="checkbox"/> 欧元 <input type="checkbox"/> 港币 <input type="checkbox"/> 日元 <input type="checkbox"/> 其他 |       | 付款方式  | <input type="checkbox"/> 现金 <input checked="" type="checkbox"/> 转账 |
|  |       | 还款方式  | <input type="checkbox"/> 普通 <input type="checkbox"/> 加急            |
|  |       | 用途  |  |
|  |       | <input type="checkbox"/> 现钞 <input type="checkbox"/> 现汇   |  |
| 金额   | 亿     | 千   | 百  |
|  | 十     | 万   | 千  |
|  |       | 百   | 十  |
|  |       | 元   | 角  |
|  |       | 分   |  |
|  |       | 37,719.00   |  |

本人已阅读本凭证背面“客户须知”，兹确认所提供的业务资料真实、有效且左方/背面栏位要素填写正确、无误，并授权银行照此办理。

客户签名：沈如菊

交易时间：2018/01/30 12:34:44 交易类型：活期转账

付款人姓名：沈如菊

付款人账号/卡号：622468001018405260

收款人户名：陆东升

收款人账号/卡号：620522001004338226

交易流水号：201801303304805002538 核心流水号：TT18030081507017

转账金额：37,719.00 CNY 手续费：0.00 汇划费：0.00

备注：

网点号：CNO015107 操作员号：330480

接着在划分后的图像中采用深度学习算法，进行文字行提取，部分文字行提取结果如图 5 所示。

图 5 文字行检测结果

个人业务凭证(填单)

OCX流水号:201801303304805002538

交易时间:2018/01/30 12:34:44 交易类型:活期转账

付款人姓名:沈如菊 收款人户名:陆东升

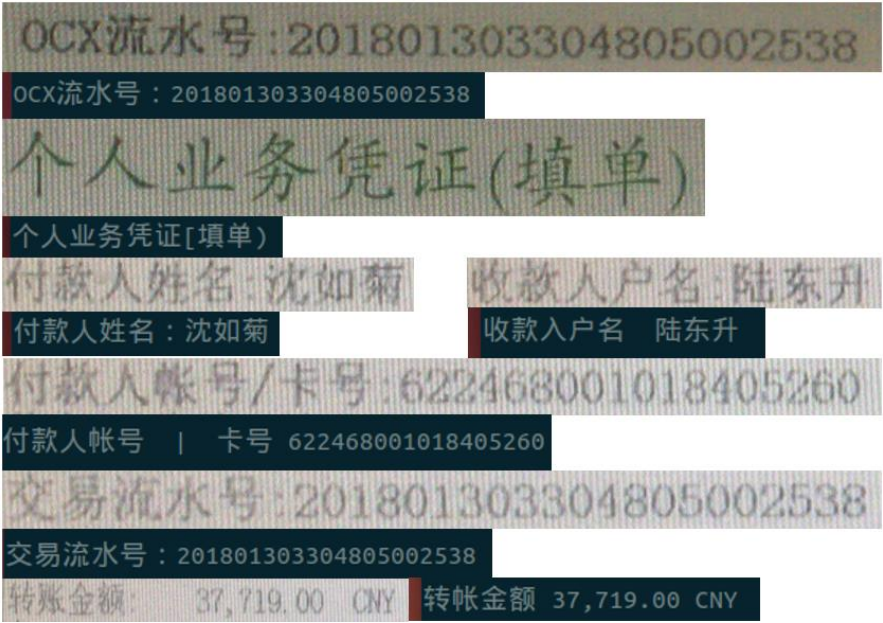
付款人账号/卡号:622468001018405260

交易流水号:201801303304805002538


转账金额: 37,719.00 CNY 手续费: 0.00 汇划费: 0.00

在文字行提取结束后，我们对每个单独的文字行进行字符识别。字符识别通常有两种思路：（1）字符分割+单个字符分类；（2）单行文字整体识别。方法（1）为传统 OCR 方法的思路，但其往往忽略了文字之间的语义关系。为了充分利用字符之间的语义关系，我们采用深度网络进行字符行识别，其识别精度相比方法（1），效果更优。其部分识别结果如图 6 所示。

图 6 字符识别结果



由于中文 OCR 目前精度仍不能达到 100%。在票据识别的过程中难免出现个别字符错误的情况。而此次任务中，我们识别的票据中四个要素的关键词相对固定，不同的要素间格式也不尽相同。所以，我们可采用模糊算法进行匹配，以修正中文 OCR 的个别字符的错误。如②中“收款人姓名：陆东升”被识别为“收款入姓名：陆东升”，我们依据字符串相似度以及其后字符串为中文字段，可将“入”模糊修正为“人”。



上海银行  
Bank of Shanghai

## 补发账务证明申请书

说明收到  
已遗失类，补清后发

本单位账户的下划数据由银行于 年 月 日记账，但账务证明已遗失类，补清后发

财务专用章

上海浦东发展银行

张才才

13331807180 张

2018年1月24日

申请账号: 0000971604

|                              |        |                   |     |        |                   |
|------------------------------|--------|-------------------|-----|--------|-------------------|
| 付款人                          | 户名     | 上海浦东发展银行          | 收款人 | 户名     | 上海浦东发展银行          |
| 开户行                          | 账号     | 310301 0000971604 | 开户行 | 账号     | 310301 0000971604 |
| 币种                           | 金额(大写) | 人民币               | 币种  | 金额(小写) | 人民币               |
| 附加信息或用途: 2018年10月 - 2017年12月 |        |                   |     |        |                   |

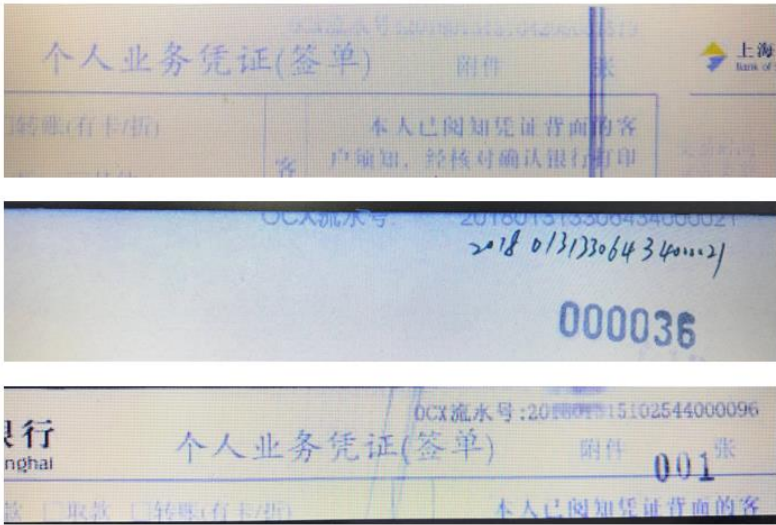
(以下内容需由申请单位填写)

上列数据确由我行于 2018 年 月 日记账，特此补发账务证明，原需各证明作废。

补发编号: 银行盖章



图 2-2 无效图像 （含流水号，人眼无法判读真值）



第一轮测试精度如表 2-2。

表 2-2

| 图像类型               | 数量      |
|--------------------|---------|
| 处理图像总量             | 17867   |
| 含流水号的图像<br>(人眼不可辨) | 215     |
| 含流水号的图像<br>(人眼可辨)  | 9508    |
| 软件识别出包含流水号图像       | 9184    |
| 软件正确识别流水号的图像       | 8843    |
| 流水号图片检出率           | 96. 59% |
| 准确率                | 93. 00% |

其中，“处理图像总数”为所有票据的正面图像（图片名为奇数），“有效图片”为包含流水号的票据图片。

“流水号图片检出率” Ratio 的计算方式为：

$$\text{Ratio} = \frac{\text{软件识别出含流水号的图像数量}}{\text{含人眼可辨流水号的图像数量}}$$

“准确率” Accuracy 的计算方式为：

$$\text{Accuracy} = \frac{\text{正确识别流水号的图像数量}}{\text{含人眼可辨流水号的图像数量}}$$

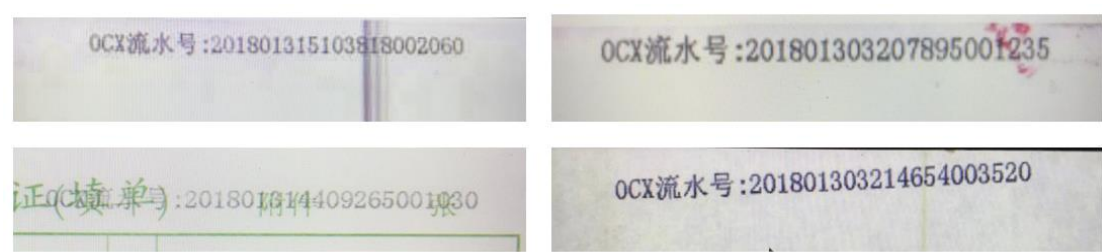
根据第一轮测试结果, 精度分析和提升计划如表 2-3。

表 2-3

| 识别错误的图像种类      | 精度提升方法     |
|----------------|------------|
| 图像旋转角度大或倒放     | 影像纠正处理     |
| 字体褪色严重         | 图像预处理      |
| 字迹被污染          | 图像预处理      |
| 字迹缺损           | 深度学习模型迭代训练 |
| 改进后预计保障精度： 95% |            |

识别错误的图像种类图例如图 2-3。

图 2-3 识别错误样例：字迹被污染，字体旋转角度大



2.3 效率分析

第一轮测试硬件环境如下表 2-3。

表 2-3

|                       |                                   |
|-----------------------|-----------------------------------|
| 系统环境                  | Win7 使用 vmware 安装的 ubuntu16 虚拟机系统 |
| CPU                   | Intel i5-4590 3.3Ghz              |
| 分配核心数                 | 2                                 |
| 分配内存                  | 3.8G                              |
| 硬盘                    | 90G                               |
| 累计运行时间                | 16.4 小时                           |
| 处理图像总量                | 19453 张                           |
| 运行速度                  | 1186 张/小时                         |
| 5 台同配置机器<br>处理 5 万张时间 | 8.4 小时                            |

提升硬件条件，数据处理效率水平如下表 2-4。

表 2-4

|     |                 |
|-----|-----------------|
| CPU | Intel i7-7700HQ |
| 内存  | 8G              |



|                       |            |
|-----------------------|------------|
| 核心数                   | 2.8GHz * 8 |
| 预计处理效率                | 2860 张/小时  |
| 4 台同配置机器<br>处理 5 万张时间 | 4.37 小时    |