ICS 03.060
CCS A 11

T/NBFS

团 体 标 准

T/NBFS 4—2022

智能文字识别技术在金融领域的应用系统设计指南

2022－08－01 发布　　2022－08－01 实施

宁波市金融学会　　发 布

目　　次

前　言

本文件按照GB/T 1.1-2020 《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由宁波市金融学会提出并归口。

本文件起草单位：中国人民银行宁波市中心支行、宁波银行股份有限公司、中国工商银行股份有限公司宁波市分行、宁波东海银行股份有限公司、东海航运保险股份有限公司、中国银行股份有限公司宁波市分行、招商银行股份有限公司宁波市分行、宁波通商银行股份有限公司、宁波鄞州农村商业银行股份有限公司、甬兴证券有限公司。

本文件主要起草人：王去非、张文元、袁冬勤、黄宪、关义生、张热弯、王巧燕、熊强、周泉、程东、崔霄翔、方诗伟、董逸飘、张芝悦、吕亚男、毛伏韬、陈建群、陈少亮。

智能文字识别技术在金融领域的应用系统设计指南

1 范围

本文件规定了在金融行业中采用智能文字识别技术进行业务单据识别的系统应具有的功能、性能和安全要求。

本文件适用于金融领域智能文字识别系统的设计与实现，对智能文字识别系统的测试、管理也可参照使用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 17961-2010 《印刷体汉字识别系统要求与测试方法》

GB/T 34080.3-2021 基于云计算的电子政务公共平台安全规范 第3部分：服务安全

GB/T 34084-2017 中文语音识别互联网服务接口规范

GB/T 37973-2019 《信息安全技术 大数据安全管理指南》

GB/T 40343-2021 智能实验室 信息管理系统 功能要求

JR/T 0185-2020 《商业银行应用程序借口安全管理规范》

3 术语和定义

3.1 自然语言处理

是人工智能和语言学的分支学科，研究如何让计算机处理及运用人类自然语言，包括对语言的认知、理解、生成等，并按人所定义和预期的目标进行正确返回。

3.2 结构化

指在OCR的文字识别结果上，结合自然语言处理、先验规则等信息提取用户所需要的字段信息的方法。

3.3 训练

指在机器学习类方法中，通过设定目标适应函数和基于此目标的一套反馈系统来从样本数据中学习得到达成目标的最佳模型的过程。

3.4 模板识别

指从待识别图像中提取若干特征向量与模板对应的特征向量进行比较，计算图像与模板特征向量之间的距离，用最小距离法判定所属类别。

3.5 错误样例

一般指异常场景。在OCR文字识别中，指经过文字识别与结构化后字段的准确率或召回率较低的样本。

3.6 识别准确率

通过图像识别后提取出的正确信息条数占提取出的信息条数的比重，定义如下：

准确率=提取出的正确信息条数/提取出的信息条数。

在OCR领域，字符识别准确率是指以单字符为统计单位的准确率；单词识别准确率是指以英文单词为统计单位的准确率；字段识别准确率是指以结构化后的字段为统计单位的准确率。

3.7 识别召回率

通过图像识别后提取出的正确信息条数占样本中存在的信息条数的比重，定义如下：

召回率=提取出的正确信息条数/样本中存在的信息条数。

在OCR领域，字符识别召回率是指以单字符为统计单位的召回率；单词识别召回率是指以英文单词为统计单位的召回率；字段识别召回率是指以结构化后的字段为统计单位的召回率。

3.8 模板识别准确率

通过模板识别分类正确的样本数占所有样本数的比重，定义如下：

召回率=分类正确的样本数/所有样本数。

3.9 恢复时间目标 Recovery Time Objective;RTO

指灾难发生后，信息系统或业务功能从停顿到必须恢复的时间要求。

3.10 恢复点目标 Recovery Point Objective;RPO

指灾难发生后，系统和数据必须恢复到的时间点要求。

3.11 JSON 数据交换格式 JavaScript Object Notation;JSON

一种数据交换格式。

[GB/T 34083-2017,定义 3.10]

4 缩略语

下列缩略语适用于本文件：

DPI：每英寸点数（Dots Per Inch）

ICR：智能字符识别（Intelligent Character Recognition）

OCR：光学字符识别（Optical Character Recognition）

5 功能要求

5.1 概述

智能文字识别系统应包含两部分：客户端和服务器端。其中客户端主要是指具有照片实时拍摄能力、且支持安装第三方研发软件的手机、平板等可携带设备上的应用程序，个人电脑、高拍仪、独立摄像头等终端不适用，服务器端是指服务器端的应用系统。

智能文字识别系统功能模块流程图如图1所示。

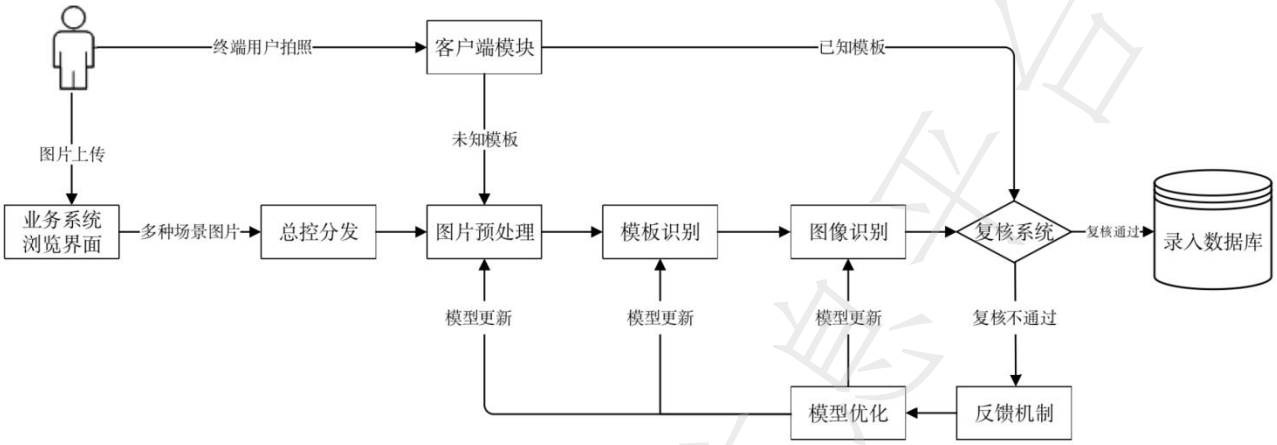

图1　智能文字识别系统功能模块流程图

注1：客户端应用程序应具有如下功能：

——实时检测已知模板位置功能，以快速提示用户拍摄是否正确；

——识别已知模板并结构化输出；

——未知模板传送到后台系统识别。

注2：服务器端应用系统应具备图像预处理、模板识别、图像识别、错误样例收集等功能。

5.2　图像预处理

5.2.1　输入图像要求

输入图像应保证在关键字段位置处文字成像无模糊、无光斑、无明显畸变、无明显噪声干扰，图像分辨率在200DPI以上。

5.2.2　输出图像要求

输出图像应满足以下要求：

a)　因模糊、块效应、噪声、畸变等因素导致的文字失真的图像应该被过滤；

b)　非单据区域部分应该被自动裁剪，只保留单据在图像中所在的主体区域；

c)　对于出现旋转的文本应被自动纠偏，输出的图像的文本应保持在水平位置；

d)　若对印章无识别要求，输入图像中的红色印章痕迹应被去除，以减少对单据正文识别处理的干扰。

5.3　图像识别

5.3.1　图像识别内容

在GB/T 17961-2010中5.1系统功能要求基础上，除了具备从表格、文字中识别可编辑的编码文本，还应具备从文本中提取出具有业务意义的字段信息的功能。

5.3.2　固定模板识别

针对识别图像的内容版面、布局等都一致的场景，采用固定模板识别，输入的单张或者多张图片应为已知模板，在输入为非已知模板情况下应在输出结果中明确告知该输入为非支持模板，在输入为已知模板情况下应输出约定结构化格式。

5.3.3 非固定模板识别

针对识别图像的内容版面、布局等都各有不同的场景，采用非固定模板识别，输入的单张或者多张图片可以为未知模板，通过文字识别结合自然语言处理等技术将输入图像内容对应到已录入的业务类型并输出约定结构化格式，在对应业务类型未录入的情况下应在输出结果中明确告知输入为非支持的业务类型。

5.4 反馈机制

ICR系统应具备反馈机制，反馈机制应包含接入复核系统进行联动，制定可持续的、自动化的错误样本收集流程，并对原有模型进行修正替代，使模型训练、识别、反馈三大环节形成闭环。反馈机制应通过错误样例对模型做优化，并根据业务发展情况提供定期或不定期的线上图像模型更新。

5.5 总控端接口与识别模式

5.5.1 总控端接口设计

总控端接口设计应遵循如下规则：

a) 对外提供统一的识别接口，各应用系统可以快速接入，使用对应的文字识别服务，并做好负载均衡和流量控制，接口详细设计可参考 JR/T 0185-2020；

b) 接口报文宜采用 JSON 数据交换格式格式，对应用唯一标识进行存储与统一管理，并根据应用唯一标识进行应用身份认证、状态校验和权限控制等。接口具备流量监控、故障隔离、黑名单控制等异常检测能力。

5.5.2 总控端识别模式设计

总控识别模式应包含：

a) 同步模式，用户发起一个识别请求后，需要等待识别结果返回才能发起第二个识别请求，识别结果通常是在提交识别请求页面直接返回。适用于图片识别速度快的场景；

b) 异步模式，用户发起一个识别请求后，无需等待就能发起第二个识别请求，识别结果以异步方式通知到用户。适用于图片识别速度慢的场景。

6 性能要求

6.1 技术指标

识别性能指标要求如表1所示。

表1 识别性能指标

流程	指标要求
模板识别	模板识别准确率应>95%
图像识别	字符识别召回率应>95%，字符识别准确率应>99%
识别速度	A4纸幅面中文识别速度不低于1000字/s,英文识别速度不低于2000字/s，识别速度较GB/T 17961-2010中5.2.4的150字/s有一定幅度提升

字符集与字体	应符合GB/T 17961-2010中5.2.1和5.2.2的要求

系统运行指标要求如表2所示。

表2 系统运行指标

指标	指标要求
可靠性	包含校验机制和解耦机制，校验机制包括但不限于人工复核，解耦机制指在ICR系统出现故障时可以通过一键切换方式切换到人工验证模式
可用性	每年非计划服务中断时间不超过4天，系统可用性至少达到99%(3级)
RTO	不超过24小时
RPO	不超过24小时

7 安全要求

7.1 系统安全

ICR系统的安全设计应遵循如下要求：

a) 服务访问安全控制应符合 GB/T 34080.3-2021 中 5.1.1 的要求；

b) 图像数据处理中的规范控制、异常告警、存储管理和关键数据流程应符合 GB/T 40343-2021 7.1.4 的要求；

c) 系统输入只接收图片类文件，对非图片类文件应在总控端进行拒绝识别处理。

7.2 样本安全

ICR系统中需要为样本数据采取安全保护措施，参考GB/T 37973-2019中的要求。

参 考 文 献

[1] JR/T 0056 《票据影像采集交换技术规范 影像采集》。
[2] 2D Attentional Irregular Scene Text Recognizer. arXiv preprint arXiv:1906.05708, 2019.
[3] Learning Shape-Aware Embedding for Scene Text Detection. In CVPR, 2019.