

# 档案云解决方案

厦门云脉技术有限公司



第一章

背景和现状

第二章

方案介绍

第三章

应用和案例

第四章

公司简介

第一章 背景和现状

2 需求分析

# 纸质文档管理现状

大数据时代,企业越来越重视从大数据中挖掘潜在价值,而大量信息资源在各种纸质文档中,文档电子化势在必行。目前市场上的文档管理系统在实践中,仍存在以下问题和挑战:

- □ 录入难:人工录入工作量繁重,易出错;
- **核对难**:电子化的文档很难快速与原文档进行匹配,电子信息准确度验证麻烦;
- 分类难: 纸质文档电子化后无法做到影像的自动分类和管理;
- 检索难: 纸质文档归类储藏后, 往往堆积在一起, 很难快速查找到需要的文档;
- ◎ 效率低:纸质信息无法及时录入系统延缓整个工作流程,工作效率大打折扣。

## 需求分析

根据纸质文档管理现状,实现纸质文档数据的集中管理与协作共享,必须彻底解决文档自动录入、自动分类、快速检索等问题,需求如下:

- ◎ 文件资料的批量扫描识别;
- □ 档案综合管理:实现文件管理、档案检索、档案利用等;
- 与第三方系统对接,便于系统升级、程序间数据的共享、节约海量存储空间以及减少网络负载等;
- □ 业务数据后台集中管理、监控;
- □ 完善的数据安全机制和备份机制。

# 第二章 方案介绍

- 1 总体设计
- 2 OCR引擎
- 3 纸质文档管理
- 4 OCRIF
- 5 方案优势

# 方案简介

为解决传统的管理手段查询效率低和档案管理难度大的问题,同时为更好地保存档案和实现档案信息共享,厦门云脉档案云解决方案运用现代信息技术对数字档案信息进行OCR快速采集、加工、存储、管理,并通过网络平台提供档案信息服务和共享利用的档案信息集成管理系统。

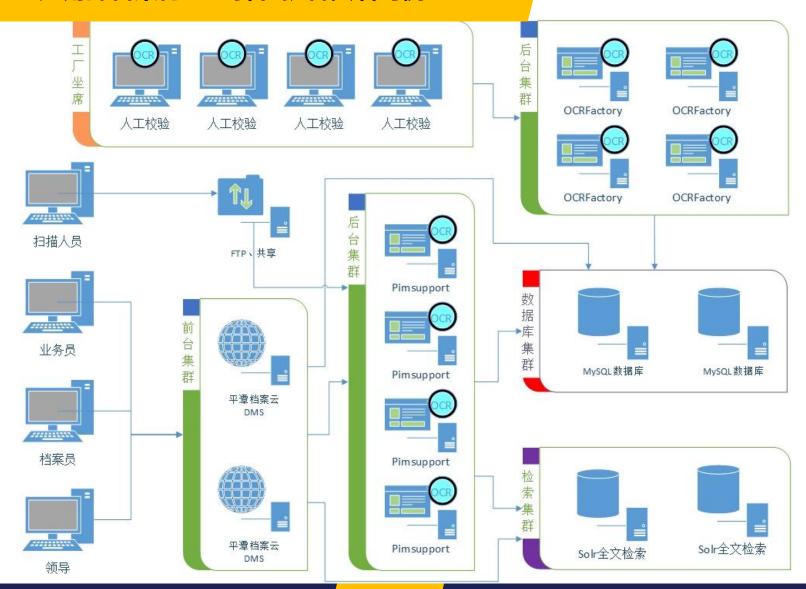
#### 建设思路

- 1、纸质文档管理系统,搭建档案数字化管理系统,建立统一的档案数据库。
- 2、OCR工厂,利用影像工作流对业务数据进行影像化处理,并实现对影像进行自动判别、处理、传递、存储等,运用工作流控制、影像处理等技术手段,实现相关业务的后台集中处理。

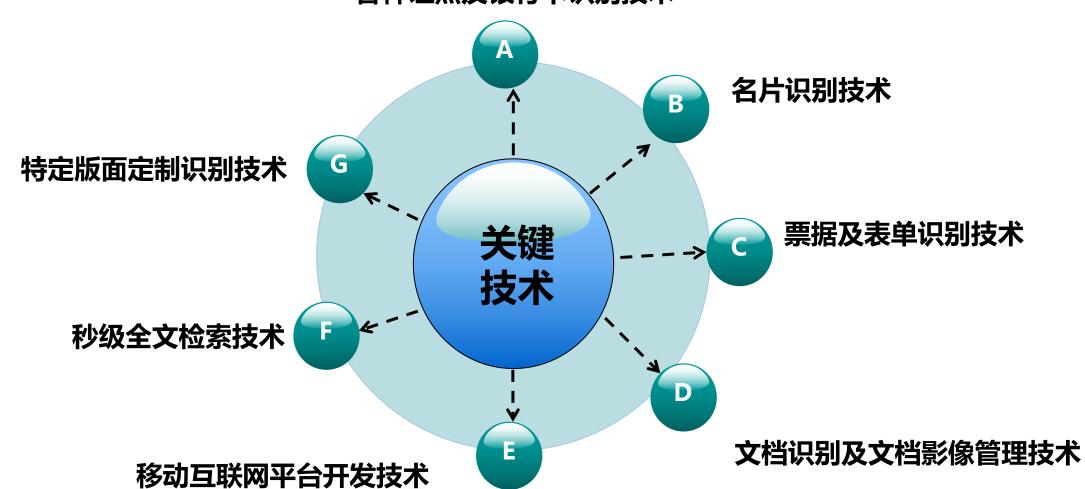
## 系统架构——模块化服务架构



## 应用部署架构——弹性扩容负载均衡



## 各种证照及银行卡识别技术



#### 技术架构

PC应 用 浏览 器

安卓

IOS

Web 前端应用

后台服务应用

Tomcat 8

**JDK 1.8** 

MySQL 5.5

Win Server / Linux Server

影像系统采用Java开发
OCR引擎采用C++开发
人工校验客户端采用.Net开发
模板管理客户端采用.Net开发

影像系统前台提供源代码 影像系统后台不提供源代码 OCR引擎不提供源代码 人工校验客户端不提供源代码 模板管理客户端不提供源代码

## OCR引擎

## 图像自动切边、倾斜校正、增强美化、压缩存储



500W像素或400dpi以上扫描仪

OCR识别+图像增强+压缩加密存储

A4彩色可压缩至200K,黑白至60K左右

## 电子图像美化

灰度/彩色切换、图像二值化、图像切边









数据中心

#### 技术参数

- 轻松识别图片、PDF、扫描件上的文字
- □ 可正常识别彩色(24BITS)、灰度(256阶)图像。支持TIFF、BMP、JPG、PDF格式的图像文件
- 🗎 可正常输出GB2312、GB18030、BIG5、ASCII、Unicode等字符集文本
- □ 自动裁边、修正倾斜、去除多余的复杂背景、图像小于20度角倾斜也能精确识别
- □ 具备自动高亮对照图像与识别文字的识别结果人工校正功能
- ᠍ 提供候选字和联想字供选择,识别结果快速校正



## 文档识别性能

## 测试集:500W中文、800W中文、中文不同类型

| 产品                            | 识别率    |  |  |  |
|-------------------------------|--------|--|--|--|
| 云脉文档识别_V4.14.160614_170630    | 90.46% |  |  |  |
| 安卓扫描全能王_3.8.0.20150711        | 88.76% |  |  |  |
| 汉王PDF-OCR文字识别8.1.0.3_20171019 | 74.08% |  |  |  |
| ABBYY_1.9.1                   | 90.39% |  |  |  |

# OCR引擎

## 表单识别性能

| 产品    | 栏目识别率   | 字符识别率  | 总识别率   |
|-------|---------|--------|--------|
| 护照    | 99.05%  | 98.84% | 97.90% |
| 身份证   | 100.00% | 99.16% | 99.16% |
| 港澳通行证 | 97.00%  | 93.76% | 90.95% |
| 增值税发票 | 96.13%  | 92.44% | 88.86% |
| 营业执照  | 94.08%  | 90.44% | 85.09% |
| 户口本   | 100.00% | 97.28% | 97.28% |
| ••••• | •••••   | ••••   | ••••   |

#### 批量上传识别

支持通过Http、Ftp、共享目录以及第三方系统接口的方式批量上传,后端队列识别。



#### 图文对照

## 电子化的文档与原文档对照显示,轻松实现电子信息的准确度验证。

之規定, 我会决定对起重机械专委会成员单位广西建工集团建筑机械 制造有限责任公司进行如下处理:

从事故发生之日起,暂停该成员单位(广西建工集团建筑机械制 造有限责任公司)在珠海市承接建筑起重机械(塔吊、施工电梯、物 料提升机)设备的安拆、顶升、维保等相关业务,并报市、区安监站 备案。

希望各施工、监理单位以及起重机械专委会成员单位深刻吸取 "5·22"事故教训,举一反三,引以为戒,并严格执行《建筑起重 机械安全监督管理规定》(建设部 166 号文)、《珠海市建筑起重机 械安全监督管理办法》、《关于进一步规范建筑起重机械专业委员会 成员单位经营业务管理工作的通知》(珠建安协【2016】006 号)的 相关规定,落实安全生产责任制和相关安全管理规定,强化安全生产 日常管理,加强安全技术指导和安全培训工作,提升作业人员的技术 水平和安全意识,防范安全事故再次发生。

特此通告。







#### 智能纠错

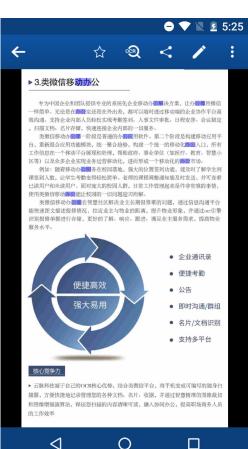
提供联想字符供选择,实现快速纠正错误。



#### 全文秒级搜索

支持精确查找、模糊查找和多条件组合检索电子图像,查找的关键字在图片和识别 文本上高亮显示。





## 导出共享

- □ 导出支持txt、word、双层pdf和以及纯文本pdf
- 通过邮件或其他应用程序分享文档
- 手机、平板、电脑等多终端随时同步,一处编辑,实时更新。



#### 自动分类

## 根据传入图像特征,引擎智能查找匹配模板,并根据模板对其进行分类。



#### 系统监控和管理

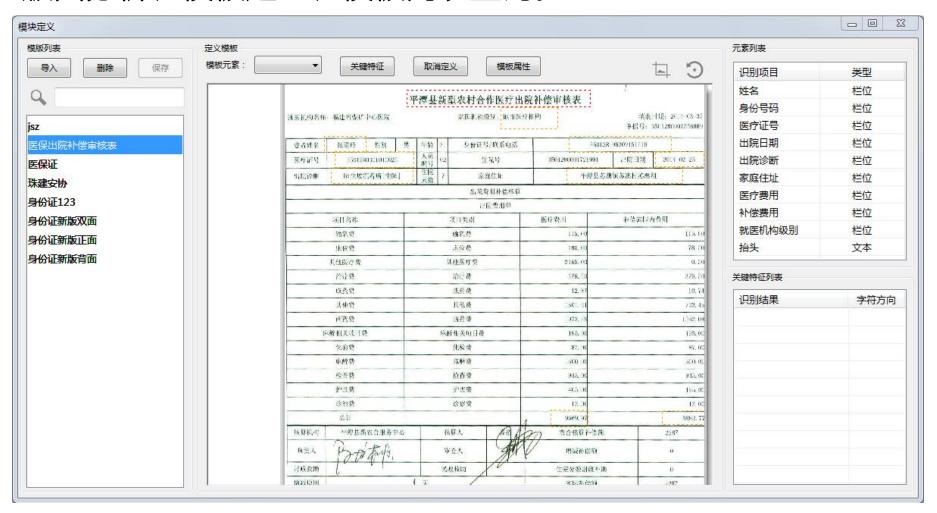


- 基于浏览器的监控和管理部署。
- 圖能够直接通过图、文、表格方式发现系统异常情况。
- ॗ能够根据预设的策略进行预警。
- ᠍ 服务器节点控制管理,进行自动部署和配置参数管理。

OCRIT

## 模板管理

## 版面分析、模板建立、模板列表查询。



OCRIT

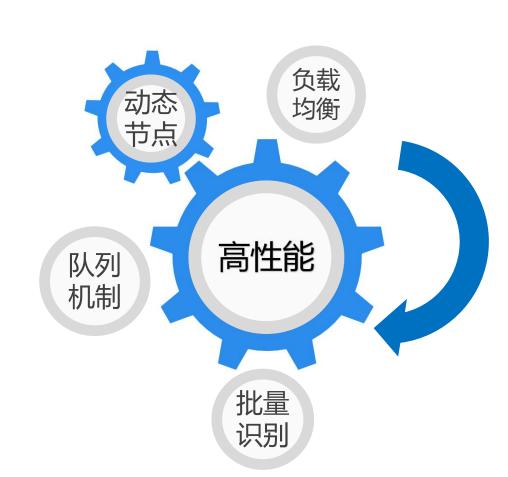
## 人工录入

## OCR 录入校对、人工定位及归类校正、手框证照区二次识别。



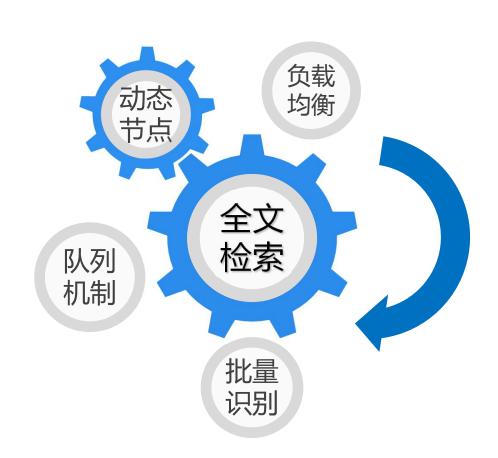
# 方案优势

#### OCR弹性高负载架构



- 客户端上传实现断点续传
- 负载均衡缓解服务器网络压力
- IO瓶颈需提升NAS性能
- 数据库瓶颈增加数据库集群节点
- 即别瓶颈扩容OCR节点数量

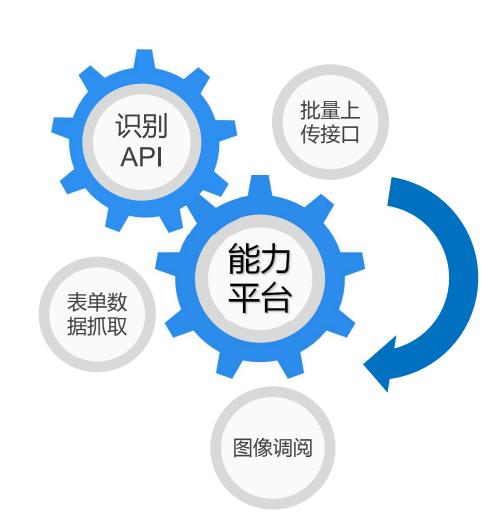
## 秒级全文检索



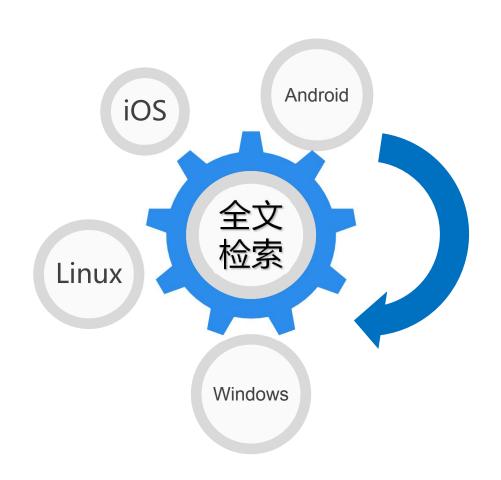
- 非结构化数据全文索引,海量资料检索秒查询
- 高亮显示关键字所在图片位置

# 方案优势

## API能力平台



- □ 批量上传
- 图像识别
- 🗎 表格数据抽取识别
- 全文检索
- 原文件图像查询调阅

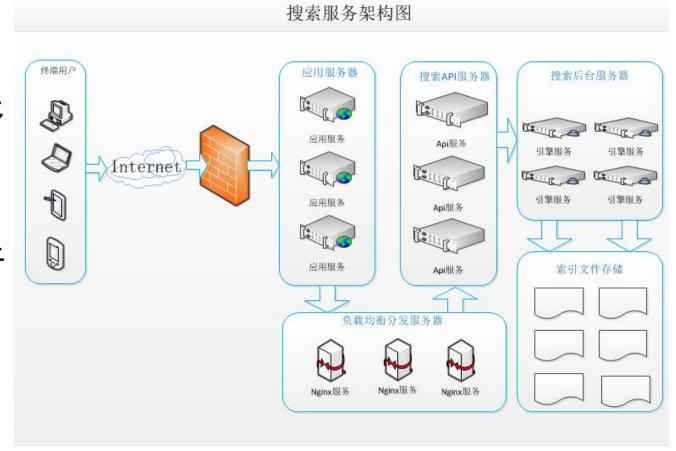


■ 支持服务端、客户端多个平台

## 方案优势

### 识别快、架构好、安全性高

- 월 1000万页的文档1秒左右反馈
- □ 支持自定义词库、自定义词组库
- 能够执行跨平台、一键式部署,多 线程处理,适应并发量
- 采用分布式、易于扩展的检索架构
- 目前该全局搜索引擎已经运用于 "平潭信用岛"项目中。



第三章 应用和案例

1 应用场景

2 案例经验

# 应用场景



数字图书馆的建设,档案 资料数字化,建立汉字文 献档案库



办公自动化中汉字文件资料自动输入,汉字图像文本的压缩存储、传输



智能全文信息管理系统、 多媒体应用系统、 自动翻译系统

## 案例经验

#### 平潭档案云项目

为平潭政府提供上亿档案影像扫描上传、海量云存储、弹性云分类云识别、秒级云检索。 为平潭信用岛等各个项目提供大数据源,实现各种证照表单分类、栏目定位、识别。

| 平潭    | "档案云  | 、"综合管理平                | 台 工作台 (             | 档案建立 | 档案管理 公文检索 报表                           | <b>查询 模板管</b> | 里 系统监控 | 系统 | <b>宣理</b>   | 欢迎:  | 批量扫描(sys)    |  |
|-------|---|------------------------|---------------------|------|--|---------------|--------|----|---|------|--------------|--|
| 类型:   | 公文  | 文号:                    | 题名:                 |      | 时间: 2000-01-01 —                       | 2050-01-01    | :宗号:   |    | 闽   |      |              |  |
| 共找到 1 | 封找到 <mark>172</mark> 条 包含 " <mark>闽岗综实管"</mark> 的数据 |                        |                     |      |  |               |        |    |   |      |              |  |
| 所     | 届部门   | 件号                     | 文号                  | 责任人  | 題名                                     | 日期            | 密级     | 页数 | 摘要  | 年度   | 操作           |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 歸省聯综合实<br>验区答顧员会文件 阃岗综实管 | 2018-02-08    | 0      | 30 | <mark>闽岚综实</mark> 管(2011) 45号 签发人: 杜源生 手灣綜合实验区管委会关于 申报 "中国服务外包示范城市" 的           | 2016 | 详情   目录   借阅 |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 論省鞴綜合实<br>验区答麟员会文件 闽岚綜实管 | 2018-02-08    | 0      | 30 | <mark>闽岚综实</mark> 管(2011) 45号 签发人: 杜源生 手潭综合实验区管委会关于 申报 "中国服务外包示范城市" 的           | 2016 | 详情   目录   借阅 |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 點省鞴綜合实<br>验区答麟员会文件 闽岚綜实管 | 2018-02-08    | 0      | 30 | <mark>闽岚综实</mark> 管(2011) 45号 签发人: 杜源生 手潭综合实验区管委会关于 由报 "中国服务外包示范城市" 的           | 2016 | 详情   目录   借阅 |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 豔省鞴綜合实<br>验区答麟员会文件 闽岚綜实管 | 2018-02-08    | 0      | 30 | 2012 ! 仃8 永久I 鼬省鞴综合实验区答麟员会文件 <mark>闽</mark><br>岚综实管(2012) 45 号 签发人 : 杜源生 平潭綜合实验 | 2016 | 详情   目录   借阅 |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 點省鞴綜合实<br>验区答麟员会文件 闽岚綜实管 | 2018-02-08    | 0      | 30 | 2012 ! 仃8 永久I 鼬省鞴綜合实验区答麟员会文件 闽<br>岚综实管(2012) 45 号 签发人 : 杜源生 平潭綜合实验              | 2016 | 详情   目录   借阅 |  |
| ×     | 办公室   | 209-WS·2016-Y-<br>0011 | 闽岚综实管(2012)<br>45 号 | 区办公室 | 2012 ! 仃8 永久I 論省鞴綜合实<br>验区答麟员会文件 闽岗綜实管 | 2018-02-08    | 0      | 30 | 2012 ! 仃8 永久I 鼬雀鞴綜合实验区答麟员会文件 闽<br>岚综实管(2012) 45 号 签发人 : 杜源生 平潭綜合实验              | 2016 | 详情   目录   借阅 |  |

#### 票据识别项目

杭州驭缘网络科技 购车发票识别定制

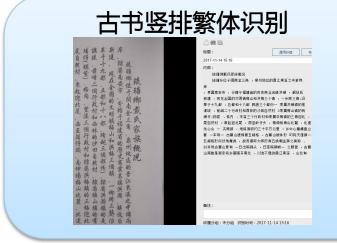
广州奔步电脑 二代身份证及银行卡识别

北京时代云海科技 购车发票识别定制

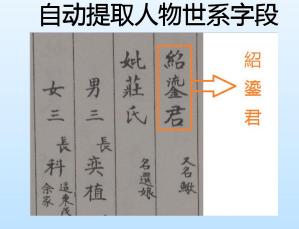


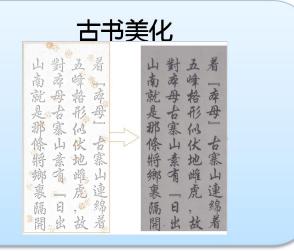
### 泉州族谱数字化项目

- ॗ族谱书籍扫描存储
- □ 古字体深度学习
- 圖影像切片安全保障
- □ 工厂坐席核对校验
- 任务优先级调度









# 第四章 企业简介

- 1 关于云脉
- 2 核心竞争力
- 3 优质客户

# 关于云脉

厦门云脉技术有限公司(原厦门思根科技有限公司)成立于2002年6月,是一家专注于多平台光学字符识别的研发与运用,同时在移动互联网技术方面处于领先地位的高新技术企业。公司凭借先进的技术实力与良好的客户关系,以OCR引擎技术为合作重点,为各行业提供以OCR技术为核心的解决方案及服务。

## \* 成功商业经验(曾合作大客户):

100多家手机及扫描器厂商(三星、索爱、ACER、ASUS、HP、MICROTEK等)

50多家电信运营商单位

100多家券商、银行等金融机构总部及其集成商

100多家政府机构及其集成商

**10多家主流互联网公司(**淘宝、支付宝、京东、网易、人人网、UBER等)

## \* 成功网络运营经验:

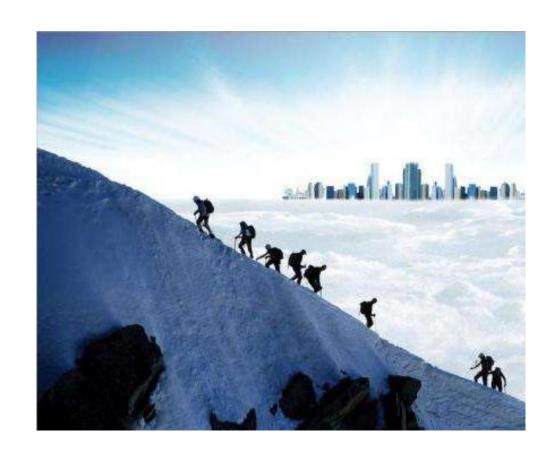
运营积累了500多万名片识别管理、文档识别管理等B2C移动互联网用户

# 核心竞争力

## 20人专业团队OCR定制能力

- 1、海内外OCR引擎专家
- 2、图像处理专家
- 3、版面智能分析专家
- 4、深度学习训练专家

云脉OCR引擎可根据行业文档的特点进行定制,通过多引擎对比识别得出结果,提升文档识别率。



# 优质客户

#### 金融行业 FINANCIAL INSTITUTIONS













#### 电信运营商 TELECOM COMPANIES







#### 互联网行业 INTERNET COMPANIES











#### 其他合作伙伴 OTHER PARTNERS



















厦门云脉技术有限公司 www.yunmai.com