A picture containing standing, woman, people

Description automatically generated



**DATABASE INSTRUCTION**

|  |  |
| --- | --- |
| Type | Image |
| Key words/ Description | **OCR image database** |

Catalog

[1. Database Introduction 3](#_Toc143158247)

[2. Data specifications 4](#_Toc143158248)

[2.1 Usage cases： 4](#_Toc143158249)

[2.2 Collecting device： 4](#_Toc143158250)

[2.3 Collecting environment： 4](#_Toc143158251)

[3. Information provided by database 6](#_Toc143158257)

# 1. Database Instruction

The dataset product can be used in many AI pilot projects and supplement production models with other data. It can improve the model performance and be cost-effectiveness. Dataset is an excellent solution when time and budget is limited. Appen database team can provide a large number of database products, such as ASR, TTS, video, text, image. At the same time, we are also constantly building new datasets to expand resources. Database team always strive to deliver as soon as possible to meet the needs of the global customers.

This OCR database consists of image data in Korean, Vietnamese, Spanish, French, Thai, Japanese, Indonesian, Tamil, and Burmese, as well as handwritten images in both Chinese and English (including annotations). On average, each image contains 30 to 40 frames, including texts in various languages, special characters, and numbers. The accuracy rate requirement is over 99% (both position and content are correct).

The images include the following categories:

* RECEIPT
* IDCARD
* TRADE
* TABLE
* WHITEBOARD
* NEWSPAPER
* THESIS
* CARD
* NOTE
* CONTRACT
* BOOKCONTENT
* HANDWRITING

# 2. Data Specification

|  |  |
| --- | --- |
| **Usage Cases** | **Image label recognition training** |
| **Collecting device** | **Mobile phone / Camera** |
| **Collecting environment** | **Multiple lights environments** |

|  |  |  |
| --- | --- | --- |
| **Database Name** | **Category** | **Quantity** |
| Korean Document OCR Images | RECEIPT | 1500 |
| IDCARD | 500 |
| TRADE | 1012 |
| TABLE | 512 |
| WHITEBOARD | 500 |
| NEWSPAPER | 500 |
| THESIS | 500 |
| CARD | 500 |
| NOTE | 499 |
| CONTRACT | 501 |
| BOOKCONTENT | 500 |
|  | TOTAL | 7,024 |
| Vietnamese Document OCR Images | RECEIPT | 337 |
| IDCARD | 100 |
| TRADE | 227 |
| TABLE | 100 |
| WHITEBOARD | 111 |
| NEWSPAPER | 100 |
| THESIS | 100 |
| CARD | 100 |
| NOTE | 100 |
| CONTRACT | 105 |
| BOOKCONTENT | 700 |
|  | TOTAL | 2,080 |
| Spanish Document OCR Images | RECEIPT | 1500 |
| IDCARD | 500 |
| TRADE | 1000 |
| TABLE | 500 |
| WHITEBOARD | 500 |
| NEWSPAPER | 500 |
| THESIS | 500 |
| CARD | 500 |
| NOTE | 500 |
| CONTRACT | 500 |
| BOOKCONTENT | 500 |
|  | TOTAL | 7000 |
| French Document OCR Images | RECEIPT | 300 |
| IDCARD | 100 |
| TRADE | 200 |
| TABLE | 100 |
| WHITEBOARD | 100 |
| NEWSPAPER | 100 |
| THESIS | 103 |
| CARD | 100 |
| NOTE | 100 |
| CONTRACT | 100 |
| BOOKCONTENT | 700 |
|  | TOTAL | 2003 |
| Thai Document OCR Images | RECEIPT | 1500 |
| IDCARD | 500 |
| TRADE | 1000 |
| TABLE | 537 |
| WHITEBOARD | 500 |
| NEWSPAPER | 500 |
| THESIS | 500 |
| CARD | 500 |
| NOTE | 500 |
| CONTRACT | 500 |
| BOOKCONTENT | 500 |
|  | TOTAL | 7037 |
| Japanese Document OCR Images | RECEIPT | 1586 |
| IDCARD | 500 |
| TRADE | 1000 |
| TABLE | 552 |
| WHITEBOARD | 500 |
| NEWSPAPER | 500 |
| THESIS | 509 |
| CARD | 500 |
| NOTE | 500 |
| CONTRACT | 500 |
| BOOKCONTENT | 500 |
|  | TOTAL | 7147 |
| Indonesian Document OCR Images | RECEIPT | 1500 |
| IDCARD | 500 |
| TRADE | 1003 |
| TABLE | 500 |
| WHITEBOARD | 501 |
| NEWSPAPER | 502 |
| THESIS | 500 |
| CARD | 500 |
| NOTE | 500 |
| CONTRACT | 500 |
| BOOKCONTENT | 500 |
|  | TOTAL | 7006 |
| Tamil Document OCR Images | RECEIPT | 356 |
| IDCARD | 98 |
| TRADE | 475 |
| TABLE | 532 |
| WHITEBOARD | 501 |
| NEWSPAPER | 500 |
| THESIS | 500 |
| CARD | 500 |
| NOTE | 501 |
| CONTRACT | 500 |
| BOOKCONTENT | 500 |
|  | TOTAL | 4963 |
| Burmese Document OCR Images | RECEIPT | 300 |
| IDCARD | 100 |
| TRADE | 200 |
| TABLE | 117 |
| WHITEBOARD | 110 |
| NEWSPAPER | 108 |
| THESIS | 102 |
| CARD | 100 |
| NOTE | 120 |
| CONTRACT | 100 |
| BOOKCONTENT | 761 |
|  | TOTAL | 2118 |
| English Handwritten Datasets | HANDWRITING | 2278 |
| Chinese Handwritten Datasets | HANDWRITING | 11118 |

# 3. Information provided by database

* Data Format：. JPG

A picture containing standing, woman, people

Description automatically generated

**数据库说明书**

|  |  |
| --- | --- |
| 类型 | 图片 |
| 数据库名称 | **OCR图片数据库** |

目录

[1.数据库简介 10](#_Toc143180964)

[2.数据规格 11](#_Toc143180965)

[2.1常见使用案例： 11](#_Toc143180966)

[2.2采集设备： 11](#_Toc143180967)

[2.3采集环境： 11](#_Toc143180968)

[3. 数据库提供内容 15](#_Toc143180972)

# 1.数据库简介

成品数据集可以运行许多 AI 试点项目，并用其他数据补充生产模型，从而提高整体性能和成本效益。无论企业规模如何，在时间和预算紧张时，选择成品数据集都是绝佳的解决方案。Appen数据集产品涵盖多种类型的数据，包括图像、视频、音频和文本。我们也在不断构建新的数据集，以满足全球客户群的需求。​

该OCR数据库为韩语、越南语、西班牙语、法语、泰语、日语、印尼语、泰米尔语、缅甸语文档类图片及中英文手写图片数据（含标注），每张图片平均30-40框，包含各语种文字、特殊字符及数字，准确率要求99%以上（位置和内容均正确）。

图片包括以下类别 ：

* 收据
* 身份证
* 贸易
* 表格
* 白板
* 报纸
* 论文
* 卡片
* 便签
* 合同
* 书籍目录
* 手写

# 2.数据规格

|  |  |
| --- | --- |
| **常见使用场景** | **图片标签识别训练** |
| **采集设备** | **手机/照相机** |
| **采集环境** | **多种光线** |
| **单元/张** | **59774** |

|  |  |  |
| --- | --- | --- |
| **数据库名称** | **分类** | **数量** |
| 韩语文档类OCR图片数据 | 收据 | 1500 |
| 身份证 | 500 |
| 贸易 | 1012 |
| 表格 | 512 |
| 白板 | 500 |
| 报纸 | 500 |
| 论文 | 500 |
| 卡片 | 500 |
| 便签 | 499 |
| 合同 | 501 |
| 书籍目录 | 500 |
|  | 总计 | 7024 |
| 越南语文档类OCR图片数据 | 收据 | 337 |
| 身份证 | 100 |
| 贸易 | 227 |
| 表格 | 100 |
| 白板 | 111 |
| 报纸 | 100 |
| 论文 | 100 |
| 卡片 | 100 |
| 便签 | 100 |
| 合同 | 105 |
| 书籍目录 | 700 |
|  | 总计 | 2080 |
| 西班牙语文档类OCR图片数据 | 收据 | 1500 |
| 身份证 | 500 |
| 贸易 | 1000 |
| 表格 | 500 |
| 白板 | 500 |
| 报纸 | 500 |
| 论文 | 500 |
| 卡片 | 500 |
| 便签 | 500 |
| 合同 | 500 |
| 书籍目录 | 500 |
|  | 总计 | 7000 |
| 法语文档类OCR图片数据 | 收据 | 300 |
| 身份证 | 100 |
| 贸易 | 200 |
| 表格 | 100 |
| 白板 | 100 |
| 报纸 | 100 |
| 论文 | 103 |
| 卡片 | 100 |
| 便签 | 100 |
| 合同 | 100 |
| 书籍目录 | 700 |
|  | 总计 | 2003 |
| 泰语文档类OCR图片数据 | 收据 | 1500 |
| 身份证 | 500 |
| 贸易 | 1000 |
| 表格 | 537 |
| 白板 | 500 |
| 报纸 | 500 |
| 论文 | 500 |
| 卡片 | 500 |
| 便签 | 500 |
| 合同 | 500 |
| 书籍目录 | 500 |
|  | 总计 | 7037 |
| 日语文档类OCR图片数据 | 收据 | 1586 |
| 身份证 | 500 |
| 贸易 | 1000 |
| 表格 | 552 |
| 白板 | 500 |
| 报纸 | 500 |
| 论文 | 509 |
| 卡片 | 500 |
| 便签 | 500 |
| 合同 | 500 |
| 书籍目录 | 500 |
|  | 总计 | 7147 |
| 印尼语文档类OCR图片数据 | 收据 | 1500 |
| 身份证 | 500 |
| 贸易 | 1003 |
| 表格 | 500 |
| 白板 | 501 |
| 报纸 | 502 |
| 论文 | 500 |
| 卡片 | 500 |
| 便签 | 500 |
| 合同 | 500 |
| 书籍目录 | 500 |
|  | 总计 | 7006 |
| 泰米尔语文档类OCR图片数据 | 收据 | 356 |
| 身份证 | 98 |
| 贸易 | 475 |
| 表格 | 532 |
| 白板 | 501 |
| 报纸 | 500 |
| 论文 | 500 |
| 卡片 | 500 |
| 便签 | 501 |
| 合同 | 500 |
| 书籍目录 | 500 |
|  | 总计 | 4963 |
| 缅甸文档类OCR图片数据 | 收据 | 300 |
| 身份证 | 100 |
| 贸易 | 200 |
| 表格 | 117 |
| 白板 | 110 |
| 报纸 | 108 |
| 论文 | 102 |
| 卡片 | 100 |
| 便签 | 120 |
| 合同 | 100 |
| 书籍目录 | 761 |
|  | 总计 | 2118 |
| 英文手写OCR图片数据 | 手写 | 2278 |
| 中文手写OCR图片数据 | 手写 | 11118 |

# 3. 数据库提供内容

* 数据格式：.JPG