Leverage the power of technology to extract crucial information from structured and unstructured documents

Intelligent Text Extractor

Marlabs

Contents

Introduction

Intelligent Text Extractor (ITE) is a computer vision-based solution for OCR using computer vision & deep learning algorithms. The solution helps recognize and extract printed and handwritten text from scanned documents. It can recognize structured, semi-structured and unstructured documents and can automatically classify different page formats. Moreover, it supports documents in multiple languages and input formats such as images and pdf documents.

The advantage of ITE is its flexibility to choose the entire recognition, extraction, and validation workflow solution or just a REST API based recognition while providing for extraction integration to existing application. The higher accuracy levels on extractions by use of ensemble of recognition algorithms, and validation logic to check type of text and its format is another unique feature built not this solution.

How ITE Works?

ITE provides a system for efficient recognition of image and PDF files using AI and computer vision and converting them into editable and searchable document format. It also provides customized extraction for priority information based on users' requirement, using location data, thereby eliminating the redundant information from the extracted document.

Image and PDF files comprising of elements such as handwritten, hand-printed and typed text, images, tables, paragraphs and other textual or non-textual properties are loaded in batches and fed into two levels of the pre-processing engine wherein the low-quality images are treated for any noise or deformities to enhance image quality. Computer vision algorithms are applied to dynamically identify all image properties and their locations and capture them. Upon extracting the image data, the system determines the context of image data for one or more attributes to classify images into templates based on characteristics corresponding to the attributes. This enables the classification of identical documents for ease in document management and improves the further scope for archive and maintenance. The resulting data is stored in JSON format and is prepared to be exported in a preferable format. Recognized images are further provided through a workflow management system for automatic assignment of recognized documents to users for one or more levels of manual validation based on custom rules.

Additionally, validation of extracted data could be done based on the confidence score. ITE provides a dedicated solution to reliable and efficient document digitization, management and maintenance, facilitating the convenience of users.

Features

- Higher Accuracy- Reliably extract Printed, hand printed & hand-written text using ensemble of algorithms
- No Upfront Template Configuration- Detect texts from unfamiliar documents without any configuration
- Auto Classification- Classify similar documents as same template
- Real Time- Recognize & extract useful data in seconds, more time for high-value work
- Integrated workflow- Seamless text extraction, validation & export for document data
- Support all popular file formats JPG, JPEG, PNG and PDF and languages

Offerings

- Web based application on mAdvisor with interface for manual validation
- REST based endpoint- Easily integrate ITE API to offer OCR capabilities
- UI Path Component- Integrate OCR capabilities within RPA workflows
- SaaS offer on Azure marketplace

Applications

1. Invoices
2. All kinds of forms- printed, hand-printed & hand-written
3. Marksheets & Transcripts
4. Credit reports
5. Contracts & Agreements
6. Financial forms and bank statements
7. Medical reports & Prescriptions
8. Timesheets