OCR

Optical Character Recognition

Kaivan Gandhi 60001160012 Rahul Jha 60001160019 Shagun vasmatkar 60001160061

Introduction

- Nowadays, a lot of paper documents are transformed to electronic form, which makes information processing easier, like searching, analysis and conversion.
- Many companies and other institutions decide to digitalize their documents. Working with files is cheaper than processing traditional documents, because there is no space required for document storage.
- There are three main steps of document digitalization: scanning, indexation (data entry) and presentation of digitalized documents.
- Researchers proved that the recognition of both barcodes and printed text through Optical Character Recognition or OCR is reliable and significantly accelerates data processing.
- On the contrary, the handwritten text appeared difficult to recognize by OCR systems.

OCR

- Optical Character Recognition or OCR is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the form.
- It is the mechanical or electronic conversion of scanned or photographed images of typewritten or printed text into machine-encoded/computer readable text.
- OCR is a field of research in pattern recognition, artificial intelligence and computer vision. It is the electronic translation of handwritten, typewritten or printed text into machine translated images.

Aim and objectives

- Stage 1
- Recognition of printed data and tables from various documents. Also conversion into text files and table generation as per the need
- Stage 2
- Recognition of handwritten text using neural networks and using in various application such as handwritten forms, attendance sheets(for automatic data generation and adding data in excel sheet) and also in various database/medical applications

Procedure

1.Pre-processing:

- grayscale conversion
- noise removal(filtering)
- binarization, skew-correction.

2. Feature extraction

3. Fetching the extracted feature(data) to Neural Network and training it to recognize character

PYTHON libraries

- PIL:

Python Imaging Library (abbreviated as **PIL**) (in newer versions known as Pillow) is a free library for the Python programming language that adds support for opening, manipulating, and saving many different image file format

- Pytesseract:

Python-tesseract is an optical character recognition (OCR) tool for python

- For mathematical and numerical calculations

NUMPY

MATH

PANDAS(data extraction and manipulation)

Applications

- Data entry for business documents, e.g. check, passport, invoice, bank statement and receipt
- Automatic number plate recognition ▫ Automatic insurance documents key information extraction
- Extracting business card information into a contact list
- More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg
- Converting handwriting in real time to control a computer (pen computing)
- Assistive technology for blind and visually impaired user