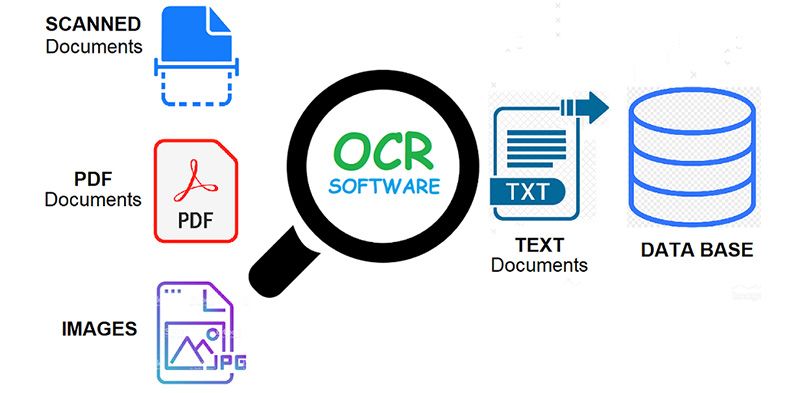
**Invoice Recognition**

**Introduction**

OCR = Optical Character Recognition. In other words, OCR systems transform a two-dimensional image of text, that could contain machine printed or handwritten text from its image representation into machine-readable text. OCR as a process generally consists of several sub-processes to perform as accurately as possible. The subprocesses are:

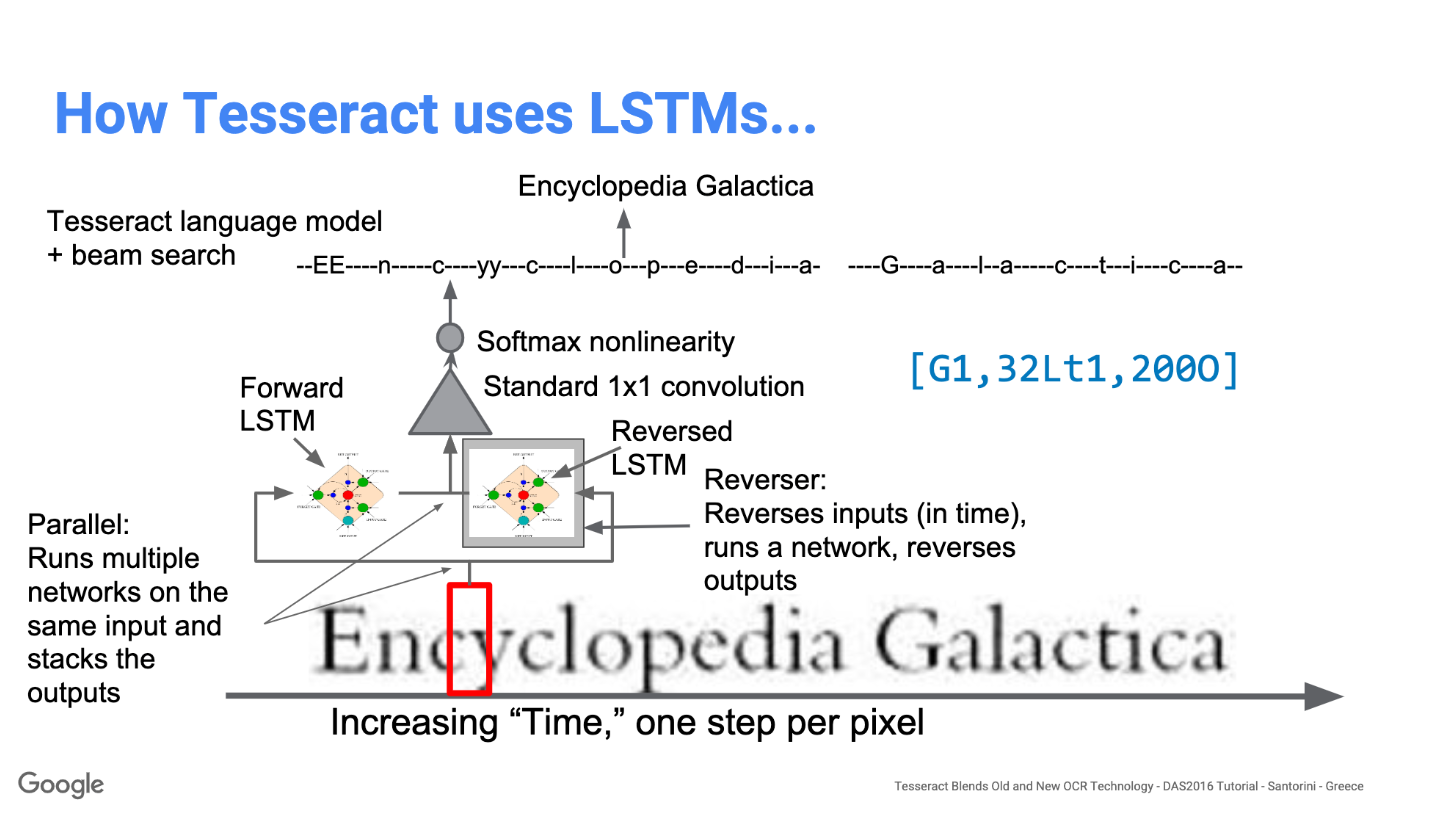
* Preprocessing of the Image
* Text Localization
* Character Segmentation
* Character Recognition
* Post Processing

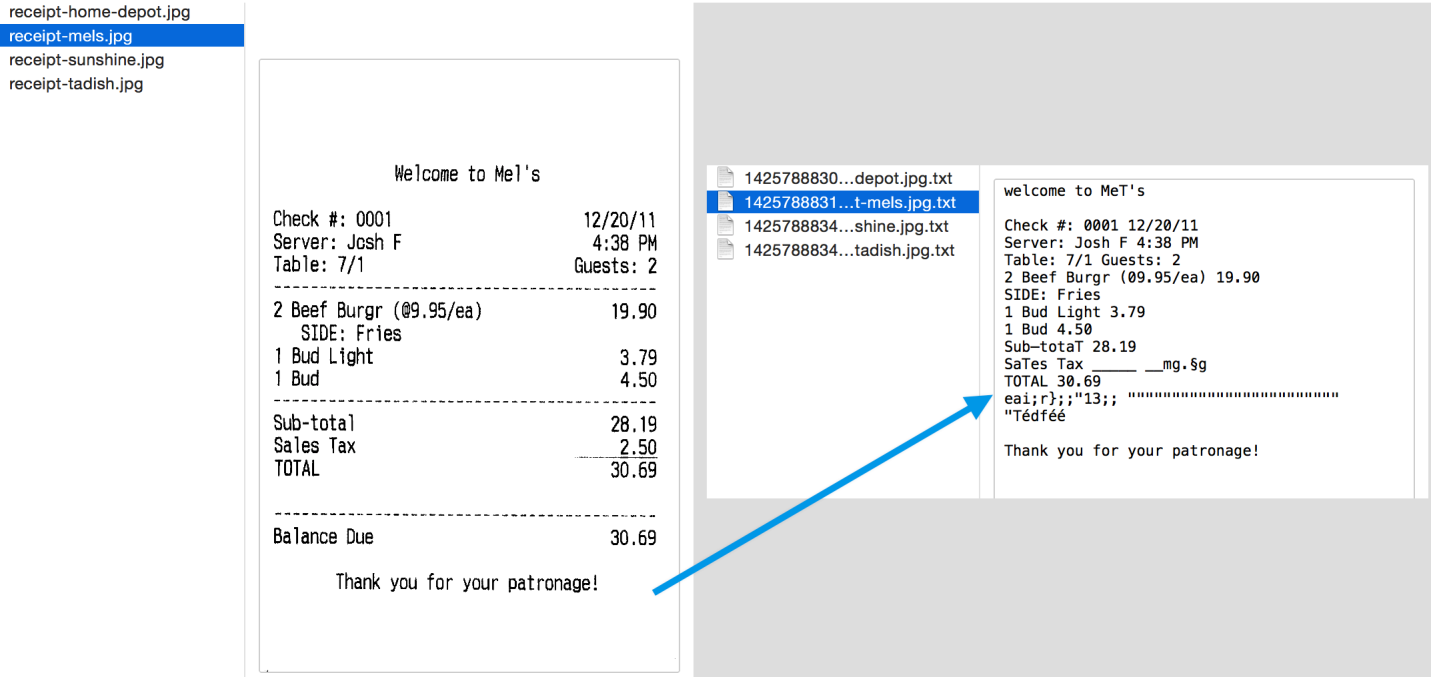


Optical Character Recognition process

## Tesseract OCR

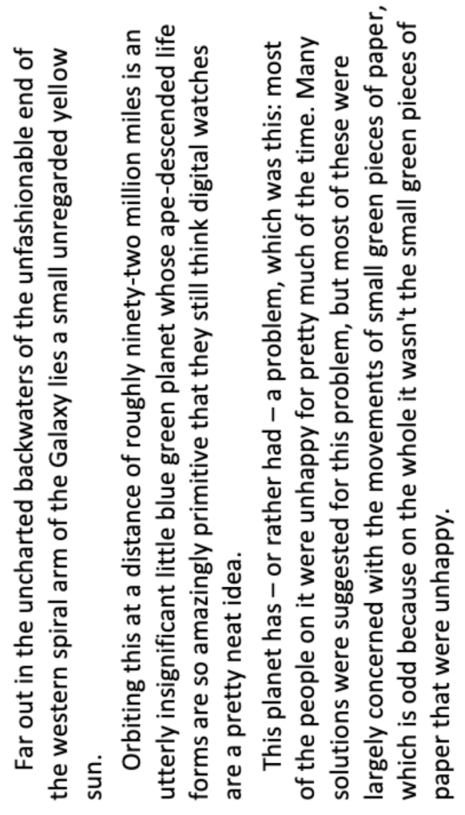
Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images. It supports a wide variety of languages. Tesseract doesn't have a built-in GUI, but there are several available from the [3rdParty page](https://github.com/tesseract-ocr/tesseract/wiki/User-Projects-%E2%80%93-3rdParty). Tesseract is compatible with many programming languages and frameworks through wrappers that can be found [here](https://github.com/tesseract-ocr/tesseract/wiki/AddOns). It can be used with the existing layout analysis to recognize text within a large document, or it can be used in conjunction with an external text detector to recognize text from an image of a single text line.



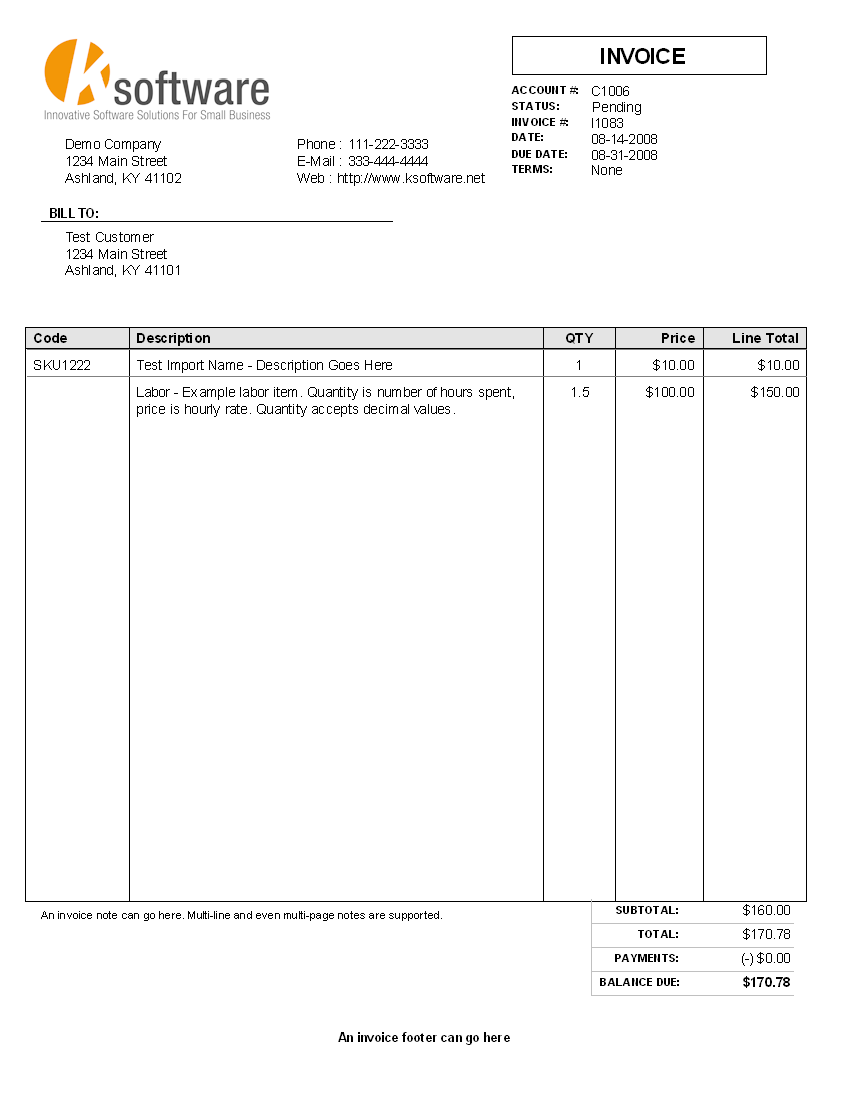


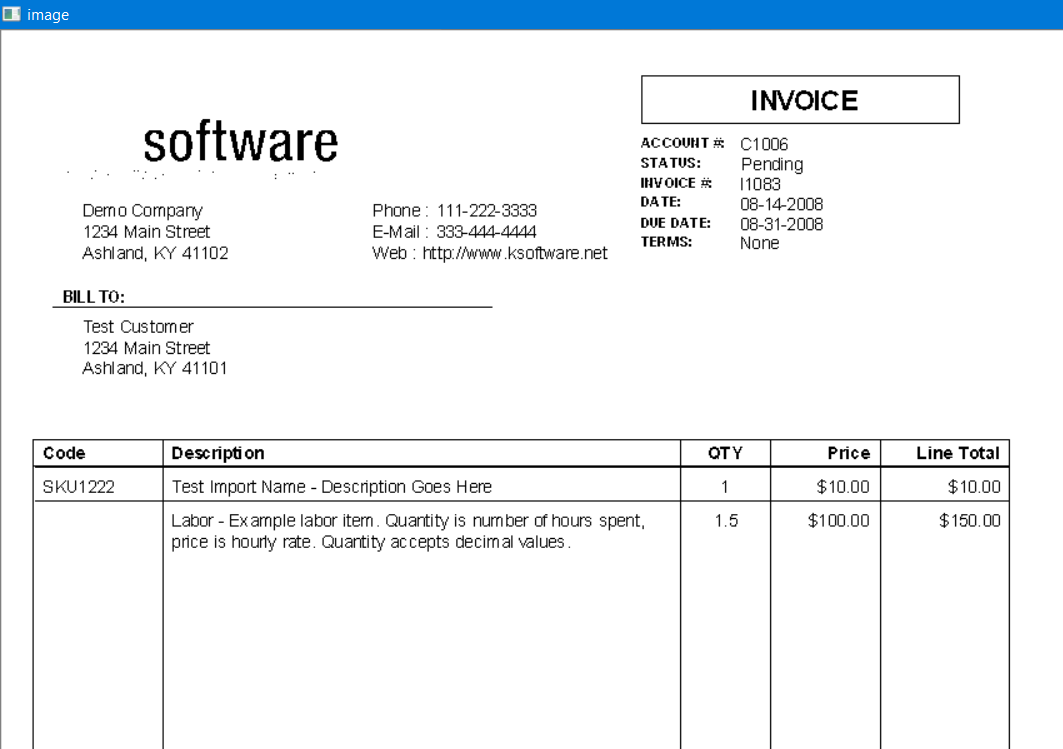
Result of the Tesseract OCR engine

You can detect the orientation of text in your image and also the script in which it is written. The following image -



Preprocessing of the Image





**OUTPUT – EXCEL FILE**

