Mariner Software

What are the factors that affect the accuracy of OCR?

Jim Henson - 2018-12-17 - OCR

Our OCR software attempts to read the receipt, amount date etc. Depending on the quality of the scan and the legibility of the actual receipt we will import it. Some scanners do better than others, some embed text information into a PDF (called text over image). We do not run the OCR software if there is embedded text.

Optical Character Recognition (OCR) is a process by which an image is translated into to computer-intelligible text. This article outlines several factors affecting the accuracy of this process.

OCR process:

1. The application **acquires an image** of an original document and does some **image pre-processing**.
2. The image goes through **OCR processing**:

- The OCR engine scans the image and finds the edges, font, and paragraphs and individual characters.
- The OCR engine determines whether each character and/or word, matches characters it is ‘trained’ to recognize.

3. The application **perfoms postprocessing** on the OCR output. Fixing typos and common misspellings.

For a more-detailed explanation of how OCR works, see this knowledge base article.

This article refers to **OCR fidelity**. This describes the accuracy of the OCR output to the visual input originally provided to the OCR engine (the "picture" that the OCR engine attempts to recognize text characters from).

1. The Quality of the Original Document
Receipts often are priting on thermal paper by a low qualtiy printer. That receipt might get wrinkled or folded and those things will play a factor in the quality of thescanned image. Try to use the highest quality document possible. If an original is of low quality--for example, the ink is too light, the paper is not flat and white (or the text otherwise does not contrast highly with the background)--an OCR engine will have a difficult time discerning the text

from any noise surrounding it.

Similarly, the quality of a document acquired by a computer for OCR will have an effect on the quality of OCR output. If the original document is:

- wrinkled, torn, or otherwise **damaged**,
- faded or otherwise **aged**,
- **discolored**,
- smudged (or **the text** is otherwise **obfuscated or distorted**),
- **printed with low-contrast or colored ink** (purple, blue, and red provide low contrast; black ink provides highest contrast),
- rendered with **nonstandard fonts or in human handwriting**,
- or printed on specific types of **paper that decrease crispness and contrast between the background and foreground** in the resulting scan,

...any scanned image of such a document (regardless of the quality of the scan) might provide extra burden to the OCR engine in recognizing text from the scan.

2. The Quality of the Scan
Scanners make a digital representation of visual input. The quality and type of output it is able to produce, will have an effect on the quality of the input provided to the OCR engine to process.

Software programs (like the Image Capture dialog in Paperless) make it possible for the user to alter the visual properties (such as brightness and contrast) of the scanner's output. One of the biggest factors is DPI or Dots per Inch. Setting the DPI lower than 200 will yield uninteligible results wheras setting it higher than 600dpi will just increase the size of the stored file without yielding much better results. We tend to recommend a 300dpi for in item.

Color vs. Black and White, vs. grayscale depends on your source material. Often times documents and receipts will not need to be stored in Color (as it makes the storage file larger) but the difference between B&W and grayscale can be profound depending on the source. In general, a grayscale image will provide a better result than a black and white image. Internally, the software performs adaptive binarization where it changes a grayscale image to a black and white image if this process is already done, presumably by your scanner software) it's possible some fidelity will be lost and the OCR results may suffer.

3. The OCR Engine
The OCR engine is the computer process that translates images to characters recognizable to the computer. For more informations on OCR engines and the OCR Process, see this Knowledge Base article.

Many factors within the OCR engine software can contribute the fidelity of the text the OCR

engine produces to the visual input the engine was originally tasked with processing.

OCR engines are provided with instructions (generally by a programmer) that provide the OCR engine with instructions on how to discern the correct letter from a set of possibilities, based on the image presented to the OCR engine. The quality of these instructions has some bearing on the fidelity of OCR output.

4. Auto-matching

Paperless attempts to match the merchant from a list of merchants that have been imported (a list of several thousand merchants is provided) or previously entered. If the item you are scanning is a place you've never been before and is not a major retailer, Papelress has zero chance of auotmatically figuring out the merchant name. The fields that are matcheda are: Merchant, date, amount and tax. Category is related to the merchant name and only if there is a merchant match will the cateogry for that merchant be used.