intex Informations-Systeme

OCR - 5 hints to increase the recognition rate

OCR - The bridge between paper and digital form

In everyday business, OCR (Optical Character Recognition) is often one of the first steps in the digitization and processing of documents. A typical use case is to read the contents of incoming (paper) invoices in order to automatically transfer them to an ERP system and process them there.

Even today, this procedure is still widespread, despite various possibilities for digital invoice exchange: Almost half of our customers (47%) still receive 70% or more of their incoming invoices in paper form today.

OCR is used to bridge the media gap between paper and digital form. The following steps are performed in detail: Although documents are already available digitally as image or PDF files as a result of scanning, the contents - let alone their significance - are still completely unknown to the target system.

At this point, the OCR. Your first task is to recognize shapes, patterns, letters or numbers in these PDF or image files. In the second step, the OCR interprets what it has recognized and tries to form words or values from it. The third step is that the OCR establishes a connection between the individual information it has found in this way, e.g. that the element "invoice number" and the associated invoice number are logically linked.

This description of what the OCR does in detail makes it clear what the challenges are:
How successful the recognition is depends strongly on the quality of the input. The following interpretation again depends on the recognition as well as on the
of the help or guidance that the OCR gets.

In the following, we would like to share with you our experience from over 1,000 successfully implemented digitization projects with customers and show you which - often very simple - measures you can use to significantly improve the recognition rate and thus the results of your OCR.

Initial situation: The typical pitfalls during document reading

The OCR can only be as good as the documents it processes. You should keep that in mind: Anything that makes it difficult for the human eye to read a document is also a hindrance to OCR.

Document quality and scanning: Preparatory work for the OCR

Before the OCR can begin its work, the paper documents must be scanned. The better the scan result, the easier it is for OCR. For the best possible scan result, you should make sure that the paper documents have no creases, folding marks or dog-ears. Stains, handwritten notes and stamps also have a negative influence on legibility. Before scanning, you should also remove staples and sticky notes and center smaller documents. It is advisable to make foreign barcodes unrecognizable. A too low resolution of the file output is also disadvantageous for the following process steps.

Business documents: visually appealing or legible?

Today, most companies adapt their business documents to their corporate design. What looks appealing to the human eye, OCR sometimes makes the work more difficult. Too small or too large fonts, playful or ornate fonts and insufficient contrasts (e.g. grey fonts on recycled paper) are obstructive. Shading, background images, watermarks and double-sided printing on thin paper (shimmering through) are also disadvantageous for OCR. Also an unconventional arrangement of the contents (e.g. header data not in the letterhead) can impair the results of the OCR.

The big hurdle: Comparison with the master data

Even if the OCR has correctly recognized all contents of a document and interpreted them as desired, there may be no result. The last hurdle has not yet been overcome - namely the comparison with the data in the target system. For example, if a vendor is created multiple times in your ERP system, the OCR cannot know which record an invoice should be assigned to, even if it has recognized all the data correctly.

The reasons for duplicates in the ERP system are manifold: different spellings or abbreviations of the company name, a move of the vendor or different locations, new contact persons at the supplier, a changed company name and new employees in your company are only some of them.

Solution: Our 5 tips to improve the results of OCR

If we look at the pitfalls during document reading, the approaches to optimization also become clear. Below are some specific tips on what you can do to improve the results of OCR.

01. pay attention to the document quality

You should treat incoming paper documents with care or scan them as early as possible. Do not apply stamps, handwritten notes, or sticky notes before scanning. Remove parentheses.

The typical traces of use that occur when a paper document makes its way across desks also worsen the scan results. To avoid this, we recommend a central inbox where documents are scanned directly before they are processed.

When scanning, you should also make sure that the settings have a sufficient resolution, we recommend 300 dpi.

02. agree the rules of the game with your suppliers

Ask your suppliers to pay attention to a few points so that invoices can be processed more quickly and with fewer queries in the interest of both parties.

Header data or payment information are often displayed more cautiously, e.g. through a smaller font size or a less contrasting font colour. However, they are the most important information for incoming invoice processing. Ask your suppliers to include these in the invoice in a large and legible form as well.

Shades, background colors, watermarks and unusual fonts make a document look high-quality and individual - but make it difficult to read. Advise the biller to use a font without serifs and to avoid graphic elements. Ask your suppliers to send invoices to a central inbox rather than to individual employees so that receipts can be scanned directly and detours can be avoided.

In the case of multi-page receipts, your supplier should refrain from stapling them. It also simplifies handling if the invoice is not accompanied by further documents (general terms and conditions, offer, product information, contract documents, etc.).

Tip: Your suppliers may also be prepared to include further information that is useful especially for you in the invoice, which is not covered by § 14
UStG are covered. This could include the purchase order number for invoices with purchase order reference or your cost center.

03. specify the frame for the OCR

A simple way to improve the results of OCR is to do some preparatory work. By presorting documents, you may have one more manual step at the beginning of the process, but you can increase the recognition rate and avoid time-consuming, subsequent cancellations of invoices in your ERP system.

On the one hand, you should make sure that only invoices are included in the OCR process and not other documents such as application documents or advertising letters. On the other hand, it also makes sense to pre-sort by company code, since, for example, specifying an incorrect company name is a typical error on invoices and the subsequent cancellation of the document in the workflow is possible, but involves more effort.

It is also advisable to separate the individual invoices by applying a barcode. Otherwise the OCR does not know with 100% certainty where one invoice ends and the next begins, or can filter out only with difficulty independently attachments (like AGB, delivery notes, etc.). It is also possible to separate or merge documents and delete pages in the scan client at a later stage, but
it's easier to do this before.

04. Clean up the master data in the target system.

Duplicates, e.g. creditors created several times in the ERP system, are confusing and hindering in various work processes. In the context of OCR, they prevent information from being correctly attributed. It is therefore advisable to clean up the master data in order to improve the results of the OCR.

05. Perfect Page

In addition to the measures mentioned above, our solutions support you with various functions to improve the recognition rate of OCR. We would like to introduce these options, features and settings to you:

Document scanning is a cornerstone of digital transformation, and choosing the right technology is essential to achieving the desired results. The image enhancement capabilities of a scanner can dramatically reduce the time spent on documents and provide more accurate information to the automated workflow. Many companies overlook the bottlenecks that can arise around image quality, but investing in the right technology right from the start can pay off immediately.
Perfect Page technology offers state-of-the-art image enhancement capabilities, even for very difficult documents and mixed document batches.

Slope correction and automatic cutting:
Letters seldom come as clearly arranged stacks, especially if they are documents of different sizes. The images must be aligned so that they can be used for subsequent processes such as automatic recognition of text, handwriting, or checkmarks.

Automatic alignment:
The ability to place multiple documents in landscape and portrait format in the scanner and still get a perfectly aligned series of images is a huge efficiency gain.

Automatic brightness:
No copier is required to adjust the brightness of a low-contrast document. The scanner technology automatically adjusts the image brightness optimally. This is done without any loss of speed or throughput for color and grayscale documents. By making the brightest color of each image as bright as possible and the darkest color as dark as possible, both image quality and human readability are greatly improved. This function is particularly useful for archiving documents.

Sharpening:
By increasing the contrast of the edges in an image, automatic sharpening makes objects in the image appear "clearer". This improves the quality of the document for a higher character recognition rate (OCR).

Intelligent smoothing of the background color:

Background colors in both color and grayscale images can be uneven. Image Smoothing minimizes color variations and provides a "cleaner" image that looks more like a digitally created document. Usually this also reduces the size of compressed images. Documents or forms whose foreground (e.g., text, lines, etc.) needs to be highlighted more clearly can be enhanced with intelligent foreground saturation.

Strip reduction:

Strip filter technology is one of the most common problems with regard to image quality. It takes care of vertical black lines on a non-aligned template. These lines are often caused by dust deposits in the scanner housing. This can be prevented by regular cleaning, but image enhancement technology can also remove it or at least reduce it should it happen.

Noise reduction:

Another known challenge when converting color documents to black and white is the appearance of small (or larger) dots, also called "noise", caused by dust or lower quality paper (such as recycled paper). Noise reduction algorithms remove individual points (individual pixels), pixel groups (majority rule) or larger pixel groups (background noise reduction) in order to better display the document.

Increased OCR read rates and the binarisation process "iThreshholding":

Binarization, e.g. the conversion of color images into black and white images, is the core of all data extraction functions and therefore the basis for all image processing. In a stack of documents of varying quality, the requirements for binarization and upgrading vary from one document to another. Intelligent technology analyzes both the foreground and background of documents, determines brightness and contrast and then dynamically determines the optimum thresholds. This optimizes the overall image quality and file size, especially for dark documents, which often have problems with character recognition.

The Perfect Page technology allows you:

Restriction of document preparation to simple things like removing staples, since absolutely no presorting is necessary anymore.
Process mixed stacks of documents without prior sorting them by orientation, size, type, or shape.
Significantly better text recognition results (OCR/ICR) and less effort with exceptions.