WHITE PAPER

The Art and Science of Kofax OmniPage OCR

Understanding the Computer Vision and Artificial Intelligence inside the OmniPage OCR engine

KOFAX

Optical Character Recognition

Optical character recognition—or as it is often referred, "OCR"—is the process of converting images of text, tables and pictures into digitally encoded documents that can be used in computer systems. This process uses computer vision (CV) and artificial intelligence (AI) algorithms.

This white paper provides an overview of how Kofax OmniPage™, the industry-leading optical character recognition (OCR) software, delivers fast, easy and accurate document conversion, instantly turning paper and digital documents into files that can be edited, searched and shared securely.

Image Pre-processing

Before an image can be accurately "OCRed," the image should be enhanced to increase the extraction rate. This is called pre-processing. Image pre-processing attempts to enhance a base image into the best version of itself so that OCR can produce the most accurate representation of the document. The OmniPage Capture SDK contains image pre-processing algorithms that have been fine-tuned over the past 20 years to achieve the highest-quality results.

There are four primary steps in image enhancement that are performed automatically during pre-processing:

Resolution enhancement – OCR works best on 300 DPI (Dots Per Inch) images. Any images significantly outside of 200-300 DPI are resampled to fit inside the desired range. This could mean down-sampling from 600 DPI images to 300 DPI or up-sampling 100 DPI images to 200 DPI.

Binarization – OCR is designed to take place on black and white images. Any color or grayscale images are converted to black and white using an adaptive binarization algorithm. This algorithm compares pixel intensities in an area to determine what threshold to use to convert pixels to either black or white.

Auto-rotation and deskewing – OmniPage uses algorithms that look for lines of text and the character patterns to both automatically deskew and rotate input images to the correct orientation.

Adaptive noise removal – Images are automatically analyzed for the presence and quantity of noise. These are the tiny "specks" often found on scanned images. An adaptive algorithm is then used to remove noise based on the size of the noise particles found on the image.

Layout Analysis and Zoning

Before OmniPage can begin to recognize text, the pre-processed input image must first be broken up into different logical zones. This is a two-step process:

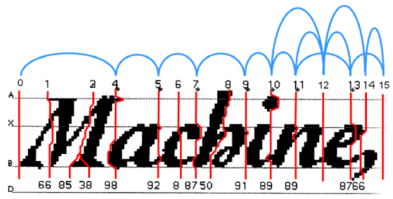
1. Locate text and pictures within the image – Using both the whitespace of the image and the patterns of the pixel shapes, text regions are segregated from images and partitioned into zones that represent columns, paragraphs and text blocks. Images and other non-text pixels are also identified so they can be omitted during the text recognition phase, but included in the final output.

2. Detect tables – Once the text regions have been identified, OmniPage looks for the presence of grid lines and structured text blocks. Using this information, OmniPage can flag text zones as tables; this allows the output to provide the same look and feel as the original document.

Text Recognition

Once the image has been pre-processed and broken up into zones, it's time to recognize the text. Text recognition is performed in multiple, interconnected steps, which are outlined below:

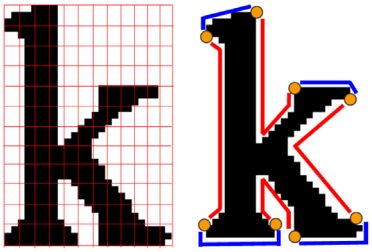
Character Segmentation – This process attempts to separate (or join) blobs of pixels into characters. The algorithms use splitters, joiners and smoothers in an iterative process to determine the best way to segment pixels into characters.



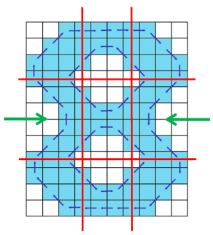
Character Classification – OmniPage Capture SDK contains three primary font-agnostic character classifiers. These character classifiers are codenamed **Paprika, FireWorx** and **Mango**. These AI and CV classifiers work in conjunction with the segmentation algorithms mentioned above to identify what characters the pixels of the image represent.

Here is a breakdown of these three character classifiers:

Paprika – This character classifier uses a CV algorithm called contour tracing. It separates character shapes into convex and concave arcs. Multiple features of these arcs, including the number of arcs, arc end points, arc lengths and direction of concavity, are extracted and fed into a k-Nearest Neighbor (k-NN) classifier (a common AI algorithm) to identify the character. The classifier also contains some built-in rules to deal with special cases.



FireWorx – This classifier has two components. The first component is a matrix matcher CV algorithm that partitions characters into a 12x12 grid (144 binary values). These grids are fed into a fuzzy AI decision-tree classifier that performs bit-by-bit comparisons of interesting leaves. The second component uses a chain code histogram to identify the contour steps. These are then fed into a neural network classifier.

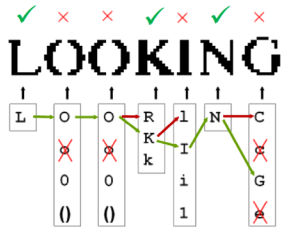


Mango – This character classifier uses 8x5 grid matrix matching to identify features classified with a k-NN algorithm. It also contains a decision tree-based "super classifier" that takes as input the output from the other classifiers to produce an even more reliable result.

Adaptive Classification – In addition to the three character classifiers (Paprika, FireWorx and Mango), there are also adaptive character classifiers. These classifiers create "templates" out of confidently recognized characters and then use these templates to attempt to segment and classify poorly recognized characters.

Neighborhood Information – OmniPage can also gather information from the "neighborhood" of nearby characters. Di- and tri-grams (groups of two and three characters) are analyzed to look for nonsensical combinations and to help resolve alpha/numeric ambiguities.

Dictionary Referencing – OmniPage takes advantage of dictionaries using a Directed Acyclic Word Graph (DAWG) and wildcard searching to attempt to match words to dictionary terms. OmniPage includes standard dictionaries for 17 different languages, as well as vertical dictionaries for the medical, legal and financial industries. Developers can also create their own custom dictionaries to support terms specific to their domain.



Voting – Finally, all of the information gathered during the character classification phase is collected and fed into a voting algorithm that selects the final character candidates and grades them with a confidence score. These results are then made available through an API and can be used during output file creation.

Snapshot:

Benchmarking Data

Below is a brief summary of some of the benchmarking of the OmniPage engine on a handful of different data sets.

Input type	Number of images	Accuracy
Western TIF images	4958	97.69%
Western Camera photos	244	97.10%
Chinese TIF (Simplified)	100	95.58%
Chinese TIF (Traditional)	82	95.56%

Conclusion

Driving productivity in increasingly mobile and home office workforces and reducing costs are challenging, particularly when paper documents add complexity to business practices. To digitally transform operations and maximize revenue, organizations need scalable and reliable solutions to streamline document processing applications with minimal programming requirements.

Whether your business only scans a few documents that require subsequent editing or it's something that occurs many times a day, OmniPage empowers teams to be more productive. And the power and versatility of the OmniPage Capture SDK allows your organization to add OCR and imaging capabilities to any platform.

Work Like Tomorrow.™

KOFAX

kofax.com

February 5, 2021 2:17 PM