How Grooper Overcomes the Limitations of OCR

When converting data from paper documents into digital text, it's important to understand the process and technology behind conversion. Without delving in too deep, this knowledge will help you learn how OCR technology works.
Additionally, you will learn why legacy or traditional OCR methods fail to capture significant amounts of a document's data, and how Grooper's technology overcomes these failings to capture data.

Optical Character Recognition

Most people who have heard of OCR think of it as a feature included with another piece of software to perform word searches. But OCR is much more than that.

Optical Character Recognition, or OCR, has essentially been around since 1913. First used to interpret Morse code and assist the blind, OCR technology has continued to evolve. Using OCR technology, a computer uses rules and criteria to compare patterns made by letters and numbers on scanned documents to a set of characters stored in the software.

The Limitations of OCR

While OCR is saving a tremendous amount of time and reducing data entry, those who use typical OCR often know that the results are not highly accurate. With unreliable data being pulled from scanned images, there is still a lot of manual intervention required.

As OCR works to recognize patterns, many things can confuse the technology and cause problems. A couple of factors that can prevent good OCR results are image color (black-and-white) and resolution (quality).

Having a scan that is only black-and-white creates high image contrast, making it easy for the computer to see clearly where characters begin and end. When it comes to resolution, a scan of a document that is low resolution creates a lot of noise around characters, confusing OCR.

Using Grooper's tools can create high contrast black-and-white images and overcome poor scan resolution. Many other factors can cause OCR problems, like text segments, text in boxes, columns, different sizes of text, and offset lines.

How Grooper Overcomes OCR Limitations for Great Results

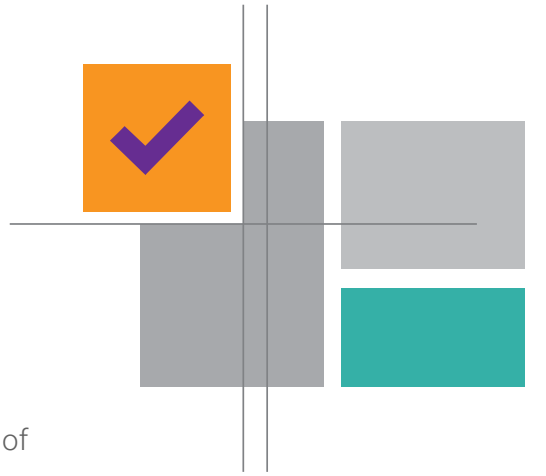
Grooper has other tools to overcome these limitations. These tools are called:

- Bounded Region Detection
- Segment Reprocessing
- Iterative OCR
- Cell Validation
- OCR Synthesis

Grooper uses these tools to hone in on the problem areas and get as good of an OCR read as possible. Let's take a look at each one.

Bounded Region Detection

A bounded region is simply words in a box. Imagine an invoice, purchase order, or delivery receipt. These kinds of documents are comprised of text in tables and boxes. Grooper finds those boxes, looks only at what's inside them, and as a result, gets a really good read on the contents. Grooper just took what has typically been a nightmare for image cleanup to get rid of and used it to its advantage.



Segment Reprocessing

Just as it sounds, a segment is a small block or line of text on a page. If any segment receives a low OCR confidence score, Grooper uses Segment Reprocessing to run OCR on those segments a second time. The outcome is much better results for each of these troublesome lines.

C

Iterative OCR

O

Documents will frequently use different-sized fonts and free-floating text that are not in alignment. OCR reads from left to right like we do. Therefore, if fonts on the left are of a different size or are out of alignment with text on the right side, typical OCR will generate poor results.

Grooper solves this problem with Iterative OCR, or by reading the document multiple times. The first time, it will read everything. The second time, using Iterative OCR, Grooper will drop out what it read well and then only read what was previously poorly read. The dropped-out text will no longer interfere with what's left, resulting in a much better read.

R

Cellular Validation

A page with two or more columns presents a challenge for typical OCR. Text in the columns may have different sizes, or the lines may be offset from one another. Many OCR processes will have a total breakdown with very inaccurate results.

However, Grooper uses Cellular Validation to create highly customizable OCR regions by splitting a document into specific rows and columns. Rows and columns are defined by the Grooper user who understands the document's structure and layout. Grooper then reads each of the individual rows and columns independently.

OCR Synthesis Puts it All Together

With OCR Synthesis, Grooper combines the data produced from Bounded Region Detection, Segment Reprocessing, Iterative OCR, and Cellular Validation into one logical text flow. Advanced font awareness re-analyzes spaces, tabs, and new line feeds during OCR Synthesis. Grooper ensures that all characters in a document are not just recognized but are assembled together in logical groupings.

These advanced tools set Grooper apart from traditional OCR systems by providing a vastly improved foundation for accurate and reliable data capture. Grooper takes the concept of OCR from simply identifying words in a document to creating an intelligent digital awareness of a document.

- What problem are you working on solving today?
- What will happen if the problem isn't fixed?

Let's talk!

Click here for a technical deep dive into the mechanics of how Bounded Region Detection, Segment Reprocessing, Iterative OCR, Cell Validation, and OCR Synthesis all work together within Grooper.