Microsoft

Transparency Note: Optical Character Recognition (OCR)

Updated 8/20/2021

Table of Contents

What is a Transparency Note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the Microsoft AI principles.

Introduction to Optical Character Recognition (OCR)

Businesses today frequently need to convert text from images, scanned paper documents, and digital files into actionable insights. These insights power knowledge mining, business process automation, and accessibility of content for everyone. Optical Character Recognition (OCR) is an AI service used to extract text from visual content such as images and documents. OCR currently supports several languages for extraction of print text (see OCR supported languages). Handwritten OCR is currently supported exclusively for English.

The basics of OCR

The OCR technology from Microsoft is offered via the Computer Vision Read API. Customers call the Read API with their content to get the extracted text, its location, and other insights in machine readable text output. They process the output within their business applications to implement content intelligence, business process automation, and other scenarios for their users.

Term	Definition
Asynchronous	Asynchronous means that the service doesn't immediately return the extracted text. Instead, the process starts in the background. The customer application will need to check back at a later time to obtain the extracted text.
Read	The Read operation is an asynchronous call that accepts images and documents to begin analysis and text extraction, which is returned via another call.
Get Read Results	While the analysis and extraction process is active, the Get Read Results operation outputs the progress status. When the process is complete, the Get Read Results operation outputs the extracted text (in the form of text lines and words) and confidence values.
Confidence value	The Get Read Results operation returns confidence values in the range between 0 and 1 for all extracted words. This value represents the service's estimate of how many times it correctly extracts the word out of 100. For example, a word that's estimated to be extracted correctly 82% of the time will result in a confidence value of 0.82.

Example use cases

The following use cases are popular examples for the OCR technology.

- **Images and documents search and archive -** Unstructured documents such as legal contracts, technical documents, and news content contain rich information and metadata that are not available for processes such as automated tagging, categorization, and search. OCR allows the text from these documents to be machine readable for analysis, search, and retrieval.
- **Image content moderation and localization** - eCommerce companies, user-generated content publishers, and online gaming and social media communities need to moderate images to be compliant with online safety regulations. In certain cases, they also need to localize content for international audiences. OCR allows you to extract text from images to apply downstream processing.
- **Business process automation -** Business process automation requires integrating user-entered data and preferences in documents and application screens with complex business processes. OCR unlocks the text embedded in documents and images and makes it available for use in the steps of the business workflows.
- **Financial and healthcare documents processing** - When used in back office processing of financial and insurance application forms, OCR helps save time and effort in document processing. Similarly, OCR applied to medical claim reimbursements and medical information forms speeds up reimbursements and qualification for services and benefits.

Considerations when choosing other use cases

Consider the following factors when you choose a use case.

- **Carefully consider when using for awarding or denying of benefits** - Using OCR output directly for awarding or denying benefits can result in errors if based on incorrect or incomplete information. For example, when filling out medical forms, users can make errors or fail to include important information. Additionally, OCR can potentially misread or not detect parts of the form. To ensure fair and high-quality decisions for consumers, combine OCR-based automation with human oversight.
- **Avoid use for signature identification** - When you extract handwritten text, avoid using the OCR results on signatures to identify individuals. Signatures are hard to read for humans and machines alike. A better way to use OCR is to use it for detecting the presence of a signature for further analysis.
- **Don’t use OCR for decisions that may have serious adverse impacts** - Examples of such use cases include processing medical prescriptions and dispensing medication. The machine learning models that extract text from prescriptions can result in undetected or incorrect text output. Decisions based on incorrect output could have serious adverse impacts. Additionally, it is advisable to include human review of decisions that have the potential for serious impacts on individuals.

Capabilities and limitations of OCR

In this section, we'll review what accuracy means for OCR and how to assess it for your context.

Word-level accuracy measure

Text is composed of lines, words, and characters. A popular measure of accuracy for OCR is word error rate (WER), or how many words were incorrectly output in the extracted results. The lower the WER, the higher the accuracy.

WER is defined as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

Term	Definition	Example
S	Count of incorrect words ("substituted") in the output.	"Velvet" gets extracted as "Veivet" because "l" is detected as "i."
D	Count of missing (“deleted”) words in the output.	For the text "Company Name: Microsoft," Microsoft isn't extracted because it's handwritten or hard to read.
I	Count of nonexistent ("inserted") words in the output.	"Department" gets incorrectly segmented into three words as "Dep artm ent." In this case, the result is one deleted word and three inserted words.
C	Count of correctly extracted words in the output.	All words that are correctly extracted.

Using a confidence value

As covered in an earlier section, the service provides a confidence value for each predicted word in the OCR output. Customers use this value to calibrate custom thresholds for their content and scenarios to route the content for straight-through processing or forwarding to the human-in-the-loop process. The resulting measurements determine the scenario-specific accuracy.

OCR system performance implications can vary by scenarios where the OCR technology is applied. We'll review a few examples to illustrate that concept.

Medical device compliance: In this first example, a multinational pharmaceutical company with a diverse product portfolio of patents, devices, medications, and treatments needs to analyze FDA-compliant product label information and analysis results documents. The company might prefer a low confidence value threshold for applying human-in-the-loop because the cost of incorrectly extracted data can have significant impact for consumers and fines from regulatory agencies.

Image and documents processing: In this second example, a company performs insurance and loan application processing. The customer using OCR might prefer a medium confidence value threshold because the automated text extraction is combined downstream with other information inputs and human-in-the-loop steps for a holistic review of applications.

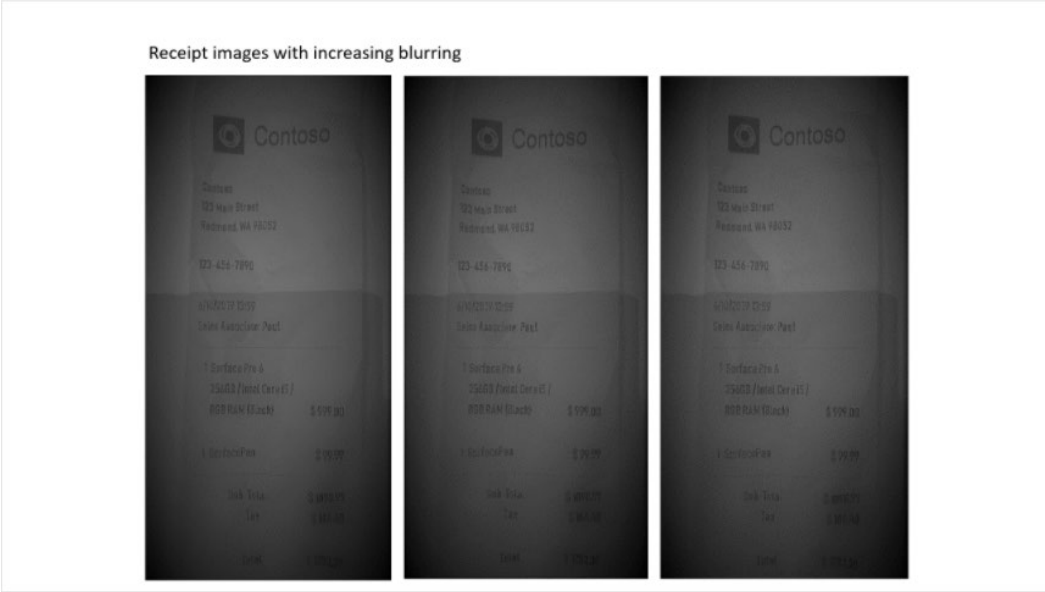
Content Moderation: For a large volume of e-commerce catalog data imported from suppliers at scale, the customer might prefer a high confidence value threshold with high accuracy because even a small percentage of falsely flagged content can generate a lot of overhead for their human review teams and suppliers.

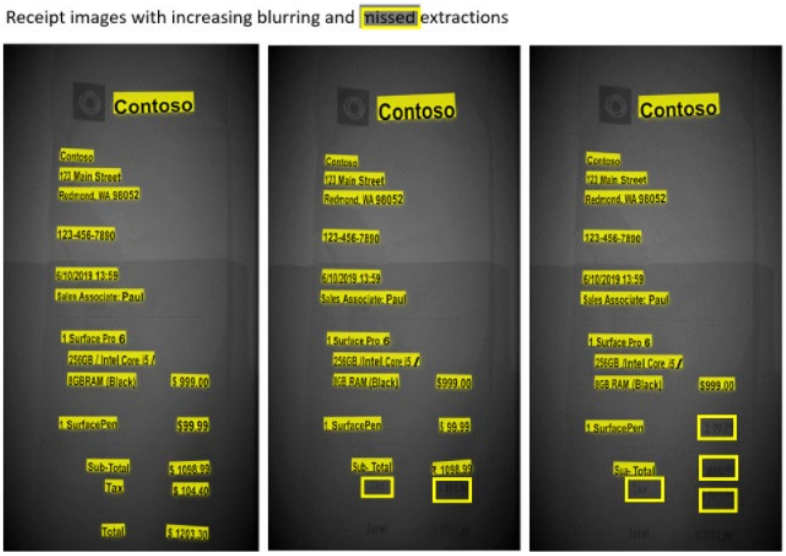
System limitations and best practices to improve system performance

The service supports images (JPEG, PNG, and BMP) and documents (PDF and TIFF). The allowable limits for number of pages, image sizes, paper sizes, and file sizes are listed on the OCR overview page.

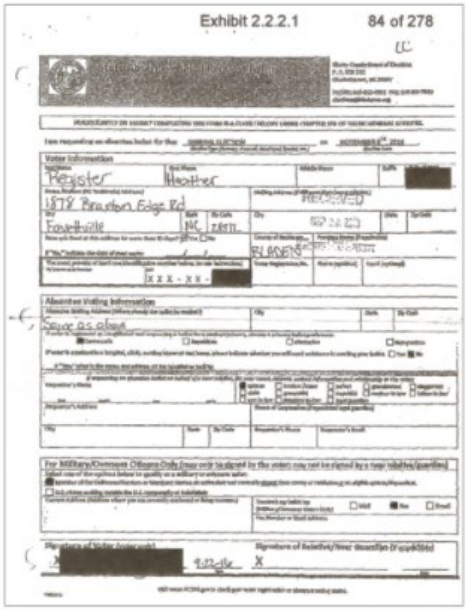
Document scan quality, resolution, contrast, light conditions, rotation, and text attributes such as size, color, and density can all affect the accuracy of OCR results. For example, we recommend the image be at least 50 X 50 pixels. Customers should refer to the product specifications and test the service on their documents to validate the fit for their situation.

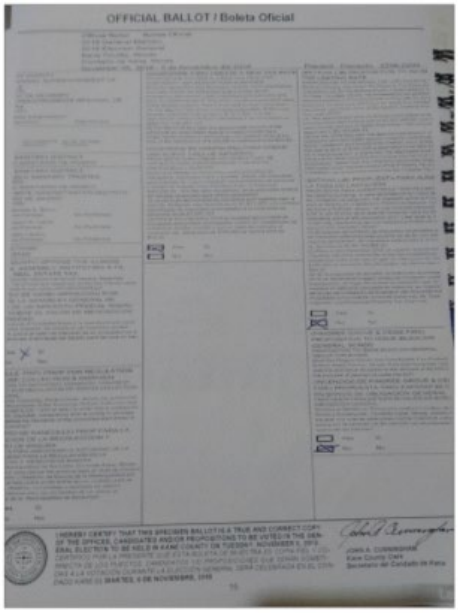
The following example shows a few difficult cases for OCR where you see missed and incorrect text extractions.



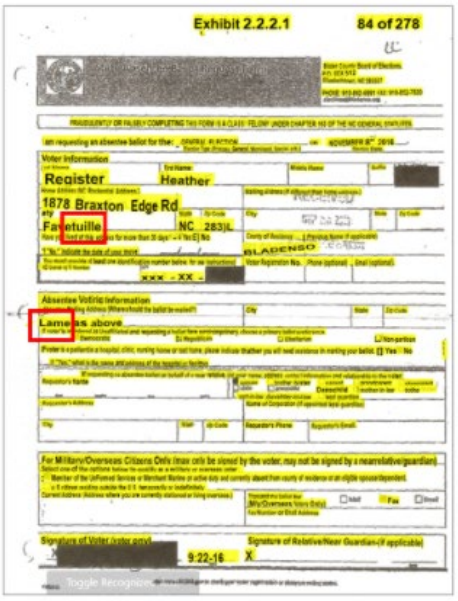


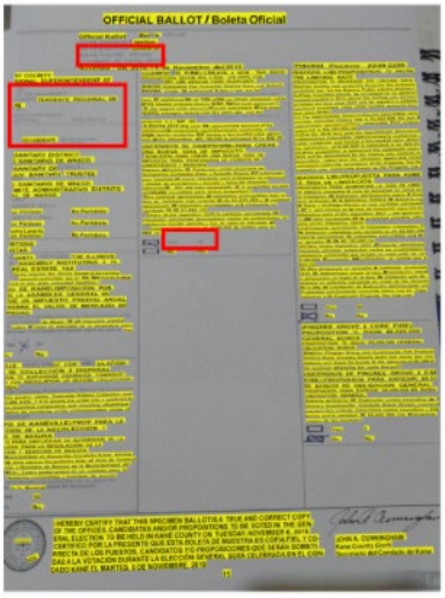
Blurry, low resolution images with handwritten text



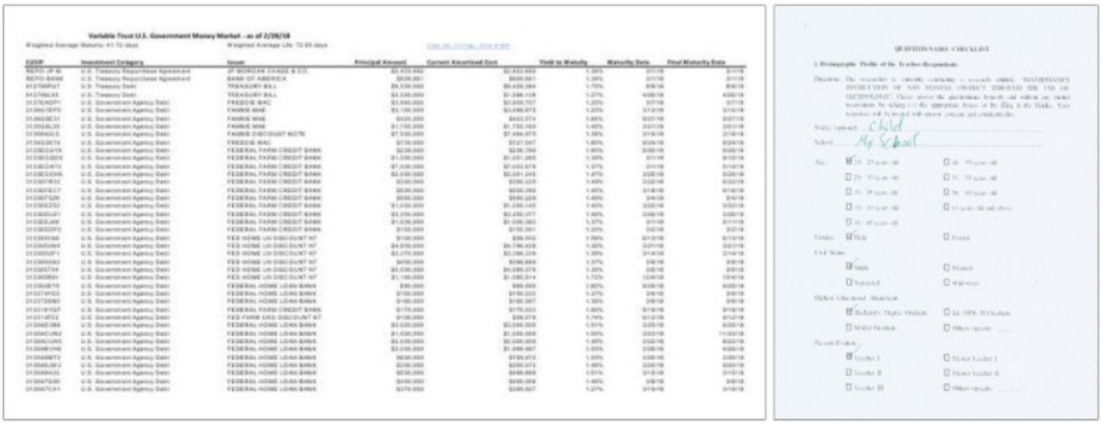


Blurry, low resolution images with handwritten text with missed or incorrect extraction

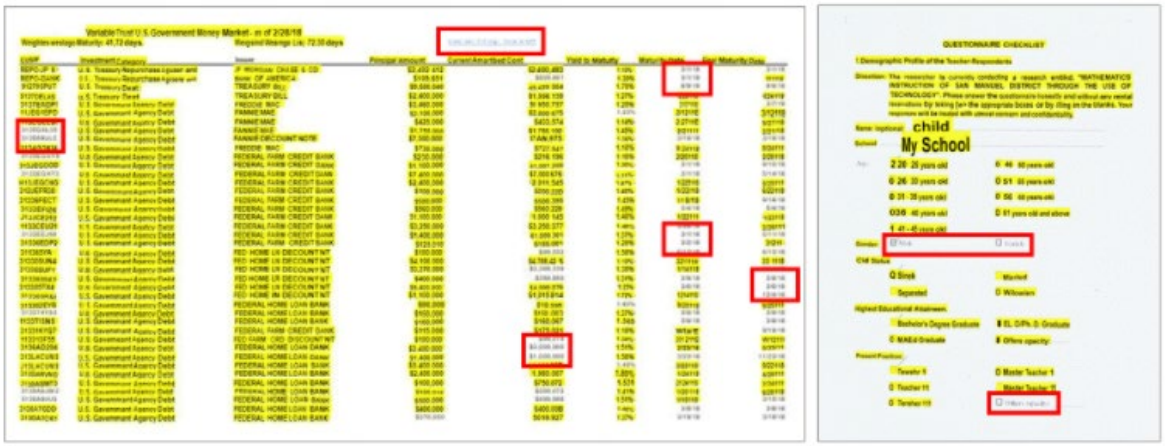




Blurry, dense, small text and low-resolution images with handwritten text



Blurry, dense, small text and low-resolution images with handwritten text with missed extractions.



Receipt images with perspective and surrounding white space impact

Receipt images with perspective and surrounding white space impact with missed extractions.

The current version supports handwriting or cursive-style text only for English. This limitation also affects any related features such as print vs. handwriting-style classification (preview) for each text line.

OCR's performance will vary depending on the real-world uses that customers implement. In order to ensure optimal performance in their scenarios, customers should conduct their own evaluations of the solutions they implement using OCR. The service provides a confidence value in the range between 0 and 1 for each detected word included in the OCR output. Customers should scan a sample dataset representing their content to get a sense of the range of confidence scores and the resulting extraction quality. They can then decide on the confidence value thresholds to determine whether the results should be sent for straight-through-processing (STP) or reviewed by a human. For example, the customer may submit results with confidence value greater than or equal to .80 for straight through processing, and apply human review to results with confidence value less than .80.

Learn more about responsible AI

Microsoft AI principles

Microsoft responsible AI resources

Microsoft Azure Learning courses on responsible AI

Learn more about OCR

What is Optical Character Recognition?

Contact us

Give us feedback on this document

About this document

Published: 08/20/2021

Last updated: 08/20/2021