Shelby Jacobs

Increased Efficiency in Archives Through Optical Character Recognition
Introduction to Optical Character Recognition

Optical character recognition, also known as OCR, is the identification of typed, printed, or handwritten characters by computer software. After the characters are recognized, they are then converted to computer text. Since the invention of OCR software, developments have been made to improve efficiency, accuracy, and flexibility. Today, there are many types of optical character recognition that can identify individual characters and words from a variety of typed and handwritten sources. This technology can be applied to digitization projects in archives to increase the efficiency of data entry, improve searchability, provide accessibility to researchers, and advance productivity.

History

Optical character recognition is one of the oldest forms of data entry. There are US patents for technology that converts images to text as early as the 1800’s. It wasn’t until computing was popularized that OCR technology was applied to data entry. Now, it is common for businesses and government offices to use OCR for data entry. For instance, the US Postal Service uses OCR software to sort mail and even correct misspelled words to safeguard correct routing.
Today, there are many examples of technology with an OCR component. Apps have been developed to translate signs from one language to another. Digital libraries like Google Books and Project Gutenberg have digitized millions of books, which would have been impossible without character recognition. There is even technology to prevent character recognition in order to distinguish between humans and computers, such as CAPTCHA, which intentionally distorts images of letters to a point that OCR technology cannot be used for spam or data extraction.

How OCR Works

Optical character recognition works slightly differently depending on the type of writing that needs to be identified and whether or not the computer is identifying individual characters or full words. Below is a table that outlines four basic variations of OCR.

<table>
<thead>
<tr>
<th>OCR Technology</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Character Recognition</td>
<td>Recognizes typewritten characters, one character at a time.</td>
</tr>
<tr>
<td>Optical Word Recognition</td>
<td>Recognizes typewritten words, one word at a time.</td>
</tr>
<tr>
<td>Intelligent Character Recognition</td>
<td>Recognizes handwritten characters, one character at a time.</td>
</tr>
<tr>
<td>Intelligent Work Recognition</td>
<td>Recognizes handwritten words, one word at a time.</td>
</tr>
</tbody>
</table>

Regardless of the variation, optical character recognition works in stages. This step-by-step process is used to ensure accuracy.
First, there is **preprocessing**, which adjusts the characters that need to be scanned to make them easier for the software to read. This step includes reducing pixelization, straightening skewed text, and removing none-characters that cannot be recognized by the software. The OCR program starts by identifying the layout of the page. The software will adjust skewed images, identify text boundaries like spacing and columns, and address complicated layouts. Then, the image is optimized to smooth and sharpen the edges of character glyphs. A good quality scan is always a good starting point. The preprocessing sets the stage for accurate character recognition.

The next step is the main part: **character recognition**. At this stage, the computer compares the image of the characters than need to be recognized to a database of glyphs. It then produces a list of likely candidates for the character. This step becomes more complicated when different fonts or scales are encountered. That is why feature extraction is important. Feature extraction accounts for variations in proportion, line directions and intersections, and other features that vary between different fonts and handwriting. Then, algorithms are employed to estimate the “nearest neighbor”, which is the most likely character that a glyph represents.

Before the software produces a final product, there is a **post processing** stage. This part of the process compares the characters or words the computer recognizes to a preset lexicon or dictionary. The lexicon contains words that are approved to appear in the document. The lexicon can include every word in a language or a selection of words relating to a topic. Occasionally, the software will identify words that are not in the lexicon, in which case the computer selects a close match. This part of the process is similar to the autocorrect feature on smart phones or word processors. Some of the more sophisticated programs also have basic grammar rules programmed into them to ensure greater precision, such as choose a noun over a verb when appropriate.

Computers cannot recognize words and grammar on their own, which means there will inevitably be mistakes. **Crowdsourcing** helps to reduce mistakes by using humans to recognize errors that a computer cannot identify. Programs can be “trainable” which means that common errors made by the software will be accounted for in future recognitions. This means that over time, the post processing stage will become more accurate and need less crowdsourcing. Using crowdsourcing to examine entire documents
is not a viable option for large-scale projects, so often humans only check the title, author, and first few lines for accuracy.

**Application in Archives**

Data entry can be a time-consuming process. Traditionally, archivists transcribed documents without the help of a computer, which wastes valuable human resources that can be best used on other activities. Optical character recognition can be applied to data entry because it extracts text that would otherwise have to be typed by hand. Instead of scanning, transcribing, and typing each and every document, archivists can now scan a record and run it through OCR software. This technology can also be useful for extracting metadata from documents. This decreases the need for an archivist to study the record, identify key identifying features and input them into the database. A downside to using computers to extract data and metadata is that computers prioritize different identifiers than humans do. For example, a computer can select the most frequently used word, but that word may not be the most important word to describing a record. This is especially relevant for the identification of concepts, which is what many researchers search for.

A useful feature of documents with live text produced from an OCR program is that they are readily searchable. Search engines rarely have the capability to search the elements of an image, so it is important to have searchable text in order to increase search results. This is where optical character recognition becomes valuable to improving document searchability. Live text means that the search engine can identify keywords, which is a common way that people search for resources online.

An important application of optical character recognition is accessibility. Researchers with visual impairment rely on screen readers to translate the written word to audible sound. Images can only be described by the metadata that is included with them in the coding, which is either limited or nonexistent. OCR technology takes it another step further by providing live text directly taken from the original image of a record that can be read by screen reader programs. Similarly, researchers may also be interested in records than are not in a language that they can read. Live text can be translated into another language, while glyphs in an image cannot.
Programs

Adobe Acrobat Pro DC is the most common program that uses OCR technology. It has all of the most popular and useful OCR features (including a smart phone app version) and accommodates 190 different languages. It is notable for its accuracy and usability. The biggest downside to Adobe Acrobat is that it cannot export to eBook formats.

Nuance’s OmniPage Standard 18 automatically detects the language on the document that is being converted. The converted file will mimic the font and layout of the original image or PDF. OmniPage Standard is offered in two forms: professional and standard. While the standard version works well for most digitization and data entry projects, the profession version has features that help with large-scale projects.

ABBYY FineReader does well with character recognition, which means it is one of the most popular programs. It does especially well with lower-quality images. However, the navigation of the features is not intuitive, which means there is a bit of a learning curve when first working with this software.

Google Drive is a free web application, which makes it perfect for individual use and small projects. It is less accurate than the other software mentioned because it has trouble recognizing anything other than basic fonts and layouts. There is also a size limit when using the OCR function of 2MB per file.
Conclusion

Optical character recognition is an important tool that should be utilized by archivists when digitizing records. Using character recognition software such as Adobe Acrobat Pro DC, OmniPage Standard 18, or ABBYY FineReader increases productivity by saving valuable time and effort retyping pages of text. In addition, live text is beneficial because it makes pages more searchable by search engines, which rely on word frequency. OCR resolves accessibility issues with PDFs and images of text because live text can be read by screen reader programs or translated into other languages. There is still room for improvement in OCR technology, but crowdsourcing and improving lexicons are resolving problems that were once more common. Optical character recognition should be employed for digitization projects.
Works Cited


