J6 Delving Into Digitally Processed Documents: How Does Optical Character Recognition (OCR) Impact Documents?

Nina Harnarine, BSc, Forensic Examiners Inc, Toronto, ON M4W 3H1, CANADA*

Learning Overview: After attending this presentation, attendees will understand the impact of OCR technology on various documents containing signatures, writing, different writing instruments, and text typed on a typewriter.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by providing information on how OCR technology may impact a document that is being examined.

OCR or optical character reader is the process of converting images containing text into machine-encoded text. The images can be scanned or photographed, and a software application converts all text components. OCR integrates three research fields: pattern recognition, artificial intelligence, and computer vision.

In the digital age of artificial intelligence and automation, OCR has become popular for various applications. Documents containing text that have undergone OCR processing can be electronically searched, edited, stored, and uploaded. In addition, this digitized text can be used in machine processes, such as cognitive computing, text-to-speech, machine translation, and text mining.

Forensic Document Examiners (FDEs) may be asked to examine digitally processed documents. Some businesses are shifting toward becoming paperless. This shift within some businesses has resulted in FDEs examining more digitally processed documents. With the aid of technology, four individuals from four corners of the globe can all sign one document without ever having to sit around the same table. Signatures from four different individuals would require digital processing of the document, likely scanning and re-scanning of the document. What does this document look like after undergoing multiple digital processing steps?

This research sets out to determine how OCR technology impacts documents with text. There are numerous OCR software applications available. The documents in this research project were processed with the OCR function in Adobe® Acrobat® Pro DC. How does Adobe® Acrobat® Pro DC OCR read different signature styles? How are these same styles of signatures reproduced when scanned and converted into a text document using the Adobe® Acrobat® Pro DC?

Documents can be written using various writing instruments. Is there any difference in how the Adobe® Acrobat® Pro DC OCR function reads and converts the text for different writing instruments?

Some documents contain a combination of typed and written text. How does Adobe® Acrobat® Pro DC OCR read cursive or printed or mixed (both cursive and print) text? How are these same writing styles reproduced when scanned and converted into a text document using the Adobe® Acrobat® Pro DC?

One may want to digitize their old records, including documents that have been composed on a typewriter. How does Adobe® Acrobat® Pro DC OCR read typewritten documents? How are typewritten documents reproduced when scanned and converted into a text document using the Adobe® Acrobat® Pro DC?

This research will also examine some of the mobile OCR applications, including Scanbot®, Dropbox®, Adobe® Scan, Microsoft® Office Lens, and Google® Keep.

In this digital age, FDEs should be cognizant of how digital processing techniques such as OCR may impact documents. Artifacts and digitization can be observed in digitally processed images. Understanding how the artifacts or digitization occur is important. If no original document is available, the best available copy should be requested. If only a digital file is available, the digital file in its original file format should be submitted for examination. An FDE may be asked to examine a hard copy of a document that has undergone multiple digital processing. It is important that FDEs understand the evidence that digital processing applications, such as OCR, may leave behind on the document and how the digitally processed document may differ from the original or even the first-generation copy.

Optical Character Recognition (OCR), Document Processing, Scanned Document

Copyright 2020 by the AAFS. Permission to reprint, publish, or otherwise reproduce such material in any form other than photocopying must be obtained by the AAFS.

*Presenting Author