Manipulating Text: Using Text for Manipulation by Storing Messages

John Kilgo

Literature Review #3 - COMP.5460

OCR - Optical Character Recognition has always fascinated me due to the ability to take documents for which there is no original, scan it, and then get an editable document. What did not dawn on me was that you could also embed information within a document, besides the information itself. OCR is not perfect, but it sure beats having to retype a document.

I looked at an overview on Tesseract which discussed this now open source OCR software product. This product was developed in the 80s and 90s and was revolutionary for its time due to the ability for the engine to look at the whole document rather than an interactive approach with very specific sections of the source document. At the same time, the product was able to distinguish and classify shapes so that it could take a “blob” of information and distinguish white text on a black source, something that was harder to do at the time.

In manipulating text by storing additional hidden information in it, having an OCR solution is what is necessary to be able to decode this information. The goal is for the information to be hidden from the human eye so as to not distract from the the original text being conveyed.

So, my primarily article builds on what was learned in the text imaging space, and chose to use Tesseract to scan their encoded text. Using fonts that were slightly modified from original fonts like the Times, but are not easily distinguishable to the naked eye, the group of researchers was able to encode hidden information in plain view! One possible use case a barcode that did not look like a barcode, but was distinguishable by their OCR app.

I can think of many more nefarious uses for this type of technology, but one constraint is how much can you store? Well, these researchers base how much can be stored on usable characters within text. And, the more text there is, the more data that can be stored. By modify each character, they can have a system of codifying these letters with a numeric pattern that can then be used to store data.

This may not be the most secure way to store information, but I can think of applications for text to be viewable to both human and artificial sources. In the future, a human could see an electronic road sign for a restaurant, but a camera hooked up to some sort of artificial intelligence could see embedded information such as the restaurant being out of enchiladas.

OCR technology and embedded data within “plain” text is something to watch and I look forward to learning more about what is in store in the future.

References:

[1] Chang Xiao, Cheng Zhang, and Changxi Zheng. 2018. FontCode: Embedding Information in Text Documents Using Glyph Perturbation. ACM Trans. Graph. 37, 2, Article 15 (February 2018), 16 pages. https://doi.org/10.1145/3152823

[2] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633.
doi: 10.1109/ICDAR.2007.4376991
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4376991&isnumber=4376969

FontCode: Embedding Information in Text Documents Using Glyph Perturbation is my primary article.