Spring 2020

Taxonomy Techniques for Holocaust-Related Image Digitization and Text

Sejin Kim
Kenyon College, kim3@kenyon.edu

Taxonomy Techniques for Holocaust-Related Image Digitization and Text

Sejin Kim | {kim3}[at]kenyon.edu
Kenyon College | Integrated Program for Humane Studies

Abstract

The Holocaust is no doubt an incredibly tragic and dark moment in human history. However, an incredible amount of primary-source material has survived and was archived, in hopes that researchers, scholars, and students could use it down the line. However, old archival and categorization techniques place like-topics together, using categories like propaganda, concentration and death camps, and Ghettos, seldom including information like resource type or size, like whether the resource is a document, image, film still, illustration, or certificate. Therefore, I intend on outlying a categorization method that can be used by researchers to break down resources into their component categories so that they can be analyzed with a greater degree of precision. With these techniques and resources, historians have a much better chance at finding applicable and important resources to their research that would have otherwise been unfindable. Computer scientists can better optimize their machine learning algorithms and tailor them to specific document types.

Overview

This project attempts to outline a strategy to categorize libraries of previously broadly-categorized datasets and databases of Holocaust-related materials, for various uses, including those by students, researchers, archivists, and librarians.

For this project, I am building a strategy to find, analyze, and categorize datasets of Holocaust-related materials so that they may be further broken down for increased accessibility, using technologies like optical character recognition (OCR) and computer vision (CV).

I attempt to solve this problem by determining the locations of available datasets, what format they are currently in, then by breaking them down into their component types in a decision-tree-like manner. Finally, I specify a manual bypass to handle unseen data distribution.

Ethics & Terminology

Studies of the Holocaust, one of humanity's darkest times, naturally brings up questions on ethics and proper terminology. I am, by no means, not the first person to ask these questions, nor am I qualified to answer them, but I present them for your consideration.
Firstly, we must recognize that some terms are often associated with Nazism, even if the author did not intend to refer to a Nazi term. For this research, I follow scholar and author Doris L. Bergen. In Bergen's *War & Genocide*, she chooses to refer to the blanket racism and hatred as "antisemitism," as opposed to hyphenated ("*anti-Semitism*"), which would assume that "semites" existed and could be classified. Likewise, terms such as "*euthanasia*" are not used; more appropriate terms are "murder" or "killing," in the interest of historical accuracy. Finally, the term "event" will be replaced with "process," indicative of the history before, during, and after the Holocaust and the killing stages.
There are plenty of people who believe that the Holocaust is an important topic, and that studies of the process are important, especially today. There are also some people who believe that the Holocaust and all of its atrocities are so horrible that they should not be spoken of except for those direct survivors who choose to speak out. These people hold no "Holocaust denier" agenda, but they believe that studies of the Holocaust are either occurring too soon or should never occur. This brings a natural question: do we, as scholars, have a right to study the Holocaust and the millions of lives that perished during the process? If the malice is so severe that it is truly unfathomable, then do we have a right to look at the images, films, and other primary sources that depict and show blatent murder? In my personal opinion, I believe that while the atrocities that occurred are no doubt tragic, they still must be studied.

Increasing OCR Accuracy

Optical character recognition relies on a good model that is not so generalized that it no longer works effectively. Most "turn key" solutions are rather generalized or generic, meaning that they return barely usable or unusable texts when presented with abnormal images.
By training an optical character recognition model on certain *types* of text, like handwriting, script letters, typewriting, press-printed pages, and other type styles, we can create a much more accurate set of recognition models that are more useful for a computer to index and search for and for a researcher to find and read.

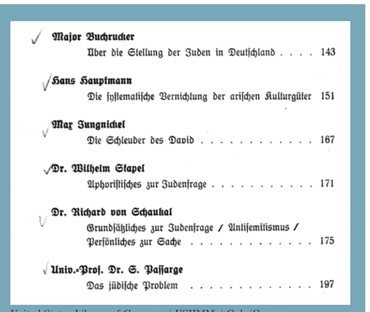
Example given: A generic OCR attempted to read and transcribe the page to the right. It returned the following text:

```
/ iajor Zud)rucder Iber bie Gtellung ber 3uben in 'Deutitd)anb . . . . 143 /ians asaupfmann Tie ft)(tematifd)e
3ernid)tung her ari[d)en Aulturgfiter 151 iar sun tgntiche Die Gd)~euber bes aib . . . . . 167 ~r. 2.
tIelm Gtapel 1pl)oriftifd)es ur Subenfrage . . . . . 171 Mr. d, tcarb oan Gd)aubal v OrunbfJdilides 35ur
Subenfrage / Qntitemitismus / 'Perfinlies aur Gad)e . . . . . 175 v UiUnt.--rof. )r. G. 3aflarge Das
jfibifcIe 3roblem . ..... . . ... . .. 197
```

It returned an OCR confidence of 78%.
For reference, a manual transcription is as follows:

Major Buchrucker - Über die Stellung der Juden in Deutschland - 143 - Hans Hauptmann - Die [unknown] Vernichtung der arischen Kulturgüter - 151 - Mar Jungnickel - Die Schleuder des David - 167 - Dr. Wilhelm Stapel - [unknown] zur Judenfrage - 171 - Dr. Richard von Schaukal - Grundfäkliches zur Judenfrage / Untilfemitismus / Persönliches zur Sache - 175 - Univ. Prof. Dr. S. Pattarge - Das jüdische Problem - 197

In this press-printed text, the page has stray markings that can distract or confuse an optical character recognition model. By adopting strategies listed below, such as letter and text grouping and word reconstruction techniques implemented in a specialized recognition model, we can create a much more accurate and complete transcription of the page. Of note is, the generic OCR believed that it had faithfully interpreted 78% of the text, not that the text generated was 78% correct.

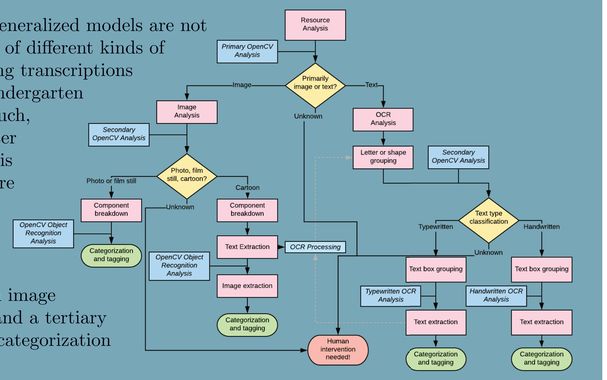


United States Library of Congress / USHMM / Gale/Cengage

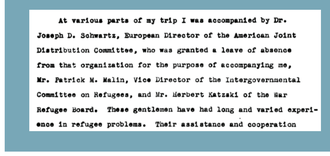
In the generic OCR's interpretation, we also notice that the OCR had attempted to transcribe words that it had no idea about, and for those words, it was willing to make wild guesses on a character-by-character basis, ultimately making up words that made no sense or which contained unexpected symbols. A better alternative would be to flag words that were illegible to the OCR. This information could be used to train future models. While this hurts the confidence level, it provides a third path for unspecified or fringe cases that prevent the OCR from making gross mistakes.

Categorization

As alluded to in the "Increasing OCR Confidence" section, generalized models are not that great at creating highly accurate transcriptions of a lot of different kinds of texts. Some OCR libraries are much better suited for creating transcriptions of textbook scans, while others are better at transcribing kindergarten writing, but importantly, they cannot be interchanged. As such, I propose a multi-tiered computer vision and optical character recognition model, as shown in the diagram to the right. This multi-tiered solution entails training each level of the software solution independently on pre-sorted and pre-analyzed materials that are indicative of the real-world work units that might be assigned. One real work unit will require a primary analysis to determine whether the resource is an image or text, a secondary analysis based on whether it's an image or text to better understand that particular image or text, and a tertiary analysis to actually transcribe or describe the image before categorization and tagging, which terminates the work unit.

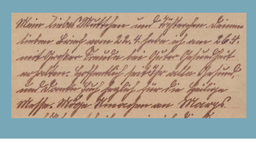


For the full-scale interactive diagram, including expanded examples, please visit https://bit.ly/2WOCsCG



At various parts of my trip I was accompanied by Dr. Joseph D. Schwartz, European Director of the American Joint Distribution Committee, who was granted a leave of absence from that organization for the purpose of accompanying me, Mr. Patrick M. Malin, Vice Director of the Intergovernmental Committee on Refugees, and Mr. Herbert Katzki of the War Refugee Board. These gentlemen have had long and varied experience in refugee problems. Their assistance and cooperation

Obviously, the texts to the left and right cannot be treated the same. A properly trained handwritten text model will be able to parse the text to the right with much greater accuracy than typewritten model could, and vice versa.



Technologies

Computer vision and machine learning are two of our most potent tools, and they can both be used to create and interpret the effectiveness of our models. OpenCV is an open-source computer vision library written for Python 3, and it allows researchers to train machine models to recognize objects or text, depending on the training mode. In order to facilitate effective and productive dataflow, we can also use tools like TensorFlow on general-purpose graphics processing units (GPGPUs), Nvidia CUDA, or on Nvidia Tensor cores.
Within the OpenCV library, we can utilize modules like the `dnn` (deep neural network) module to execute a single shot detection (SSD) on objects if we prioritize speed or a region-based convolutional neural network (R-CNN) if we prioritize accuracy. Both of these approaches allow us to categorize items seen in an image for tagging purposes. This is useful for researchers that want to find certain types of images by visual feature (person, sign, food, insignia), rather than by the keywords assigned to the image by the original archivist.
Additionally, several OCR libraries are needed or need to be trained: one for typewritten text and one for handwritten text. Further research could split the types of text into more subtypes, with less generalization in each OCR library. For example, certain press-printed texts look much different from typewriter-printed texts, and generalizing between the two printed texts could lead to a higher error rate with a falsely high OCR confidence rate.

Data

As with any research project, data is instrumental in the research process. My starting-off point was the Bulmash Family Holocaust Collection at Kenyon College, a curated collection of Holocaust-related resources. This collection contains high-quality scans or images of approximately 1.4k artifacts and resources, and the researcher believes that this is representative of other types of Holocaust documents that may be carried by other collections, such as those in the United States Holocaust Memorial Museum. Resources in the USHMM's collection were found using Gale/Cengage Publishing's Archives Unbound Holocaust database, which hosts for the USHMM. Collections include pre-Holocaust artifacts, such as German antisemitic propaganda and the Nuremberg Laws or the Nazi annulment of German-Jewish nationality to Holocaust correspondence from German concentration camps, and government memos from Western countries in response to Jewish cries for help.

Conclusion

As seen in this categorization framework, the world of optical character recognition is not as simple for historians and researchers in digital humanities as just tweaking a pre-tuned model to increase the confidence level in a given OCR work unit. A wide variety of resources in different forms and different formats influence a computer's ability to fully and completely interpret an image that is presented to it in a way that is human-readable. Digital humanities offers an excellent ingress point for merging high-tech machine learning and deep learning with history, allowing historians the ability to find the most relevant resources and increase legibility and accessiblity to all. Of course, this does not and cannot replace a historian or an archivist; rather, it is only an aid. Categorization frameworks like these have the potential to return much more accurate and precise information than a generic model cannot.

Discussion & Future Work

This project is only an exploratory analysis of the available datasets of historical documents available and only scrapes the surface of the many difficulties associated with research in this field. One key factor is that while mayny of these datasets are available from various sources, as described above, almost all of it is untranscribed and largely unexplored. Much more work can be performed on this topic, including training specific models on pre-made datasets and employing crowd-sourcing techniques to build reliable datasets, such as through platforms like Zooniverse. Future production software could be developed that is more generalized to historical documents and resources from different eras and is capable of interpreting more than one language.

Future research might entail:

- Developing and running OCR and OpenCV deep learning models
- Building a word reconstruction pipeline (on a per-language basis)
- Optimizing analysis time
- Parallelizing and multi-threading computing tasks for improved efficiency

Acknowledgements

I thank Dr. Jon Chun and Dr. Katherine Elkins (Kenyon College Digital Humanities) for providing extensive guidance on the data analytics, computer vision, and categorization strategies. I also thank Dr. Leo Riegert (Kenyon College Department of Modern Languages) for his indirect assistance in the background research and Dr. James Skon (Kenyon College Department of Mathematics, Scientific Computing) for his indirect assistance in software engineering. Lastly, I thank the Kenyon College Greenslade Special Collections and Archives and the United States Holocaust Memorial Museum for access to their datasets.