

I'm Not a Murderer! Issues with AI for OCR

*WANTED FOR GRAND LARCENY: A woman styling herself here as Mrs. F. W. Williams, Test Medium and Clairvoyant, is badly wanted by this Department for grand larceny for fleecing a number of 'soft' female residents out of various amounts of jewelry and coin... She prevails on the victim to leave in her hands jewelry worn by the person in quest of knowledge, also as much gold coin as possible as "then the Spirits will be more apt to reveal."*¹

The above is one of hundreds of wanted ads from *The Detective*, the "Official Journal of the Police Authorities and Sheriffs of the United States," published from 1885 through the mid 1930s. As well as wanted ads, the journal offers its readers job postings for detective agencies, highlights recent arrests, and includes advertisements for such wonders as "Zeglen's bullet-proof fabric... it should be worn by the police and by all others who risk death by shooting—almost as light as gauze, it affords protection equal to that of a stone wall."²

It is a fascinating read and a wonderful reflection of its time and circumstances. It is also available digitally from only one archive; it is not in the Library of Congress;³ it has seemingly never been the subject of any significant scholarly work; and the OCR quality of the available issues is lacking. This last issue, OCR, is understandable: the scans are sometimes incomplete, there are some pencil notes on top of various segments, column structures are inconsistent, and images are sometimes irregularly shaped and break structural boundaries.

This project broadly has two goals. The first is data-oriented: new and better digitization/OCR, including, given the irregular arrangement of pages in the magazine, layout-sensitive text and image recognition. The second uses this new dataset as a base: novel examination of discourse around crime using a number of different methods. These include network analysis of depictions of criminals and law enforcement, as both are highlighted throughout, including under multiple pseudonyms; tracking evolution of crime patterns as depicted through topic modelling; and language use in context, in relation to race and gender.

¹ "Wanted Ad," *The Detective*, September 1901, No. 189.

² "Advertisement," *The Detective*, September 1901, No. 189.

³ Tarrant County College, which holds the collection, is not listed by the Library of Congress. This collection can be found at Tarrant County College due to a donation.

For the SSHA conference, the particular focus will be on contrasting different pipelines, with a critical focus on potential boons and harms of AI OCR. While plenty has been noted about AI hallucinations broadly, crime-related text brings its own issues. Namely, the training data used seems inclined to make events more lurid. In tests thus far, AI OCR outputs murder more often than it is found in the actual text; a pharmacy robbery was changed to a street car heist; and totally novel lines, such as “They are a couple of the most dangerous crooks engaged in this kind of work” are pulled from the ether. The issue, naturally, is that the utter confidence and coherence of this type of bad OCR can pass muster; they are not simply the broken and fragmented text of poor OCR of the past. Gen AI is also willing to offer an answer that it feels it should have even if there is nothing to support it: when prompted that it had missed a wanted ad for the author of this paper, it indicated that Liam Downs-Tepper had escaped from the Kansas State Industrial Reformatory.⁴

The data preparation portions of this work are in and of themselves valuable, as they offer insights into the available methodologies for layout detection and OCR, particularly in light of the recent boom in AI-based approaches in scholarship. This project on the whole offers a critical look at AI OCR and takes a different approach; methodological selection is intended to maximize both transparency and accessibility by both minimizing use of black-boxes and focusing on free tools and resources. This necessarily means turning away from projects like NewsEye, which is based on Transkribus - a paid tool.⁵ This methodological workflow, as both a general model for similar projects and as a concrete code offering, should be considered an additional goal/output of this project.

The actual method is broken into a few deceptively simple steps:

- 1) **Page layout parsing.** Building upon John Scancelli’s work on newspaper layout detection and OCR, although the inclusion of images in *The Detective* - one of the selling points of the journal⁶ - adds complications.⁷ The figure below shows how “smearing” (dilating) the text can be used to isolate segments.
- 2) **Page splitting and OCR of each segment.** Segment OCR decreases error rate by removing surrounding noise and allows for easier error correction. This is particularly true for AI OCR, where the “noise” introduced from other articles and segments can become jumbled.

⁴ Claude.ai, for example, includes enough details that are correct that it could pass a surface muster; further examination shows that its transcriptions are completely inaccurate, changing charges, dates, and descriptions.

⁵ Some other results of this project, like the ONB Annolyzer, are non-functional.

⁶ Issues of *The Detective* include a note on the cover page touting that there are almost 100 images included.

⁷ John Scancelli, “Better OCR for Newspapers,” *Medium* (blog), August 20, 2020, <https://medium.com/@blacksmithforlife/better-ocr-for-newspapers-c7c1e2788b7a.sc>

- 3) **OCR error correction.** Both manual and automated error detection and correction are necessary; automated correction has a robust historiography of its own.⁸ This is, however, one area where AI tools can prove useful, and thus are examined as well.⁹

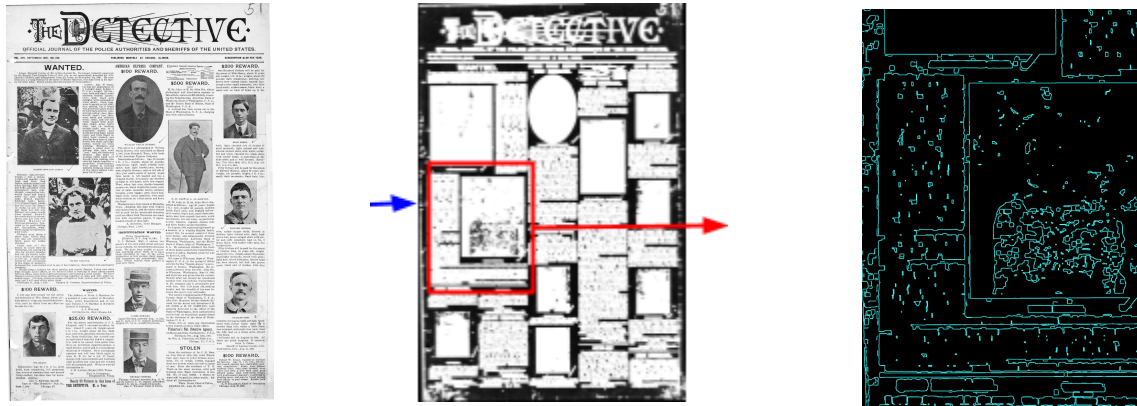


Figure 1: Use of dilation and edges to find individual text segments

This pipeline is intended to be reproducible for any to use on their own materials, creates greatly improved machine-readable text of *The Detective* for broader use, and provides the text data for the content-centered portion of this work.

For this paper, the substantive work is generally more exploratory. Naturally, the data included in *The Detective* is far from objective; the journal is primarily notices submitted by different law enforcement agencies. Examining the text, therefore, should not be viewed as providing insights into crime patterns themselves, beyond the broadest of generalizations. Instead, work on *The Detective* can be used to illuminate discourse on crime itself, and through such, provide a window into public fears and concerns at the time. There is a surprising lack of literature on this subject in a historical context, with the notable exception of Caroline Stabile’s *White Victims, Black Villains: Gender, Race, And Crime News in US Culture*.¹⁰ This paper expands upon some of Stabile’s ideas by applying a distant-reading approach.

⁸ Quoc-Dung Nguyen et al., “OCR Error Correction Using Correction Patterns and Self-Organizing Migrating Algorithm,” *Pattern Analysis and Applications* 24, no. 2 (May 1, 2021): 701–21, <https://doi.org/10.1007/s10044-020-00936-y>; work goes as early as Xiang Tong and David A. Evans, “A Statistical Approach to Automatic OCR Error Correction in Context,” in *Fourth Workshop on Very Large Corpora*, ed. Donia Scott (VLC 1996, Herstmonceux Castle, Sussex, UK: Association for Computational Linguistics, 1996), <https://aclanthology.org/W96-0108>.

⁹ Anderson, Carol, and Phil Crone. “Segmenting Messy Text: Detecting Boundaries in Text Derived from Historical Newspaper Images.” In 2020 25th International Conference on Pattern Recognition (ICPR), 5543–50, 2021. <https://doi.org/10.1109/ICPR48806.2021.9413279>.

¹⁰ Carol A. Stabile, *White Victims, Black Villains: Gender, Race, And Crime News in US Culture*, 1st edition (New York ; London: Routledge, 2006).

There are a number of different specific sub-methods and subtopics included in this distant reading. The first step is the extraction of persons from the text, with attendant information; demographic data like race and age, known aliases, associates, number of mentions across issues, location, reward offered, and so on. This, in conjunction with Named Entity Recognition, allows for much more elaborate work:

- Network analysis of persons included in *The Detective*, which would reveal whether the focus of the journal was isolated individuals or something more coordinated.
- Impact of race and gender on language selection, including potential changes in standards over time.¹¹
- Contrasting demographics with crimes committed and rewards offered: do men and women who have committed the same crime have the same reward offered for capture?
- Changes in time over which crimes are reported; the digitized span of *The Detective* covers, among other things, the emergence of the Model-T and all of the significant societal changes that come with it.
- Spatial analysis of reported crimes, which, layered in contrast to law enforcement recruitment, could provide a look at the split between which crimes are being discussed and where there may be enforcement shortages, show urban/rural splits in crime reporting, and illuminate the interaction of race and place in reporting.

All of these are applications of proven methods, but the combination thereof offers a unique blend of approaches and insights.

This project offers three distinct outputs:

- 1) A streamlined and accessible pipeline for layout and text extraction using freely accessible tools.
- 2) High quality OCR text for an as-of-yet broadly ignored historical resource for study.
- 3) Exploration and examination of an understudied period and subject, cross-applying a variety of DH approaches which have been shown effective for other cases.

Each would be a worthwhile project; the combination is so good that it's basically criminal.

¹¹ A variety of methods come into play here, but are developed following models and ideas put forth in (among others): Lorella Viola and Jaap Verheul, "Mining Ethnicity: Discourse-Driven Topic Modelling of Immigrant Discourses in the USA, 1898–1920," *Digital Scholarship in the Humanities* 35, no. 4 (December 1, 2020): 921–43, <https://doi.org/10.1093/llc/fqz068>; Haifeng Hui, "What Can Digital Humanities Do for Literary Adaptation Studies: Distant Reading of Children's Editions of Robinson Crusoe," *Digital Scholarship in the Humanities* 38, no. 4 (November 30, 2023): 1564–76, <https://doi.org/10.1093/llc/fqad059>; Elad Segev, "Textual Network Analysis: Detecting Prevailing Themes and Biases in International News and Social Media," *Sociology Compass* 14, no. 4 (2020): e12779, <https://doi.org/10.1111/soc4.12779>.

Works Cited

- Anderson, Carol, and Phil Crone. "Segmenting Messy Text: Detecting Boundaries in Text Derived from Historical Newspaper Images." In 2020 25th International Conference on Pattern Recognition (ICPR), 5543–50, 2021.
<https://doi.org/10.1109/ICPR48806.2021.9413279>.
- Hui, Haifeng. "What Can Digital Humanities Do for Literary Adaptation Studies: Distant Reading of Children's Editions of Robinson Crusoe." *Digital Scholarship in the Humanities* 38, no. 4 (November 30, 2023): 1564–76.
<https://doi.org/10.1093/llc/fqad059>.
- Nguyen, Quoc-Dung, Duc-Anh Le, Nguyet-Minh Phan, and Ivan Zelinka. "OCR Error Correction Using Correction Patterns and Self-Organizing Migrating Algorithm." *Pattern Analysis and Applications* 24, no. 2 (May 1, 2021): 701–21.
<https://doi.org/10.1007/s10044-020-00936-y>.
- Scancella, John. "Better OCR for Newspapers." *Medium* (blog), August 20, 2020.
<https://medium.com/@blacksmithforlife/better-ocr-for-newspapers-c7c1e2788b7a>.
- Segev, Elad. "Textual Network Analysis: Detecting Prevailing Themes and Biases in International News and Social Media." *Sociology Compass* 14, no. 4 (2020): e12779.
<https://doi.org/10.1111/soc4.12779>.
- Stabile, Carol A. *White Victims, Black Villains: Gender, Race, And Crime News in US Culture*. 1st edition. New York ; London: Routledge, 2006.
- The Detective*. "Advertisement." September 1901, 189 edition.
- The Detective*. "Wanted Ad." September 1901, 189 edition.
- Tong, Xiang, and David A. Evans. "A Statistical Approach to Automatic OCR Error Correction in Context." In *Fourth Workshop on Very Large Corpora*, edited by Donia Scott. Herstmonceux Castle, Sussex, UK: Association for Computational Linguistics, 1996.
<https://aclanthology.org/W96-0108>.
- Viola, Lorella, and Jaap Verheul. "Mining Ethnicity: Discourse-Driven Topic Modelling of Immigrant Discourses in the USA, 1898–1920." *Digital Scholarship in the Humanities* 35, no. 4 (December 1, 2020): 921–43. <https://doi.org/10.1093/llc/fqz068>.