The Journal of Open Source Software

Handwritten Analog Tabular Text-Recognition and Input Checker (HATTRIC) Pipeline

Eduard Faisal Steiner [1] **and Bryan Robert Carlson** [2]

1 Department of Crop and Soil Sciences, Washington State University, Pullman, Washington, USA **2** US Department of Agriculture–Agricultural Research Service, Northwest Sustainable Agroecosystems Research Unit, Pullman, Washington, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 26 September 2025
Published: unpublished

Summary

HATTRIC is a hybrid optical character recognition (OCR) pipeline designed to process digital archives of handwritten data tables. It integrates image processing, text extraction, and post-processing steps to improve OCR accuracy and efficiency. The pipeline first segments the image by defining grid lines corresponding to table rows and columns. It then breaks the image down into individual tabular cells for text extraction. Each cell is processed using the Google Vision API (Google Cloud, 2025) to extract text and output it into comma-separated value (CSV) format. Finally, users can review the extracted CSV data with an interactive error-checking user interface that flags potential issues based on outlier detection and user-defined minimum and maximum bounds.

Statement of Need

Accurately digitizing tabular data from analog sources remains a significant challenge. While many optical character recognition (OCR) tools exist, few perform with high accuracy on messy or irregular tabular data (Ramesh et al., 2024), especially when handwriting and inconsistent structure introduce noise and distortions (Placeholder, 2025). This often results in data being incomplete or inaccurate after extraction, limiting its usability.

At the same time, there is a vast amount of valuable scientific data locked in physical documents and digital archives. Many scientific records exist in tabular form, such as weather observation logs, experimental results in laboratory notebooks, and astronomical observation logs. These are critical for research and analysis but are not machine-readable and remain inaccessible to modern data processing tools. Manual transcription is often used as a robust and reliable approach, but it is labor-intensive, error-prone, and costly (Placeholder, 2024).

HATTRIC addresses the gap between manual transcription and full OCR by providing a hybrid approach tailored specifically for handwritten and analog tabular text recognition. It integrates manual image segmentation, automated OCR, and interactive error checking to allow efficient, customizable, and reliable digitization of complex tabular documents. By facilitating the conversion of analog tabular data into machine-readable formats, HATTRIC supports preservation, analysis, and reuse of important datasets that would otherwise remain difficult to access. The pipeline outputs extracted data in standard CSV format, ensuring easy integration with downstream data processing, visualization, and analysis tools.

The Journal of Open Source Software

Functionality Overview

The HATTRIC pipeline consists of three main phases: image selection and segmentation, text extraction, and input checking.

Segmentation

Users begin by selecting the "Start Segmentation" button on the chosen image. They then define grid lines on the scanned tabular image to segment it into individual cells. The process requires users to add row gridlines first, followed by column gridlines. Gridlines are added by left-clicking on the image, with right-click undoing the last added line. Users can enter a rotate mode by pressing the `R` key, allowing rotation of the image clockwise (`R`) or counterclockwise (`L`) to correct alignment issues. Pressing `Enter` exits rotate mode. Once row gridlines are set, pressing any key moves to column gridline definition. After defining column gridlines, pressing any key finishes segmentation. The image is then broken down into singular tabular cells for further processing.

Text Extraction

Upon completing segmentation, users can initiate OCR by clicking the "Run OCR" button, which sends each segmented cell image to the Google Vision API (Google Cloud, 2025) for text extraction. The extracted text is aggregated and output into a CSV table, representing the tabular data in a machine-readable format.

Input Checking

Users launch the error checker via a dedicated button, opening a new window to select the CSV file for review. The interface displays the CSV contents alongside the corresponding segmented image cell, allowing users to step through the table and manually correct errors in real time. The tool flags potential errors based on criteria such as missing values (NaN), statistical outliers, and customizable minimum and maximum bounds. Users can save their progress at any time by clicking the "Save Now" button.

Acknowledgements

This research was a contribution from the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture.

References

Google Cloud. (2025). *Google cloud vision API client library for python*. https://github.com/googleapis/python-vision.

Placeholder, A. (2024). Manual transcription in historical datasets: Labor, accuracy, and cost tradeoffs. *Historical Language and Corpus Studies*. https://hlcs.nl/article/view/15456

Placeholder, A. (2025). Handwritten text line extraction in historical documents using hybrid methods. *International Journal on Document Analysis and Recognition (IJDAR)*. https://doi.org/10.1007/s10032-025-00543-9

Ramesh, P., Bao, L., Lee, J., Ahmed, K., & Yang, Y. (2024). RAGTable: Retrieval-augmented table recognition with table structure memory. *arXiv Preprint arXiv:2404.10305*. https://arxiv.org/abs/2404.10305