Optical Character Recognition for Group Quarters Residents in the 1990 U.S. Census

Douglas E. Rosin, David A. Bleckley, and J. Trent Alexander

https://censusdigitization.org

info@censusdigitization.org

Corresponding author:
rosind@umich.edu

INTRODUCTION

The 1990 Decennial Census "group quarters" (GQ) records document individuals living in communal settings like dorms, prisons, and hospitals. Originally stored as 65,000 microfiche images of condensed font printouts, these records cover about 6.6 million people across 142,000 GQs. We digitized the microfiche and extracted the data into linked data tables, capturing person- and GQ-level data.

Optical character recognition (OCR) formed the backbone of our digitization, using Tesseract OCR, a long short-term memory neural network-based engine. To overcome the lack of character and line spacing in these images, we developed a method that uses pixel density to deconstruct the image into isolated data fields within the table and improve OCR accuracy.

We add to the literature on historical document processing. Open-source OCR software applying deep learning methods to simplify document image analysis (Shen et al. 2021) has recently been used to successfully process historical newspaper records (Dell et al. 2024) and extract structured administrative data from government records (Lafia et al. 2023). Our approach is also similar to the work of Dahl and associates (2025), who locate a table grid's cells by identifying the intersections of horizontal and vertical lines.

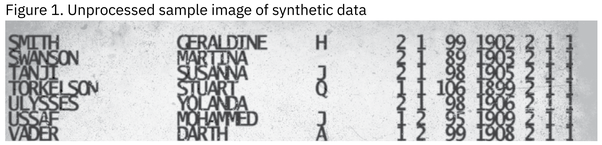
METHODS

GQ Record Data Description

The GQ records of the 1990 U.S. Census are semi-regularly structured tabular data. While the general format of each GQ record is the same, there is substantial variation in data field length, number of persons, and the presence of supplemental, unformatted rows within the records. Each scanned GQ image consists of a page header over two columns of data, containing one or more GQ records. Each GQ record begins with a header that contains GQ-level data. Person-level records have 11 columns of individual data fields. Records are tabular and fixed width, with whitespace separating columns. Supplemental rows, providing respondent-filled race and ethnicity responses, disrupt the tabular structure, creating a critical challenge to digitizing and parsing the data.

For this paper, we use a sample image of synthetic data (Figure 1), created to portray the challenging nature of the GQ image but not based on actual census responses. The methods and workflow applied remain consistent, but parameter values and results may differ from the real GQ data.

Figure 1. Unprocessed sample image of synthetic data

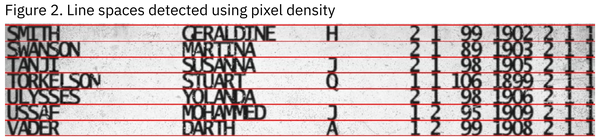


Digitizing the GQ Records

We digitized grayscale TIF images of the GQ records using a five-phase process:

1. **Initial preparation:** We corrected the images for skew and rotation, converted them into binary numpy arrays, and cropped excess whitespace. This phase optimizes the image for segmentation.
2. **Segmentation:** We separated the page header from the body and divided the body into two columns. For each section, we detected densities of black pixels. We used the densities to separate GQs and detect and divide GQ-level data from person-level records.
3. **Row detection and separation:** We first applied preprocessing steps (thresholding and erosion) to eliminate black pixels from characters, enhancing the minimal space between rows. We then measured pixel density to detect line spacing (Figure 2). We split the original image into rows and padded each with white pixels. We used pixel density to determine if each row was a main data row or a supplemental row.
4. **Image preprocessing:** We applied a second round of preprocessing steps (thresholding and closing), with a goal of increasing the accuracy of the OCR process (Figure 3).
5. **Optical Character Recognition:** We split each row into individual cells, facilitated by the records' fixed-width layout. We created custom Tesseract configurations for each field in a record, with dictionaries, patterns, and whitelists/blacklists for different field types. We also designated the Page Segmentation Mode (PSM) and OCR Engine Mode (OEM) for each field, which tell Tesseract what type of text to digitize and whether to use neural networks or traditional models, respectively. We compiled the OCR output into a database.

Figure 2. Line spaces detected using pixel density



Optimization of Parameters

We optimized model accuracy by tuning three key areas: image preprocessing, line pixel density threshold, and Tesseract OCR configurations. Since these parameters interact (e.g., increasing the threshold for black pixels makes the text wider, which can increase readability, but also decreases character space between words), we used a multivariate approach. Accuracy was measured using character error rate (CER), which accounts for substitutions, insertions, and deletions (Neudecker, et al. 2021).

Figure 3. Individual rows of text preprocessed for OCR

SMITH	GERALDINE	H	2	1	99	1902	2	1 1
SWANSON	MARTINA		2	1	89	1903	2	1 1
TANJI	SUSANNA	J	2	1	98	1905	2	1 1
TORKELSON	STUART	0	1	1	106	1899	2	1 1

RESULTS

Table 1 shows CER results. With optimal settings, CER drops below 1%. While a base Tesseract model has a CER of 45-65%, adding whitelist and blacklists to the Tesseract configuration cuts the error rate to about 5%. While the newer neural network model's accuracy is high, **using the original Tesseract engine (OEM 0) along with dictionary and pattern files yields optimal results, reaching 0.2% CER.**

Table 1. CER Results for different parameters in our model

	Kernel: 3x2 Threshold: 85			Kernel: 4x1 Threshold: 80			Kernel: 3x2 Threshold: 80		
Line pixel density threshold	0.05	0.07	0.09	0.05	0.07	0.09	0.05	0.07	0.09
Base Config	56.6	56.8	61.6	56.8	58.5	58.9	44.6	56.8	65.0
PSM-only	8.8	9.8	34.4	10.3	9.8	9.3	13.3	10.5	47.1
PSM+ Whitelist+ Blacklist	5.0	5.7	32.0	6.4	6.9	6.0	10.9	6.4	43.7
PSM+ Whitelist+ Blacklist+ OEM 0+ Lists & Patterns	0.2	1.0	26.5	1.0	1.4	0.7	7.8	1.0	38.8

Note: These accuracy numbers (character error rate) are derived from the test image that we construct to resemble a GQ census image. PSM (Page Segmentation Mode) refers to how Tesseract assumes the layout of the text.

Figure 4. Relationship between line pixel density threshold and character error rate

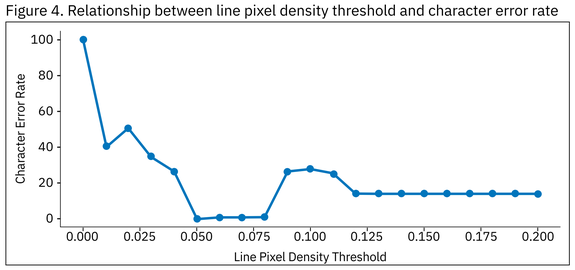


Figure 4 shows that a line pixel density threshold of 0.05 yields optimal row detection and lowest error. Accuracy drops if the threshold is too low (missed rows) or too high (too many false rows). The graph plateaus because we hard-coded row separations if detected rows were too close. For preprocessing, **higher thresholds combined with thicker kernels perform best,** improving text retention and row separation for OCR.

DISCUSSION

Our research offers two main contributions:

❶ We introduce a novel pixel density method to segment dense, whitespace-divided tabular data—useful for any tables lacking clear ruling lines.

❷ We identify optimal OCR settings for irregular data using Tesseract; older models with dictionary and pattern files yield better accuracy for single characters and numbers than neural network models designed for general text.

Finally, it is important to note that comprehensive data documentation is key for building effective patterns and dictionaries and understanding table structure (Bleckley et al. Forthcoming)

WORKS CITED

Bleckley, D.A., Sullivan, J., Rosin, D. Forthcoming. "Using U.S. Census Bureau Documentation to Revisit and Rebuild 20th Century Data Complexities." Social Science History Association 2025 Annual Meeting, Chicago, Illinois.

Dahl, C.M., S. Hwang, T. Johansen, and M. Squires. 2025, August 6-7. "Improving Historical Census Transcriptions: A Machine Learning Approach." 2025 Data-Intensive Research Conference, Minneapolis, Minnesota.

Dell, M., J. Carlson, T. Bryan et al. 2024. "American stories: A large-scale structured text dataset of historical us newspapers," Advances in Neural Information Processing Systems, 36.

Lafia, S., D.A. Bleckley, J.T. Alexander. 2023. "Digitizing and parsing semi-structured historical administrative documents from the G.I. Bill mortgage guarantee program." *Journal of Documentation* 79(7): 225-239.

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., Pletschacher, S. 2021. "A survey of OCR evaluation tools and metrics." *The 6th International Workshop on Historical Document Imaging and Processing,* 13–18. https://doi.org/10.1145/3476887.3476888

Shen, Z., R. Zhang, M. Dell, B.C.G. Lee, J. Carlson, and W. Li. 2021. "Layout Parser: A Unified Toolkit for Deep Learning Based Document Image Analysis."

U.S. Bureau of the Census. 1996. 1990 *Census of Population and Housing—History CPH-R-2.* Washington, D.C.: U.S. Government Printing Office.

Tesseract OCR. 2025. Tesseract User Manual. https://github.com/tesseract-ocr/tessdoc

ACKNOWLEDGMENTS

This project was undertaken as part of the Decennial Census Linkage Project, funded by a U.S. National Science Foundation Human Networks and Data Science - Infrastructure grant (Award Number 2023639).

We are grateful to Jonathan Morgan for his consultation and technical support in bringing this project to production and to the Junior Professional Researcher Program for providing me with this opportunity to pursue a career in research.