Reading the Scale Bar: A Computer Vision Pipeline for Automated Metadata Extraction from Historical Maps

Maxwell Donaldson [1]*

[1] Department of Geography/Geology, University of Nebraska at Omaha, Omaha, United States
* [mrdonaldson@unomaha.edu]

Keywords: computer vision, historical maps, Library of Congress, object detection, scale bars, vision-language models, YOLO

The Library of Congress (LOC) managed 21 petabytes of digital content, comprising 914 million unique files in 2022 with over 5.5 million of this data being cartographic material that has been digitized and made accessible through the International Image Interoperability Framework (IIIF). Maps carry a plethora of important spatial metadata, however, scale metadata remains largely uncaptured in the existing catalogue records. Scale values are one of the most prominent elements on a map and can contain important spatial and historical context. Manual extraction of map scale at the LOC collection size is unrealistic, however, this metadata can be essential for enabling new spatial analysis and more insight for historical research. AI-driven tools are increasingly being incorporated into cartographic workflows, but how accurately can automated systems interpret the various, ambiguous visual elements of historical maps and where does human expertise remain? This research examines a multi-stage, end-to-end computer vision pipeline for automatically detecting, extracting, and interpreting scale bars on historical maps from the Library of Congress, and evaluates its accuracy against a manually created dataset. By finding scale bars accurately we are one step closer to capturing map scale metadata for historical maps.

Recent work has shown the capability of YOLO architectures for use in the detection of elements from historical documents (Junior et al., 2025) and the capacity of VLMs to interpret cartographic elements like the scale bar from map imagery (Ung et al., 2025). However, no existing work combines object detection with VLM interpretation in a multi-stage pipeline focusing on scale bar metadata extraction from historical maps.

The map scale metadata pipeline works in three stages. First, a pretrained object detection model from the YOLO family, trained on manually annotated map images, locates scale bars on full map images from the LOC collection. Second, the bounding box coordinates are extracted from the first stage and are used to request TIFF images cropped to the scale bar location. These images are pulled directly from the Library of Congress IIIF Image API to preserve the fine typographic detail that would be lost in other formats. Third, these cropped scale bar images are passed to a Qwen3-VL, which is a vision-language model (VLM). This model will interpret the visual and textual content of each scale bar and output structured metadata in JSON format. The metadata framework is predetermined and includes transcribed text, language identification, units, unit classification (historical or modern), modern equivalents, and unit range.

This work creates a novel dataset of approximately 1,000 Library of Congress map images with over 900 annotated scale bars, including manually created, corresponding scale bar metadata. The map dataset includes maps from multiple centuries containing diverse languages, and various cartographic traditions and design. This includes maps with multiple scale bars, images with no scale bars, and maps containing multiple scale bars often including several historical units (versts, sazhens, leagues). Automated scale bar metadata extraction shows clear promise through this pipeline. YOLOv11m locates scale bars accurately across map types, with performance being strongest on standardized maps (mAP50 0.992). Bounding box geometry is a limitation as crops of scale bars adjacent to other map elements carry unwanted content to the interpretation stage. Interpretation divides sharply between perception and inference fields of the metadata. Reading what is printed on the scale bar is largely reliable. However, it degrades upon converting historical units to modern equivalents and inferring which text should be treated as a caption. This can be seen across both prompt engineering methods; zero-shot and few-shot with few-shot demonstrating only slight improvement. On standardized cartography, interpretation reaches a level that could be substantial enough to significantly reduce manual metadata extraction. On heterogeneous maps it does not, and human involvement remains necessary where cartographic judgement is required.

References

Júnior, E. S. dos S., Paixão, T., & Alvarez, A. B. (2025). Comparative performance of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 for layout analysis of historical documents images. *Applied Sciences, 15*(6), 3164. https://doi.org/10.3390/app15063164

Ung, H. Q., Habault, G., Nishimura, Y., Niu, H., Legaspi, R., Oya, T., Kojima, R., Taya, M., Ono, C., Minamikawa, A., & Liu, Y. (2025). CartoMapQA: A fundamental benchmark dataset evaluating vision-language models on cartographic map understanding. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)*, 440–453. ACM. https://doi.org/10.1145/3748636.3762755