Reading Historical Maps: End-to-end Map Text Detection, Recognition, and Label Grouping

Dimosthenes Karatzas and Jerod Weinman

March 4, 2024

Abstract

Localizing and recognizing text on historical maps presents several difficulties: widely spaced and overlapping words, complex text-like backgrounds, and strongly rotated or curved text. This project continues to build on robust reading models with additional features tailored to the map reading task and will develop a jointly-trained neural method for grouping words into label phrases.

1 Task Background

A brand new benchmark data set created for the ICDAR 2024 Competition on Historical Map Text Detection, Recognition, and Linking has created a fresh research opportunity. Extending a long line of robust reading competitions that ask systems to localize and recognize text amid challenging contexts, the new MapText competition targets a wide range of historical maps. In addition to requiring end-to-end localization and recognition of words on the map, this challenge follows the recent ICDAR 2023 Competition on Hierarchical Text Detection and Recognition in requiring systems to assemble words into higher-level semanatic groups. In this case, the phrase labels. Figure 1 illustrates a subset of words and links between them that must be extracted.

Training data sets curated for the challenge include 700 images with 80,000 words and 11,000 positive links. Test data sets will be even larger. Synthetic data may also be available.

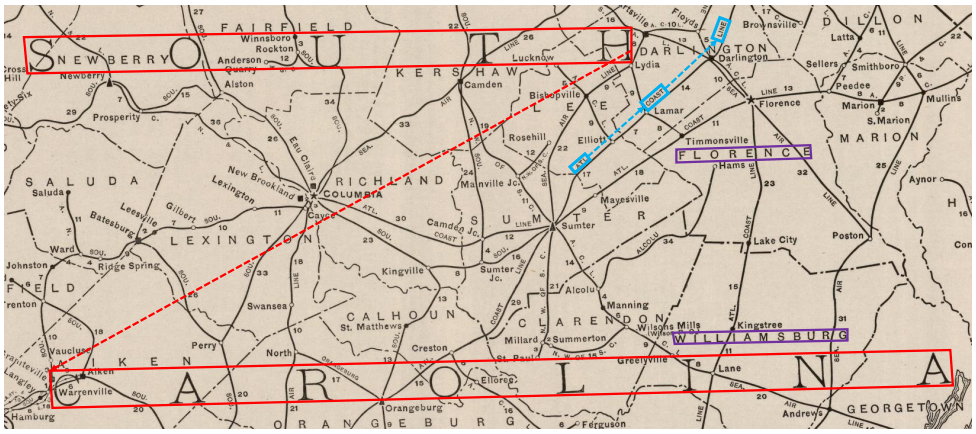

Figure 1: Word detection and linking (red, cyan) and non-linking (purple) examples. Image credit: Rumsey Collection [1] Image 5028.054 (Rand McNally and Company, *South Carolina*, 1924).

2 Proposed Work

The Detection Transformer Model (DETR) [2] and its derivative variant the deformable DETR [5] have become well-established tools anchoring many important computer vision tasks related to object detection and segmentation.

Recent work in end-to-end text spotting has also been building upon the basic structure of these models. In particular, we highlight the TESTR model [4] shown in Figure 2. Whereas the original DETR models have a single decoder to process a set of "atomic" object queries with prediction heads that give the category and location of each object detected, TESTR uses dual decoders with cross-attention to predict both the tightly cropped location of a word and the constituent "sub-atomic" characters. This tailored approach gives state of the art performance on scene text benchmarks with significant proportions of curved text. Recent work by additional collaborators (under review) extends TESTR, resulting in even better performance on historical maps.

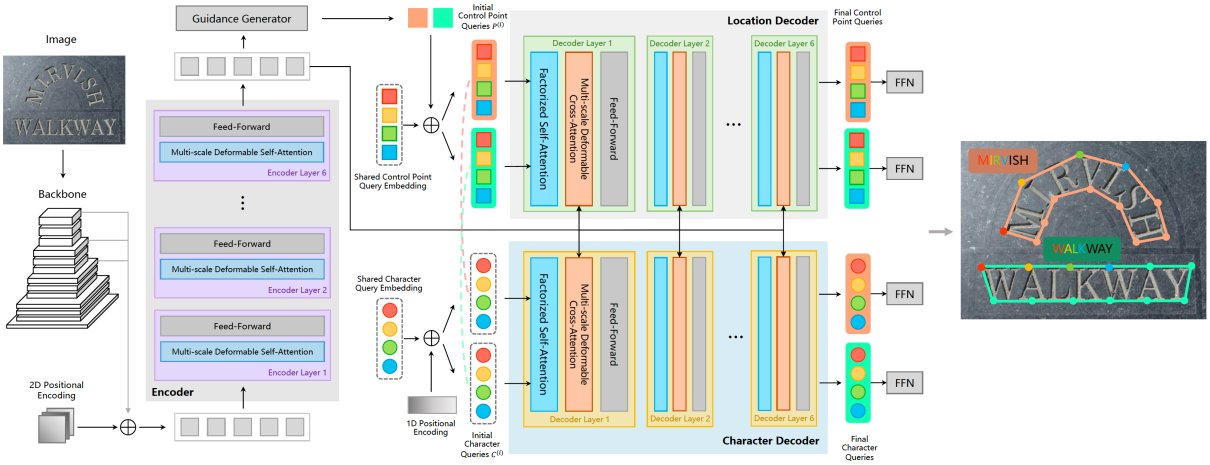

Figure 2: Overview of the TESTR model [4]: The guidance generator produces bounding box proposals which anchor deformable detection transformer [5] attention mechanisms for joint boundary control point and character prediction. (Figure from Zhang et al. [4].)

This project will have the opportunity to further extend TESTR derivatives to improve its performance on historical map text spotting.

In particular, one important open problem involves the linking of detected words into the underlying label phrase groups (cf. Figure 1. This is a special case of the image-to-graph problem, which has been addressed in a general way by the joint model Relationformer [3]. Also derived from the DETR family of models, the Relationformer inserts a relation query token and jointly decodes it with the other object query tokens. The relation prediction head then classifies the links between all pairs of detected objects. Factoring the tasks this way allows the objects (nodes) and relation (edges) to specialize their representations while collaborating on the graph generation process.

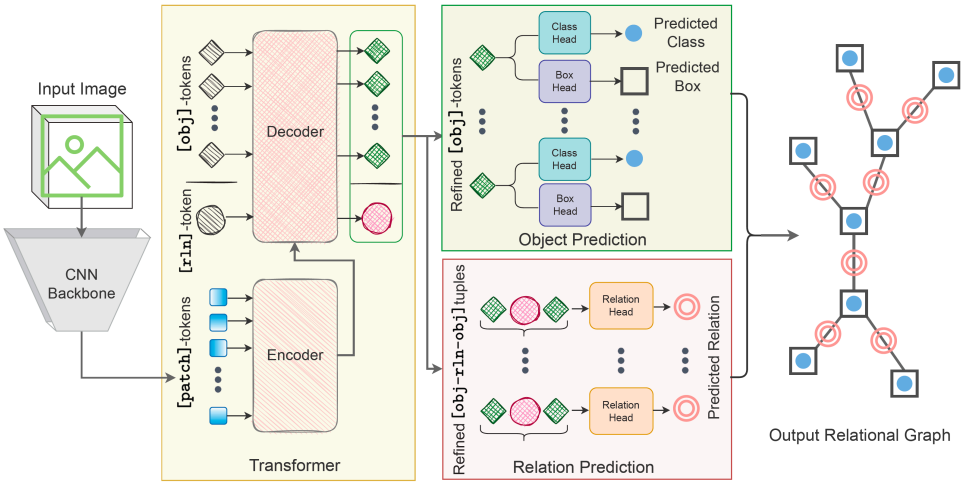

Figure 3: Overview of the Relationformer [3]: One additional relation query token `[rln]` is jointly decoded with the object detection tokens `[obj]` and used in a separate prediction head to identify relations among the detected objects. (Figure from Shit et al. [3].)

To address the map text linking problem jointly in an end-to-end fashion, we propose the master's project fuse the Relationformer's [3] additions to DETR, which enable its image-to-graph synthesis, with the dual text location/recognition decoders of a model like TESTR. By injecting an additional relation query into TESTR's character and location decoders, the model should be able to learn a "follows" relation that links words in reading order.

In addition to developing, training, and testing the model, we anticipate the need to implement a link-level evaluation tool that will provide more fine-grained feedback than the existing competition benchmark.

Excellent students might be given the possibility to continue for a PhD at the Computer Vision Center following the successful completion of this project.

References

[1] Cartography Associates. David rumsey map collection. https://www.davidrumsey.com.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[3] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, and Bjoern Menze. Relationformer: A unified framework for image-to-graph generation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 422–439, Cham, 2022. Springer Nature Switzerland.

[4] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9509–9518, 2022.

[5] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.