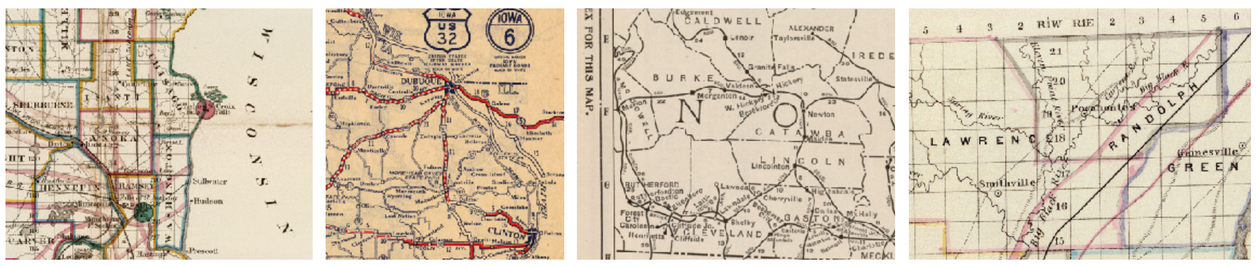
Deep Neural Networks for Text Detection and Recognition in Historical Maps

Grinnell College
Computer Science Department

Jerod Weinman, Ziwen Chen, Ben Gafford, Nathan Gifford, Abyaya Lamsal, and Liam Niehus-Staab

Problem and Motivation

Maps intertwine text and graphics, with text appearing in nearly any orientation, many sizes, and wide spacing.



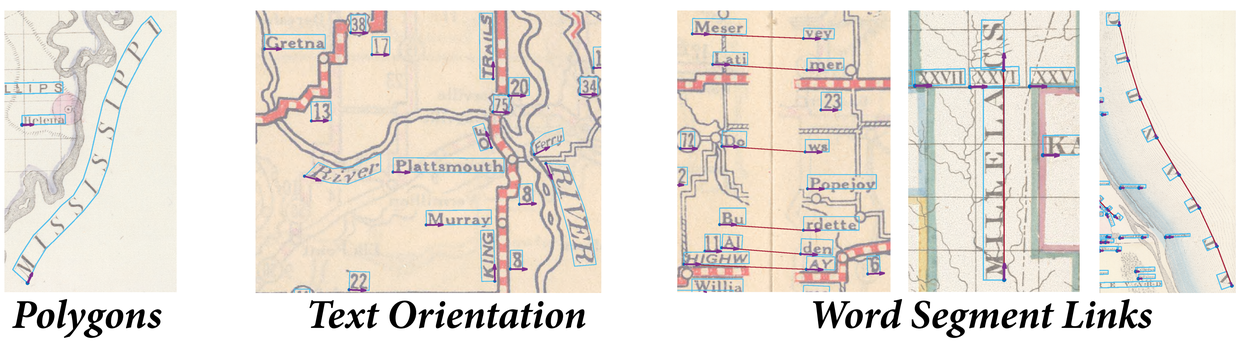
Adapting deep-learning methods from scene text, we tailor models to the map text detection and recognition problems.

Contributions

- Detects semantic baseline orientation
- Structures CNN+LSTM for robustness to graphical distractors
- Dynamic training data synthesis prevents overfitting

Data

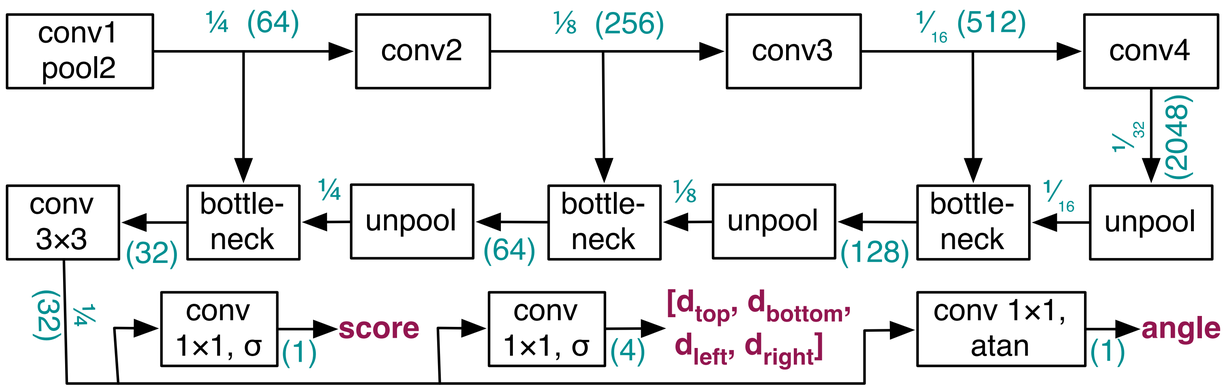
- 31 U.S. maps (city, multi-county, state/multi-state), 1866–1927
- 33,868 annotated bounding text polygons and transcriptions



Polygons ***Text Orientation*** ***Word Segment Links***

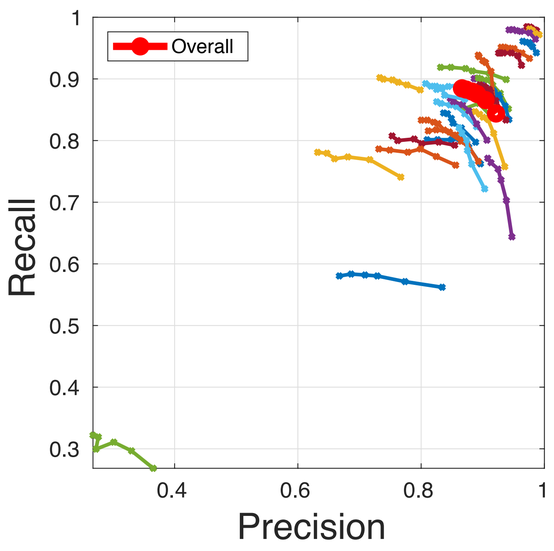
MapTD: Map Text Detection Network

MapTD is inspired by EAST (Zhou *et al.*, CVPR 2017) and shares similarities with the FOTS text detection branch (Liu *et al.*, CVPR 2018).

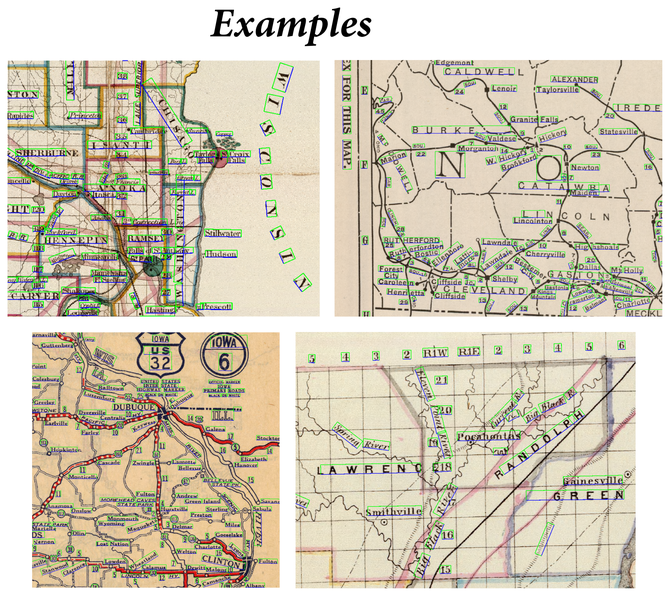


- Extracts feature maps using ResNet50 as the backbone network
- Repeatedly upsamples, concatenates, and applies bottleneck and feature fusion convolutions for each feature map scale
- Trained with ten-fold cross-validation using Dice loss (score), IoU loss (rotated rectangle), and cosine loss (angle)
- Predictions filtered with locality-aware NMS (Zhou *et al.*, CVPR 2017)
- Semantic baseline orientation increases F-score by 3.6%

Detection Results



Examples



Map Text Recognition Network

The recognition network modifies the layer dimensions and configurations of CRNN (Shi *et al.*, TPAMI 2017).

	Op	Krn Sz	Strd v,h	Out Dim	H	ΔW	Pad
0	Input			1	32		
1	Conv	3×3	1,1	64	30	−2	valid
	Conv	3×3	2,2	64	30		same
	Pool	2×2	1,1	64	15	⁄⁄2	valid
2	Conv	3×3	1,1	128	15		same
	Conv	3×3	1,1	128	15		same
	Pool	2×2	2,1	128	7	−1	valid
3	Conv	3×3	1,1	256	7		same
	Conv	3×3	2,1	256	7		same
	Pool	2×2	1,1	256	3	−1	valid
4	Conv	3×3	1,1	512	3		same
	Conv	3×3	1,1	512	3		same
	Pool	3×1	3,1	512	1		valid
5	Bi-LSTM			512	1		
	Bi-LSTM			512	1		
6	Output			62	1		

Avoids spurious edges

Deepens initial layers

Preserves width details

Eliminates 2×2×512 conv

15% fewer total convolution parameters

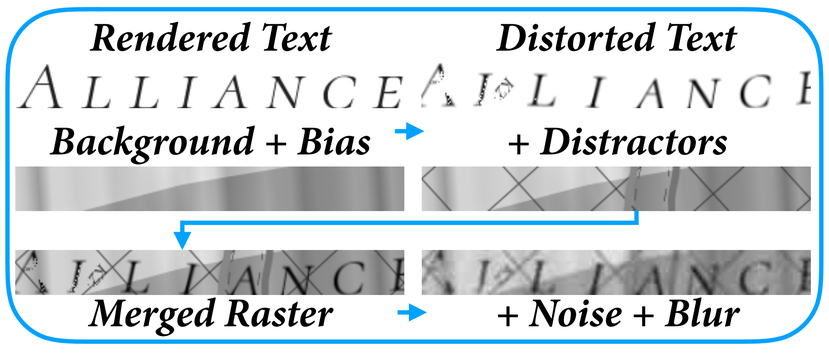
	Op	Krn Sz	Strd v,h	Out Dim	H	ΔW	Pad
	Input			1	32		
	Conv	3×3	1,1	64	32		same
	Pool	2×2	2,2	64	16	⁄⁄2	valid
	Conv	3×3	1,1	128	16		same
	Pool	2×2	2,2	128	8	⁄⁄2	valid
	Conv	3×3	1,1	256	8		same
	Conv	3×3	1,1	256	8		same
	Pool	2×2	2,1	256	4	−1	same
	Conv	3×3	1,1	512	4		same
	Conv	3×3	1,1	512	4		same
	Pool	2×2	2,1	512	2	−1	same
	Conv	2×2	1,1	512	1	−1	valid
	Bi-LSTM			256	1		
	Bi-LSTM			256	1		
	Output			36	1		

- Trained with CTC loss, adjusting step size and batch size
- 1.82% word error rate on MJSynth (Jaderberg *et al.*, IJCV 2016) is 14% lower than CRNN (case insensitive, closed lexicon)
- MPE predictions operate in open vocabulary mode while integrating a lexicon with beam search (Scheidl *et al.*, ICFHR 2018)

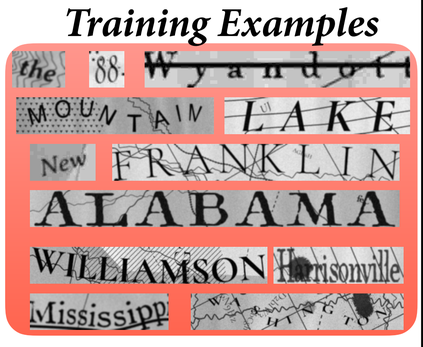
$$\hat{\mathbf{c}}_U \triangleq \arg\max_{\mathbf{c}} P(\mathbf{c} \mid \mathbf{x})$$
$$\hat{\mathbf{c}}_L \triangleq \arg\max_{\mathbf{c} \in L} P(\mathbf{c} \mid \mathbf{x})$$
$$\hat{\mathbf{c}} = \begin{cases} \hat{\mathbf{c}}_L & \lambda P(\hat{\mathbf{c}}_L \mid \mathbf{x}) > (1-\lambda) P(\hat{\mathbf{c}}_U \mid \mathbf{x}) \\ \hat{\mathbf{c}}_U & \text{otherwise.} \end{cases}$$

Dynamic Training Map Text Synthesis

To prevent overfitting to a small dataset with inappropriate features, we synthesize map-like training images on-the-fly.



Training Examples

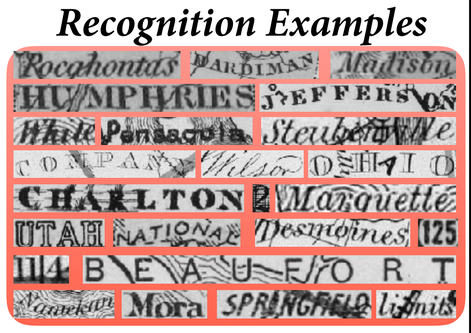


Experimental Results

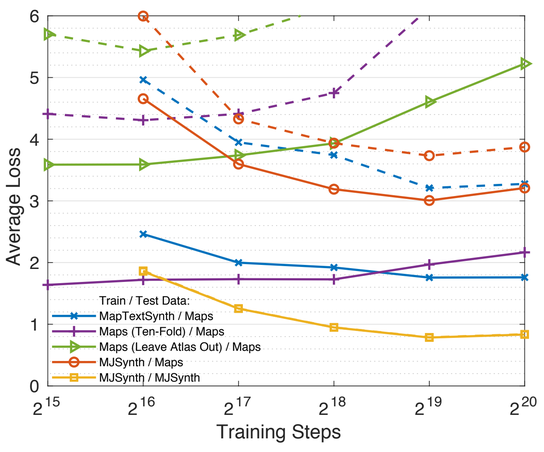
Case-Sensitive Cropped Word Results on Maps

Training Data	Character Error (%) Open	Closed	Mixed	Word Error (%) Open	Closed	Mixed
MJSynth	21.22	18.03	17.88	50.77	37.59	37.24
Maps (Atlas)	22.12	17.25	16.92	53.40	39.02	38.60
Maps (10x)	13.03	9.19	8.28	36.61	23.53	21.54
MapTextSynth	13.64	9.88	9.04	37.88	24.05	22.13

Recognition Examples



Cropped Word Recognition Validation and Test Results



End-to-End Results