Scaling-Up Whole-Book Recognition

Pingping Xiu and Henry S. Baird

Motivation

Unsupervised high-accuracy recognition of the textual contents of an entire book

Using internal evidences within a long, isogenous book, we can do automatically model adaptations to improve OCR rates.

Disagreements between independent models suggests improvements for models

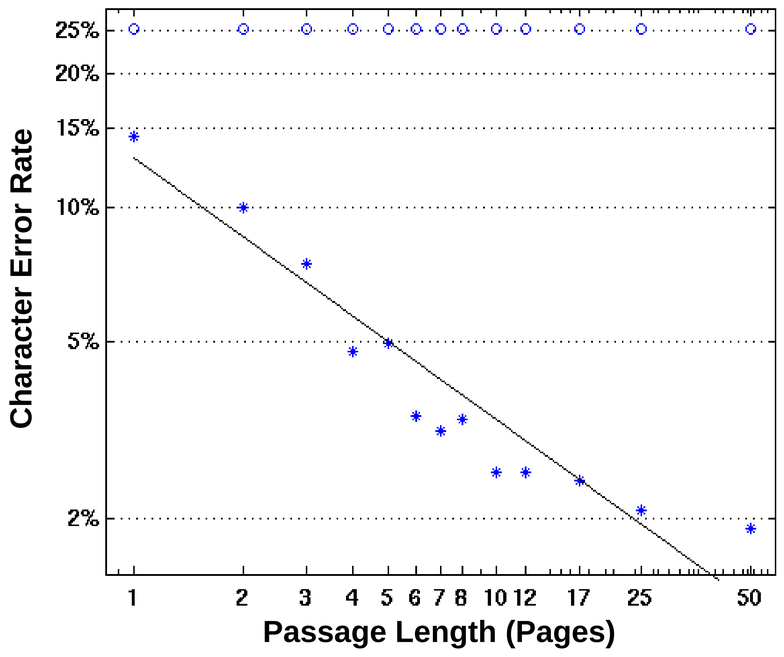
Any character classifier is inaccurate, and any dictionary is incomplete; the two models can cross-check and complement each other to get improvements on recognition accuracy.

Model Adaptation Guided by Mutual Entropy

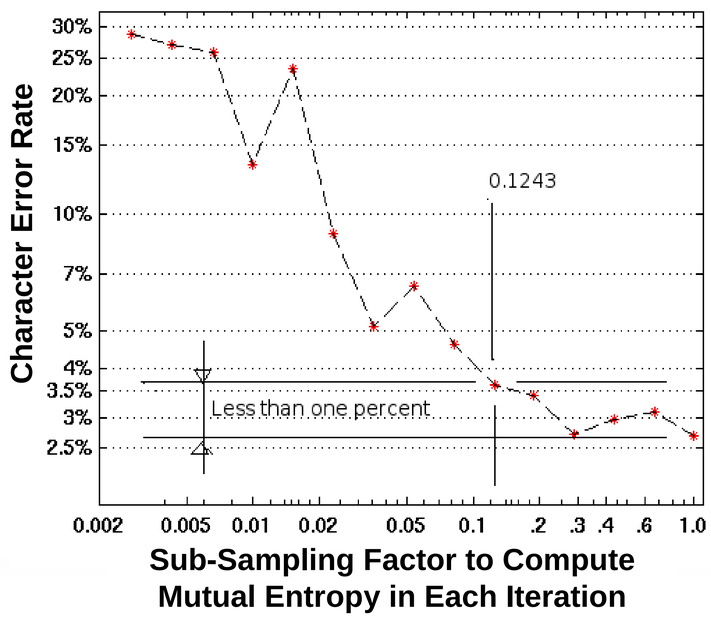
***A correct model adaptation will *presumably* lower the overall passage-level disagreement; the wrong one will *probably* increase it.**

The longer the book, the more confident the previous statement * will be.

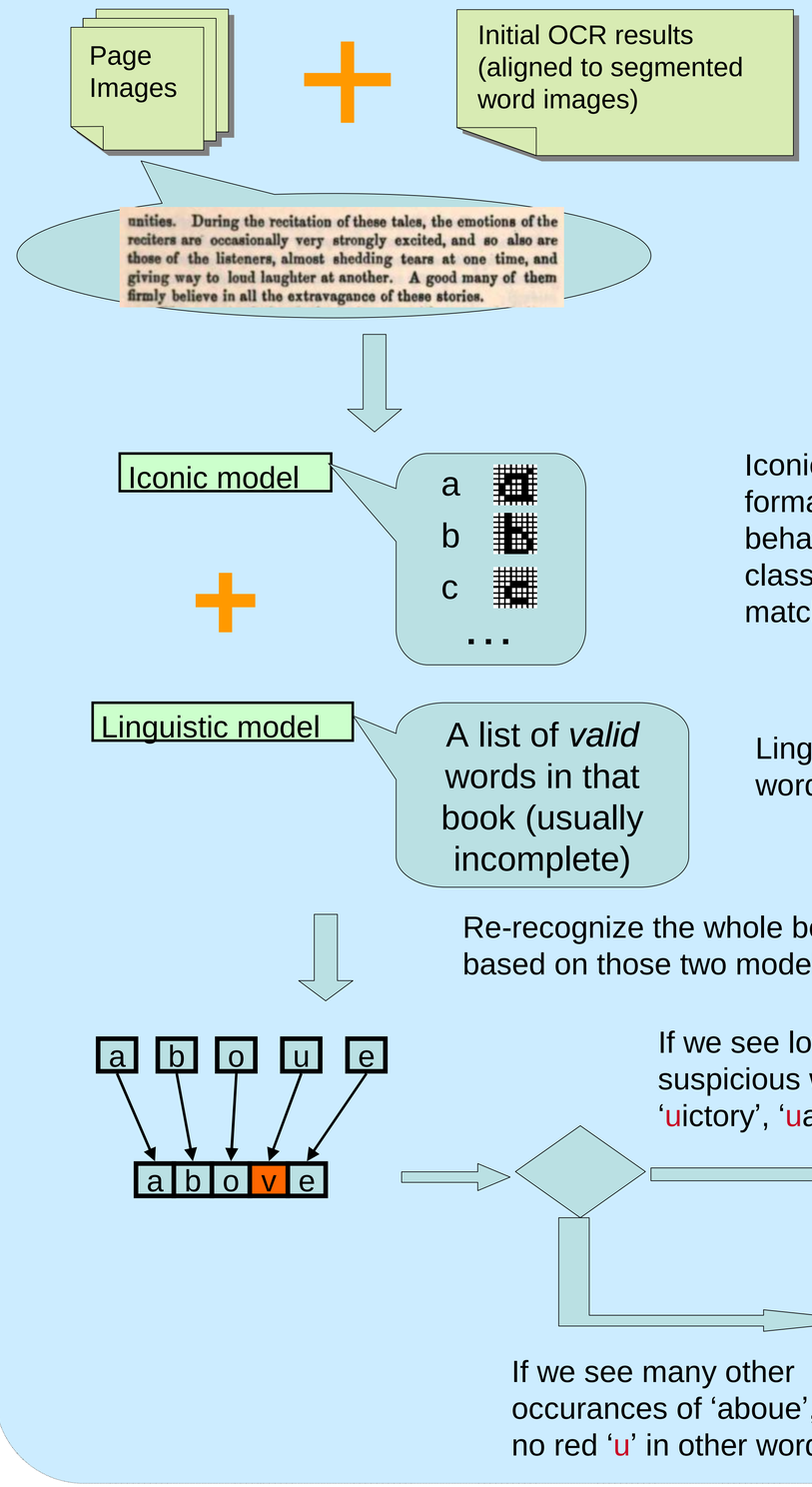
The longer the passage, the better our algorithm performs



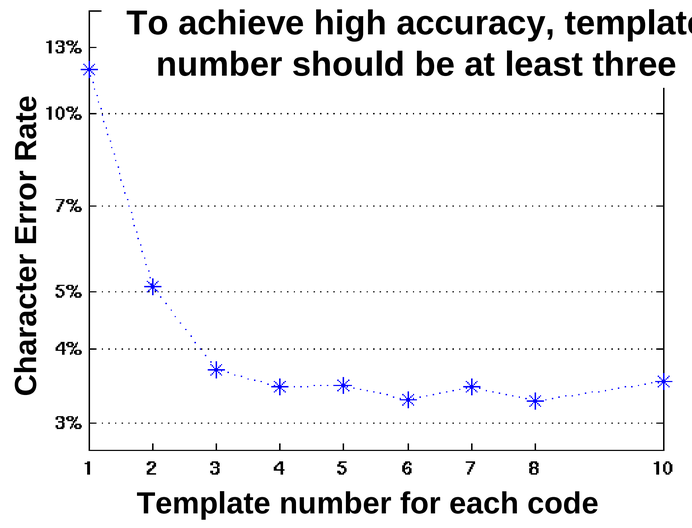
Our algorithm can be randomized so that it speeds up, e.g., by a factor of eight with little loss of accuracy



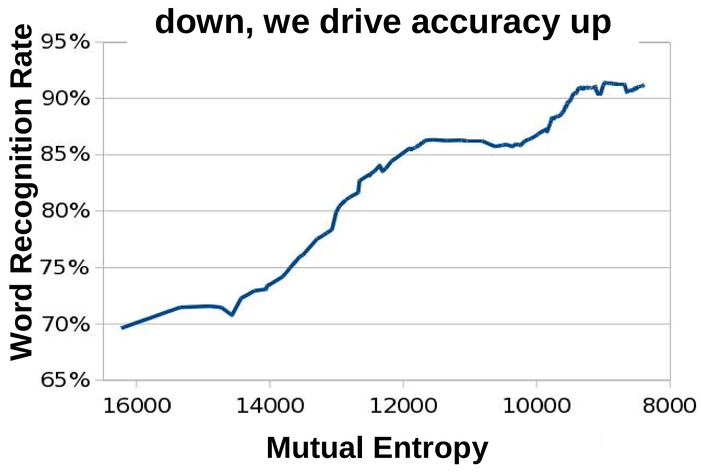
Task: Given a book's images and an initial buggy OCR transcript, derive two independent models and adapt those models to that book's images to get a better transcript.



To achieve high accuracy, template number should be at least three



By driving mutual entropy down, we drive accuracy up



Conclusions

It is highly encouraging that word accuracy and character accuracy improve *monotonically* as passage-length increases.

The algorithm can *tackle a much larger scale beyond 50 pages* through the randomization technique.

Passage-scale mutual entropy is strongly *negatively correlated* with accuracy.

Future Work

Scale-up experiments to an entire book.

Investigate different policies for applying corrections to the iconic model.

COMPUTER SCIENCE & ENGINEERING

LEHIGH UNIVERSITY

P.C. Rossin College of Engineering and Applied Science

Computer Science and Engineering CSE

Henry S. Baird & Daniel Lopresti
Pattern Recognition Research Lab