**Goal**

*Fully automatic improvement of a recognition system for the textual contents of an entire book*

Using internal evidence within a long and isogenous book, we can automatically adapt the recognizer to improve OCR accuracy.

**Disagreements among independent models can suggest improvements to each model**

If the character classifier is inaccurate, and the dictionary is incomplete, the two models can cross-check and complement each other to improve recognition accuracy.

**Strategy:** 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 improve the transcript.

- **Iconic model**
  - Describes image formation and determines the behavior of a character-image classifier. Here we use template matching as our classifier.
  - Suppose the iconic model suggests:
  - But the linguistic model prefers:
  - Then this disagreement suggests two possible corrections:

- **Linguistic model**
  - Describes word-occurrence probabilities or, more simply, merely a lexicon.
  - (1) To correct the iconic model
    - replace the label ‘u’ with ‘v’.
    - choose this if there are many suspicious words such as “vehicle”, “victory”, “vocabulary” that have a “u” with a high disagreement.
  - (2) To correct the linguistic model
    - add a new entry “aboue” to linguistic model.
    - choose this if there are many other occurrences of “aboue”.

**Future Work**

- Scale up to an entire book
- Compare competing policies for correcting the models
- Explore branch-and-bound strategies

**Experimental Results: A Monotonic Improvement**

Recognition rate as a function of passage length

![Graph showing recognition rate improvement](image)

Accuracy on 4-character and longer words, measured in a single page:

- Word Accuracy
  - Top-1 Word Hit Rate
  - Top-5 Word Hit Rate
  - Accuracy on 4-character or longer words

**Tests on Much Longer Passages**

An improved version of our algorithm can handle longer passages

- Cope with segmentation errors: one split, or one merge, per word.

**Algorithm Enhancements Described in the Paper**

Using multiple templates for each character label in iconic model.