Boosted Decision Trees for Word Recognition in Handwritten Document Retrieval

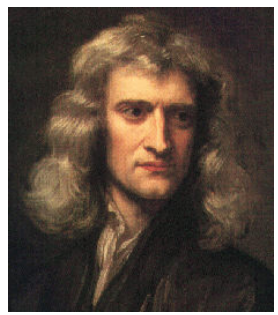
Nicholas R. Howe
Smith College

Toni M. Rath
R. Manmatha
University of Massachusetts
Center for Intelligent Information Retrieval

Inaccessible Treasures

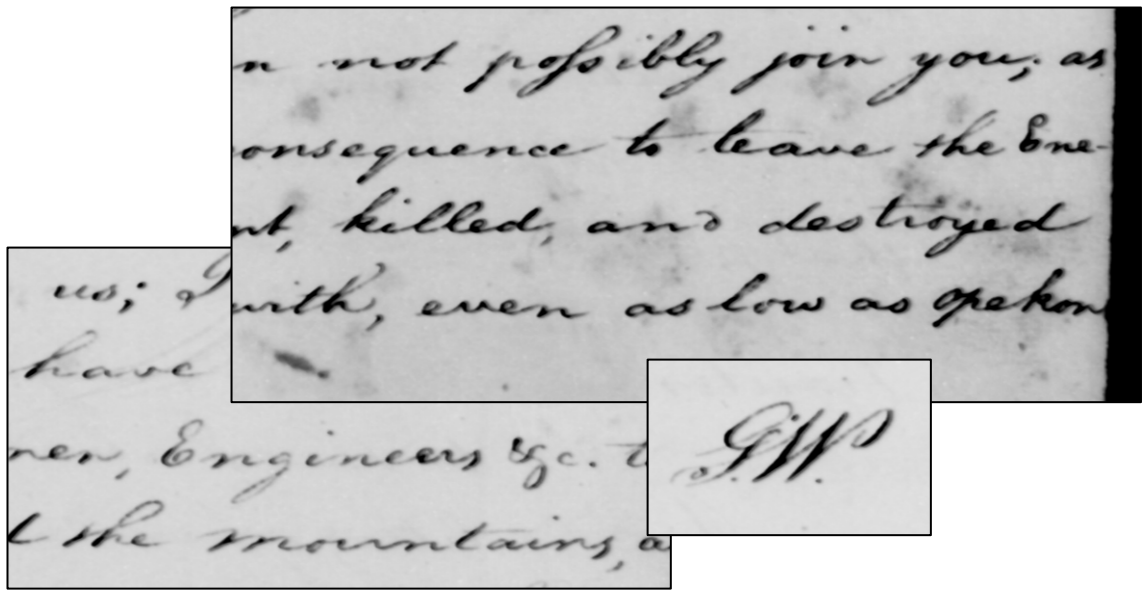
- Washington's letters: 140K pages
  - Scanning project complete ($$)
  - Transcription prohibitive ($$$)
  - Unprocessed format limits use
- Similar problem with other collections:
  - Isaac Newton's manuscripts
  - Scientific field notebooks

Goal: automated search/retrieval



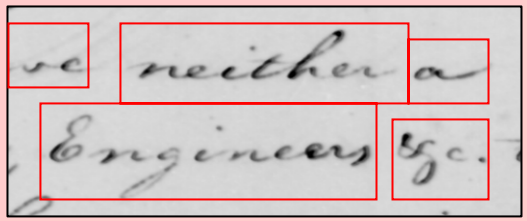
Challenges of Historical Documents

- Handwriting recognition: success in constrained domains
  - Postal addresses, bank checks, etc.
- Historical documents are much harder
  - Fewer constraints
  - Fading & stains
  - Hyphenation
  - Misspellings
  - Ink bleed
  - Slant
  - Ornament



Analysis Sequence

Segmentation



Segment words, not characters

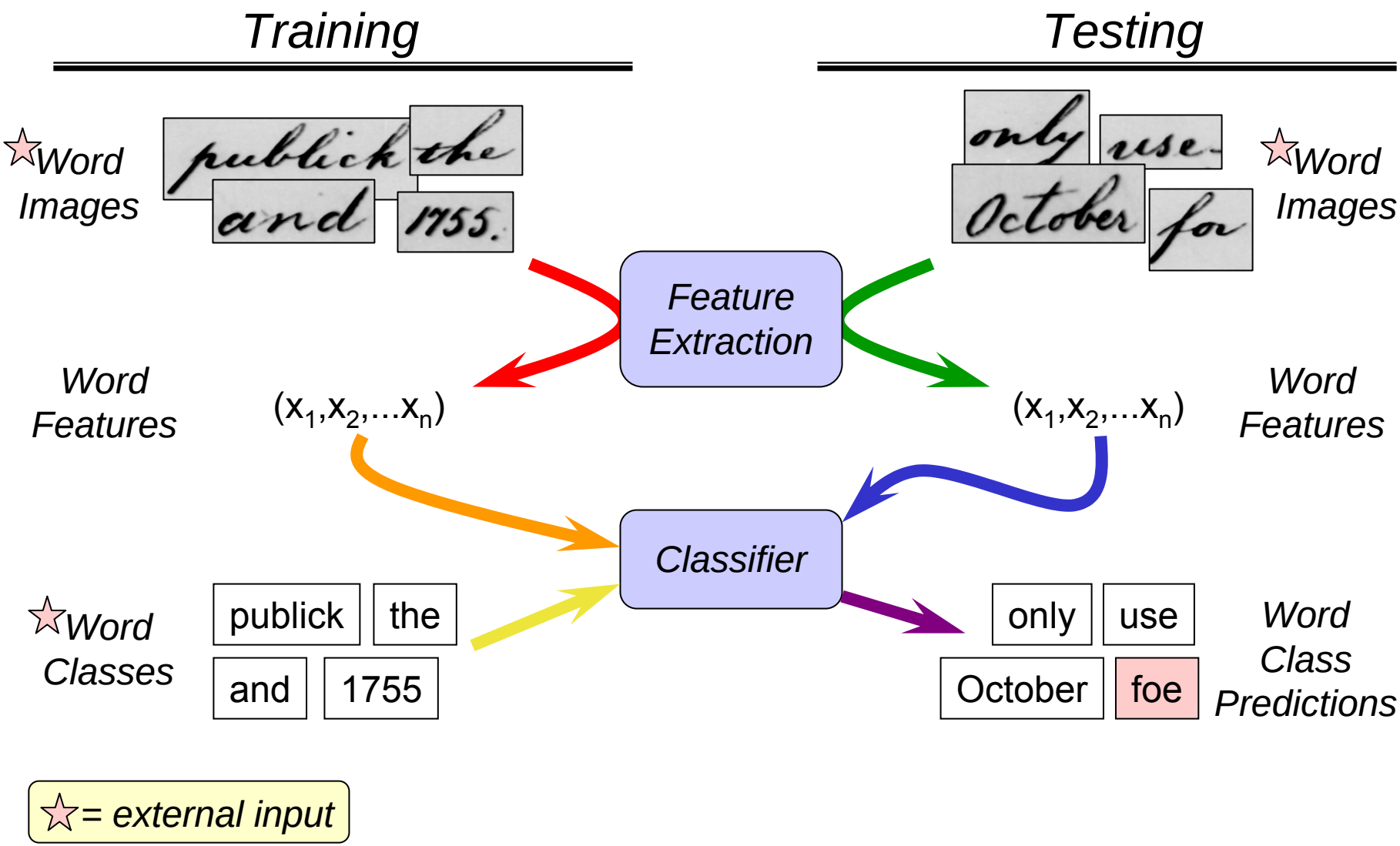
Recognition

Orders

Transcription & Retrieval

Focus of this talk

Recognition = Supervised Learning

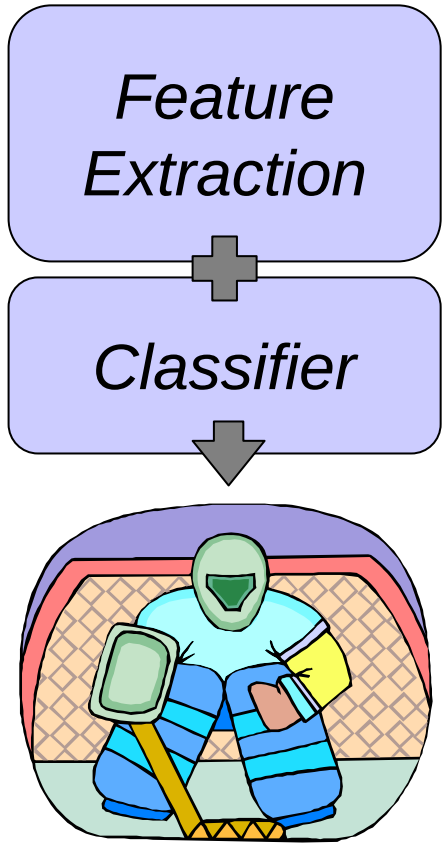


Recognition Game Plan

- Identify appropriate features
- Apply boosting classifier

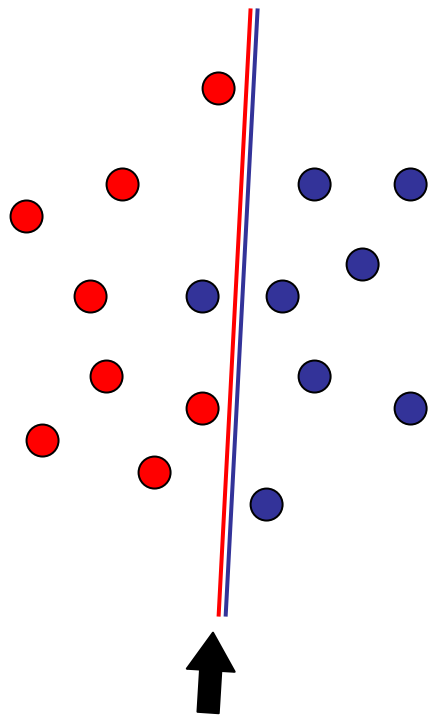
- Previous work (27 features):
  - 40% words correctly identified
  - 55% correct with language model

(Boosting on 27 features → 51%)



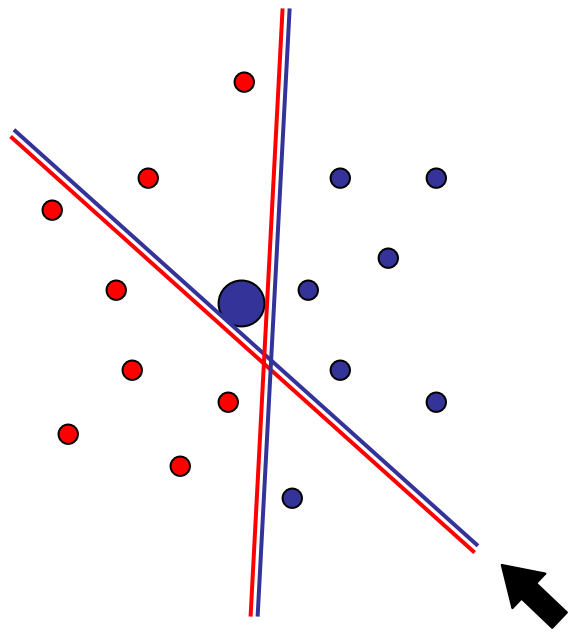
Example: Boosting

- Base rule must classify at least half of examples correctly.



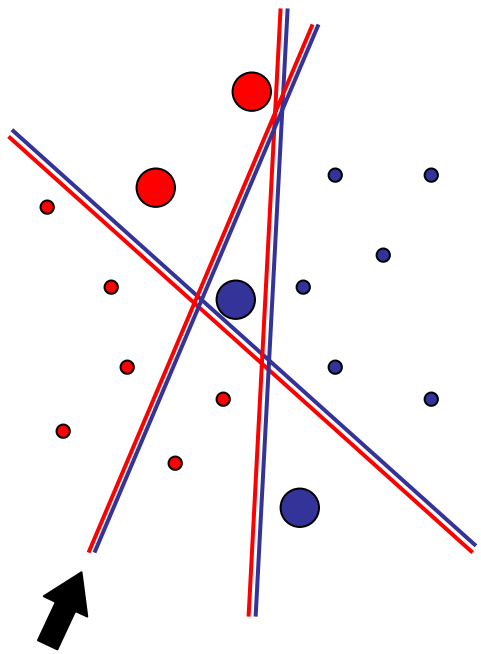
Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)



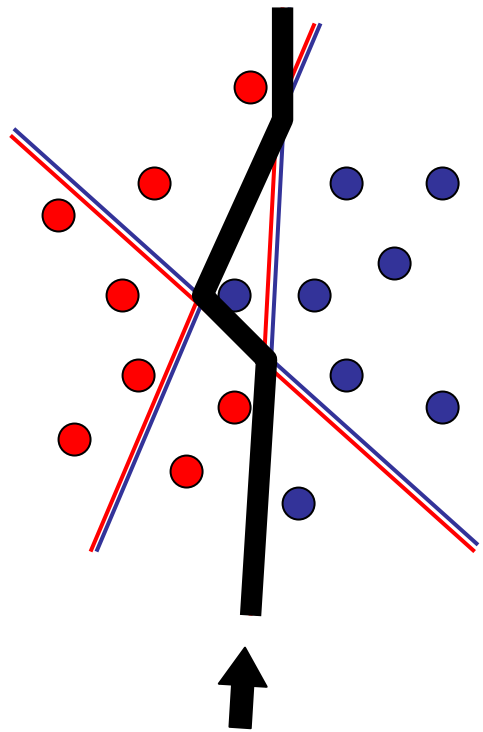
Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)
- Each new rule has different viewpoint



Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)
- Each new rule has different viewpoint
- Combined predictions are better than single classifier alone.

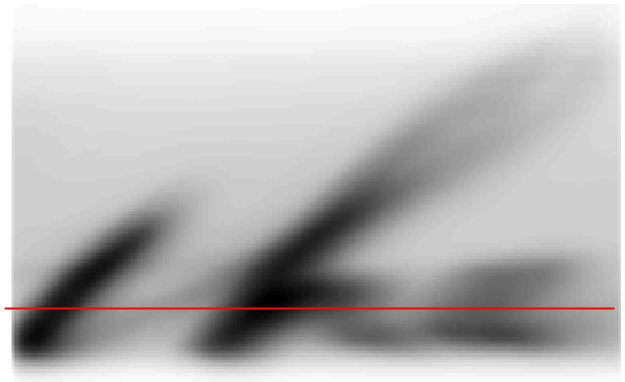


Boosting with Word Images

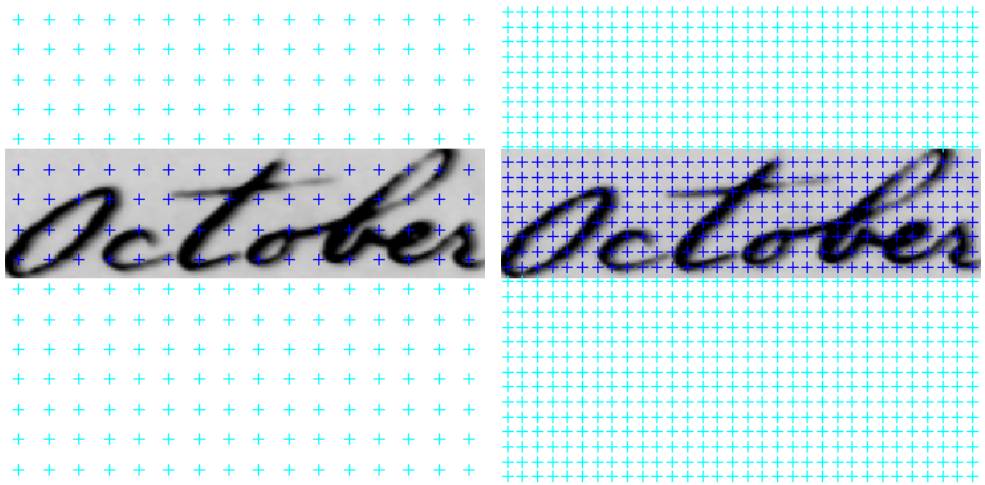
- Many more than two word categories
  - Must still make less than 50% error per step
  - Need sophisticated base classifier
- Feature choice will be important
  - Complex features $\Rightarrow$ extraction errors
  - Simple features $\Rightarrow$ less relevant individually
  - Boosting strength: wheat from chaff

Features: Spatial Samples

- Aligned words are spatially consistent
  - Images scaled & translated
  - Midline mapped to [0,1] interval
- Feature = sample at fixed point in aligned image

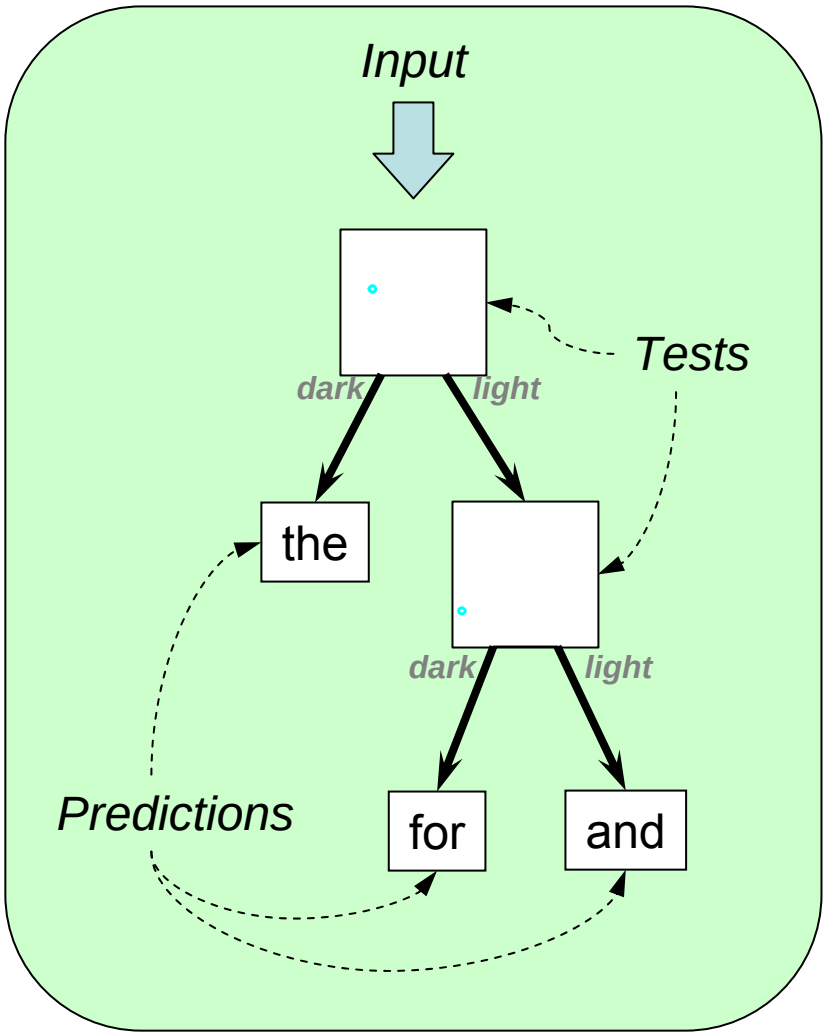


Superposition of 238 versions of 'the'

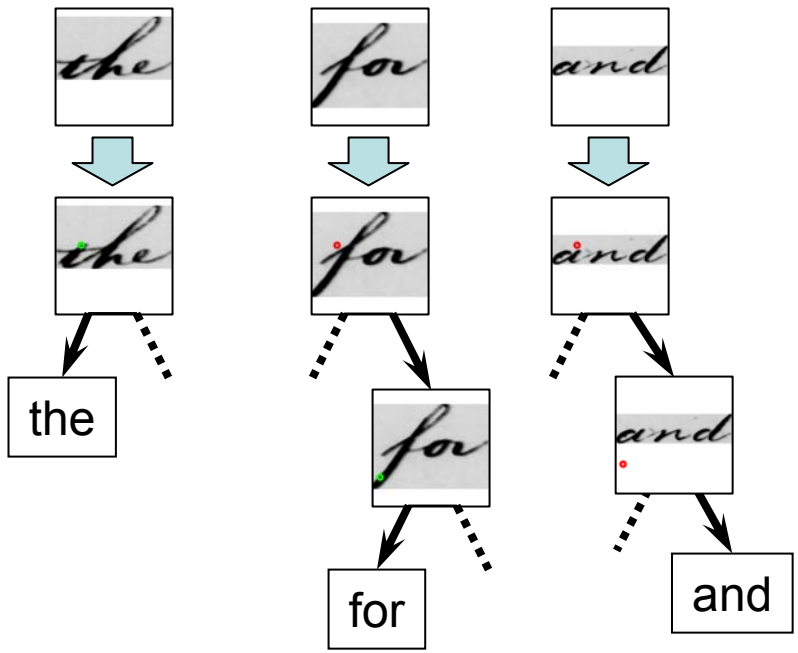


← Sample arrays with different densities

Base Classifier: Decision Trees



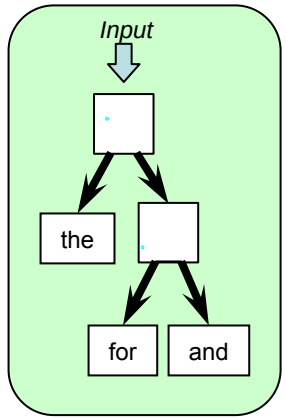
Sample Classifications



(Actual trees have ~2000 branches)

Building Decision Trees

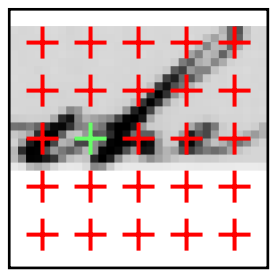
- How to choose good tests?
  - ☞ Exhaustive Search ☜



(80K candidate features)
x(5K images)
x(2K nodes per tree)
x(255 thresholds)
x(200 trees)

} = *too much searching!*

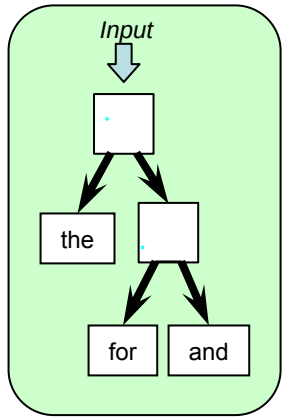
Solution: “Pyramid” search



Coarse Grid

Building Decision Trees

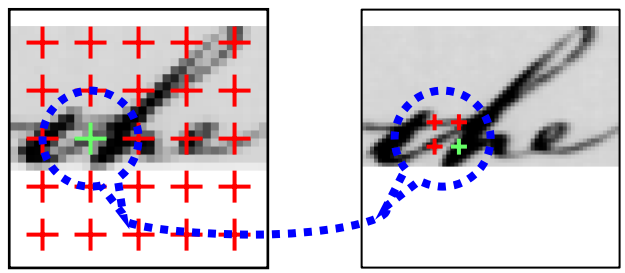
- How to choose good tests?
  - ☞ Exhaustive Search ☜



(80K candidate features)
x(5K images)
x(2K nodes per tree)
x(255 thresholds)
x(200 trees)

} = *too much searching!*

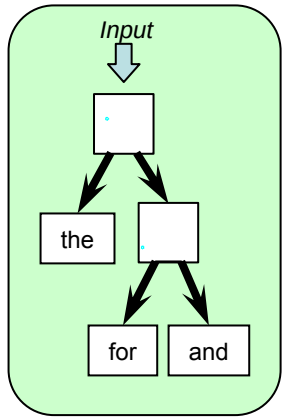
Solution: "Pyramid" search



Coarse Grid *Refined Grid*

Building Decision Trees

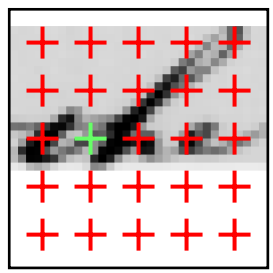
- How to choose good tests?
  - ☞ Exhaustive Search ☜



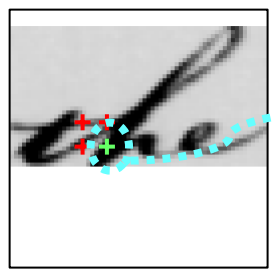
(80K candidate features)
x(5K images)
x(2K nodes per tree)
x(255 thresholds)
x(200 trees)

} = *too much searching!*

Solution: “Pyramid” search



Coarse Grid



Refined Grid

etc.

Problem: Rare Classes

- Zipf's Law: frequency of i^{th} most common word proportional to i^{-1}

$\Rightarrow$ Most words appear only rarely

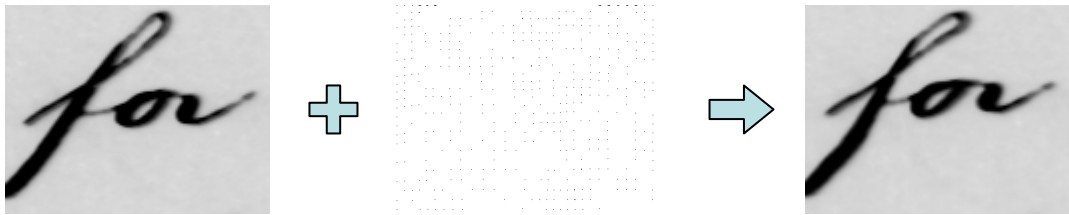
57% of vocabulary: <u>single</u> example

George K. Zipf

- Very hard to learn a class properly from one example!

Augmented Training Data

- Solution? Simulate new training examples.

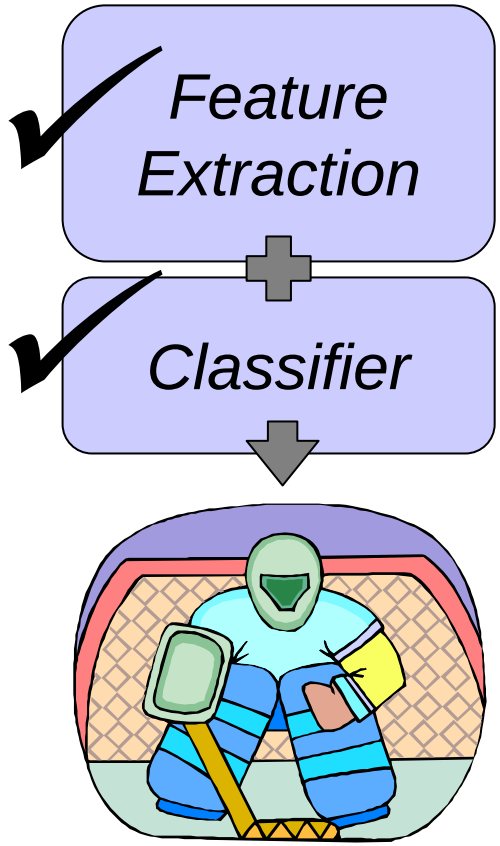


Synthetic Image Warping

- Unusual tactic. Why might it work here?
  - Not simply adding variance to features
  - Result reflects spatial neighborhood
- Only rare classes need augmentation.

Game Plan Revisited

- Features: aligned samples
- Classifier: boosted decision trees
- Bonus: augmented training data



- Testing program:
  1. Word classification error rate
  2. Retrieval using classifier labels

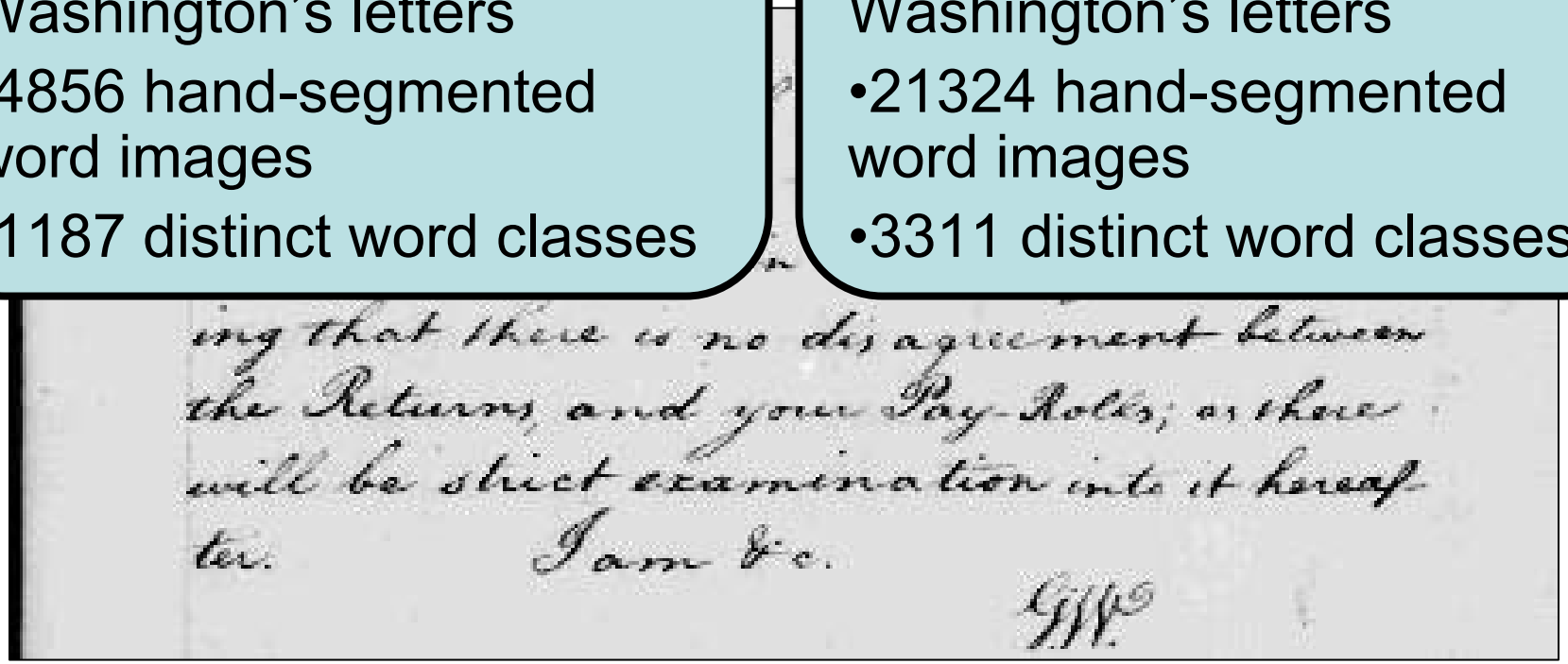
Data Sets

GW20

- 20 pages of George Washington's letters
- 4856 hand-segmented word images
- 1187 distinct word classes

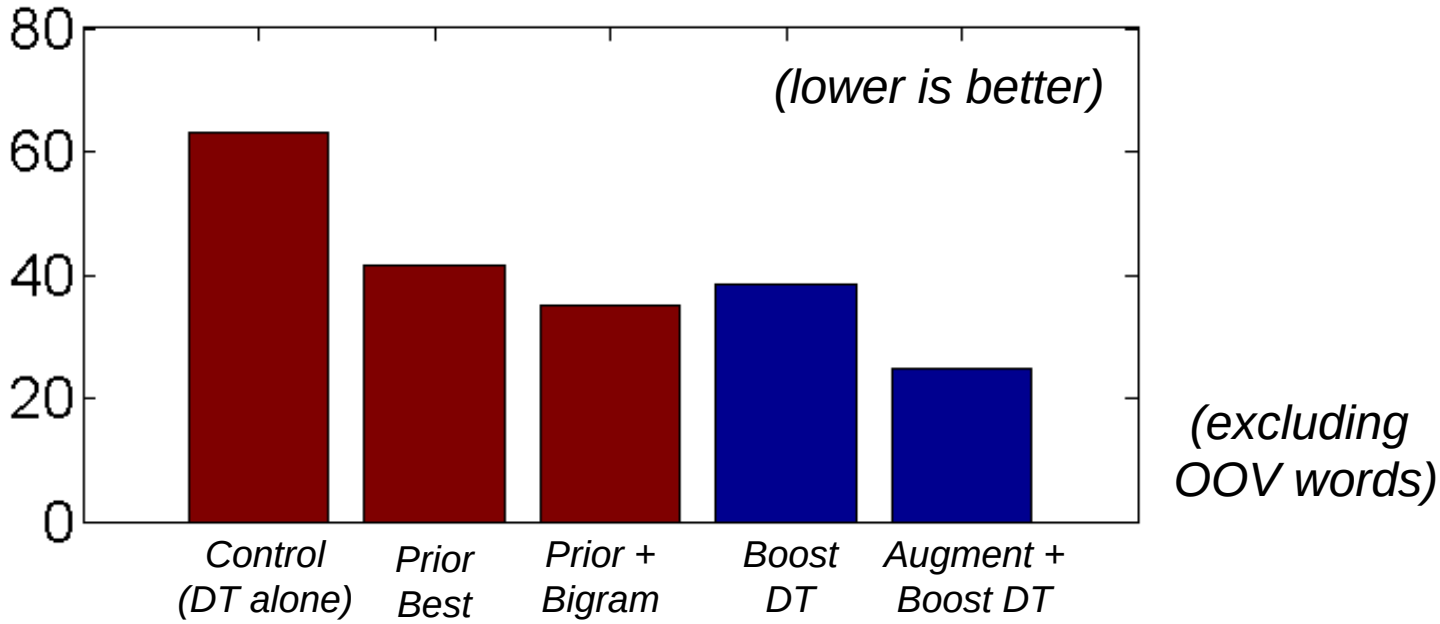
GW100

- 100 pages of George Washington's letters
- 21324 hand-segmented word images
- 3311 distinct word classes



★ *GW100 is harder than GW20.* ★

GW20 Classification Error Rate



- Boosting with augmented training improves error rate, 35% $\rightarrow$ 25% over previous best

Retrieval Experiments

- Language Modeling approach to retrieval:
  - Estimate unigram language model $P(\cdot \mid M_D)$ for each document D
  - Use query-likelihood ranking: score of document D using query $Q=w_1, ..., w_k$ is

$$P(Q \mid M_D) = \prod_{i=1}^{k} P(w_i \mid M_D)$$

Two Ways to Estimate P(· | M_D)

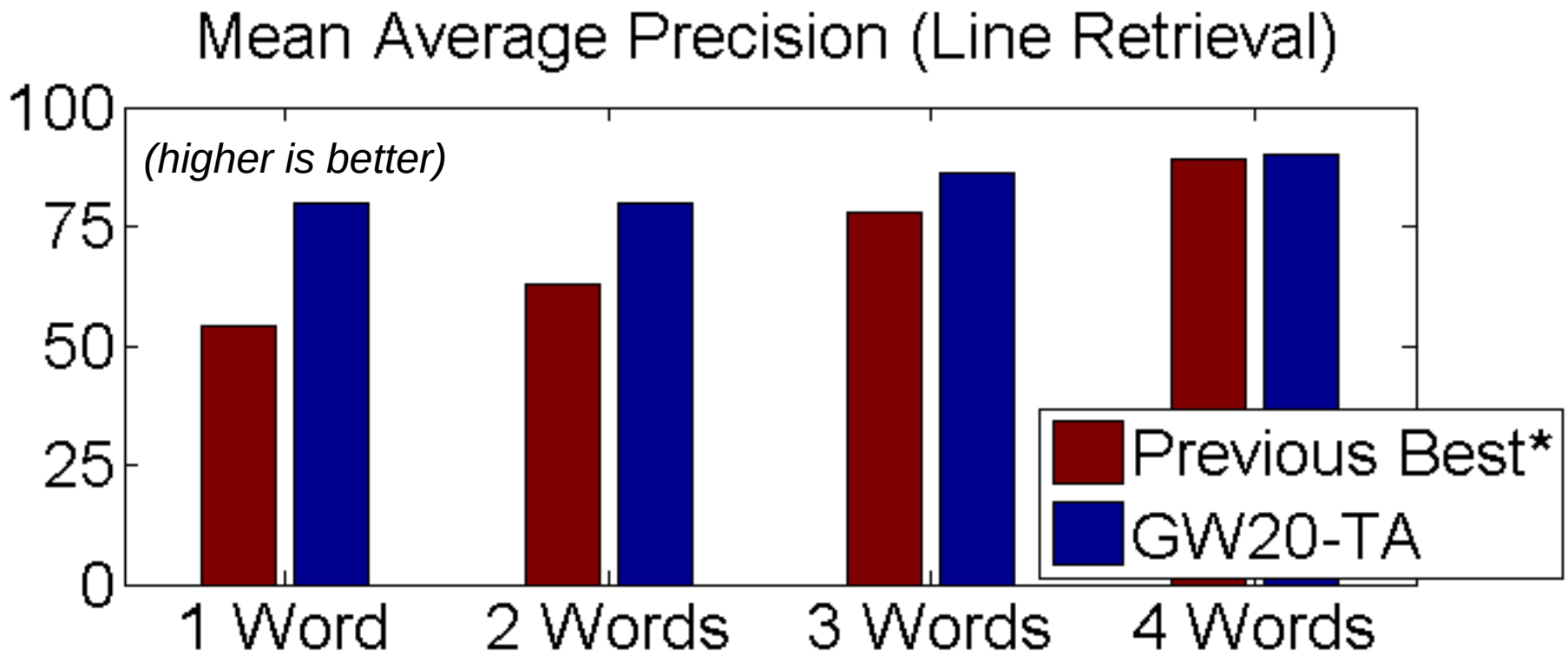
TA: Ignore word classification errors

- Assume word label output = actual text
- Maximum likelihood → P(· | M_D)

PA: Estimate misclassifications

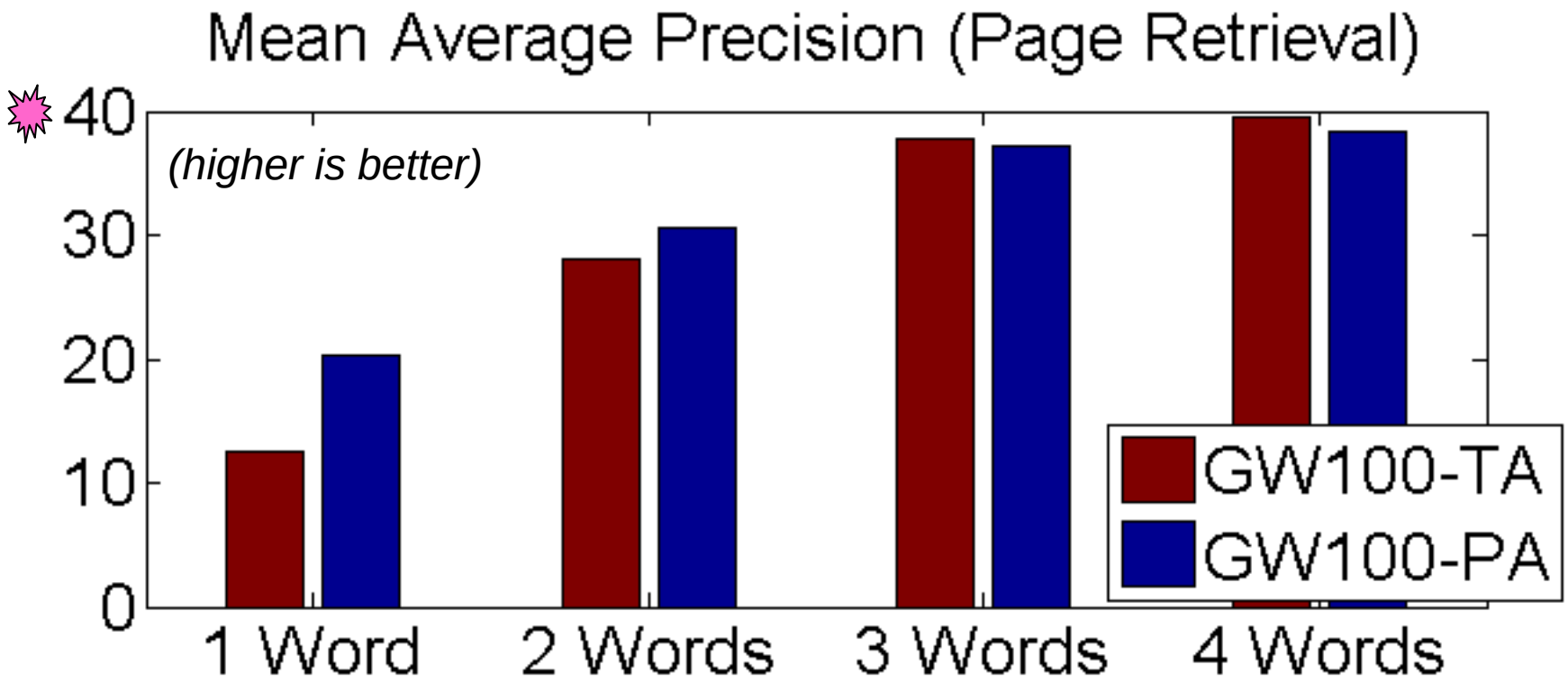
- AdaBoost scores ≠ probabilities (in general)
- Assign probability P(*w* | *Img*) to top *n* labels *w* for each word image *Img*:
  - Fit Zipfian distribution.
  - Label at rank r is assigned *P(r)* = *Z/r* (*Z* gives normalization)
- Estimate $P(w \mid M_D) = \frac{1}{|D|} \sum_{img \in D} P(w \mid Img)$

GW20 Experiments



- Used top annotation (TA model)
- Line retrieval, 10-fold cross-validation design
- Ran all 1- to 4-word queries (no stop words)

GW100 Experiments



- GW20 = training set, GW100 = test set
- Ran most common 1- to 4-word queries
- Compared TA & PA models

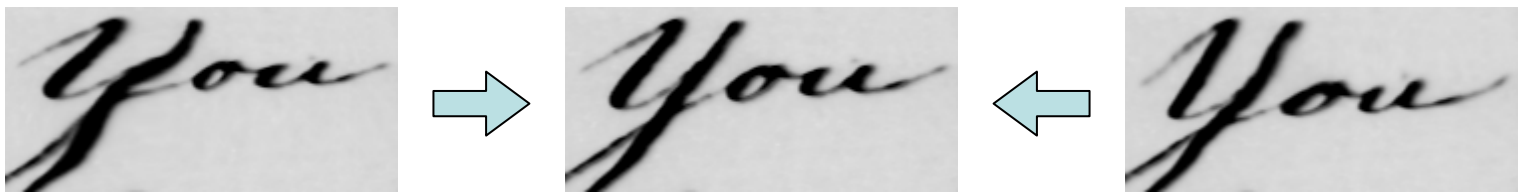
Conclusions

- Boosted trees $\rightarrow$ accurate classification
  - Best reported for GW20
  - Key step: Augmented training data
- Boosting drawback: no probability estimates
  - Can't combine with bigram model
  - More difficult to estimate $P(\cdot \mid M_D)$
    - Choice between TA & PA approaches.
    - PA helps mitigate classification errors
- <u>Best word recognition results so far.</u>

Future Work

- Many test words never seen during training
- Can we create training data out of thin air?


- Global alignment not precise
  - Local alignment possible?



The End

Comparison: GW100 vs. GW20

- 25.7% of GW100 words do not appear in GW20
- More style variation (additional authors, less temporal coherence)
- More ink fading/variance.
- 1/5 train/test split vs. 19/1
- More retrieval units (100 pages in GW100 vs. 66 lines in GW20)

provide all other necessaries for the Expedition, which you know will be wanted

As there are several Contracts made by me to have Cattle delivered here &c. by the 1st. of next month, — I desire that for such as you receive upon that account, if you have money in your hands, you make immediate payment.

Given &c.

Winchester October 29th. 1755. G.W.

29th. Winchester October 29th. 1755.

Parole Williamsburgh.

One Subaltern, one Sergeant, one Corporal, one Drummer and twenty five private men, the Guard to-day — Captain Bacby is ordered to take upon him the command of the Recruits which arrived here under Lieutenant Hall and Ensign Price; who are also ordered to act under him, until further orders — Ensign Hedgeman, and the Recruits which arrived with him, are ordered to join Lieutenant King, and be under his command until further orders — Lieutenant Eustace, and the eight men with him are to join (as soon as they arrive at Fort Cumberland) the Company which Captain Waggoner commands at present, — and the Party left with Sergeant Shaw, is to return to their respective Companies, so soon as they reach the Fort — The Commissary is to see that the Magazine is secured, by fastening up the windows &c. better than they now are; The Officers are to see that the men are cleanly

compt. If that quantity can not be procured, send any lesser quantity that can be got.

I beg you will lose no time herein; by which you will oblige

Yours

G.W.

April 21st. 1756.

To Ensign Hubbard.

Commanding at Enocks Fort.

Sir,

You are hereby desired if possible, to retreat with what men and provision you have to Edwards's; and to Escort what families have put themselves under your protection — But if you find this impracticable without a reinforcement, on your applying to Captain Harrison at Edwards's, a Detachment will be sent to assist you. You are not to fail in bringing off all the Stores you can.

I am &c.

G.W.

April 21st. 1756.

To Captain Harrison,

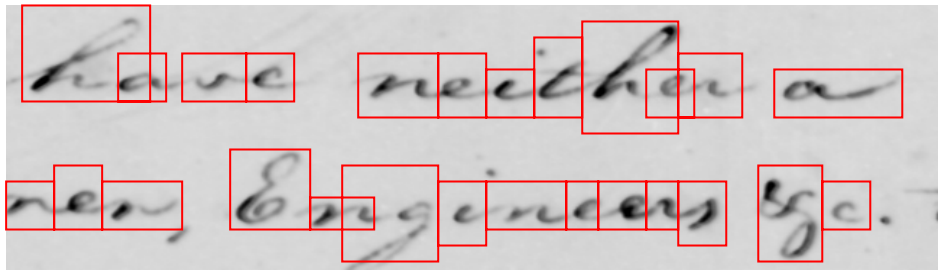
Commanding at Edwards's.

Sir,

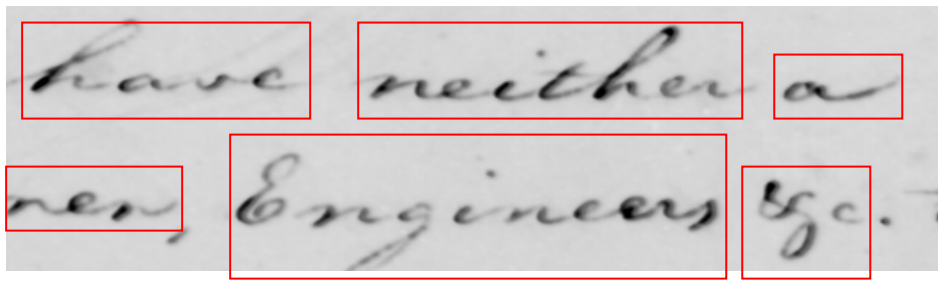
It is out of my power at this juncture to supply you with any Provision. Therefore I would have you apply to Edwards, to whom I write. Acquaint him, that whatever he expends, he shall receive a reasonable satisfaction for: and hint to him, that without his compliance

A Note On Segmentation

- Historically, text recognition has segmented & recognized individual letters

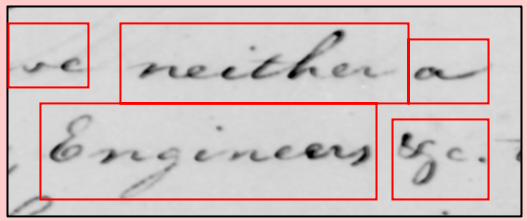


- New work focuses on entire words (easier)



Analysis Sequence

Segmentation



Recognition

Orders

Transcription & Retrieval

Focus of this talk