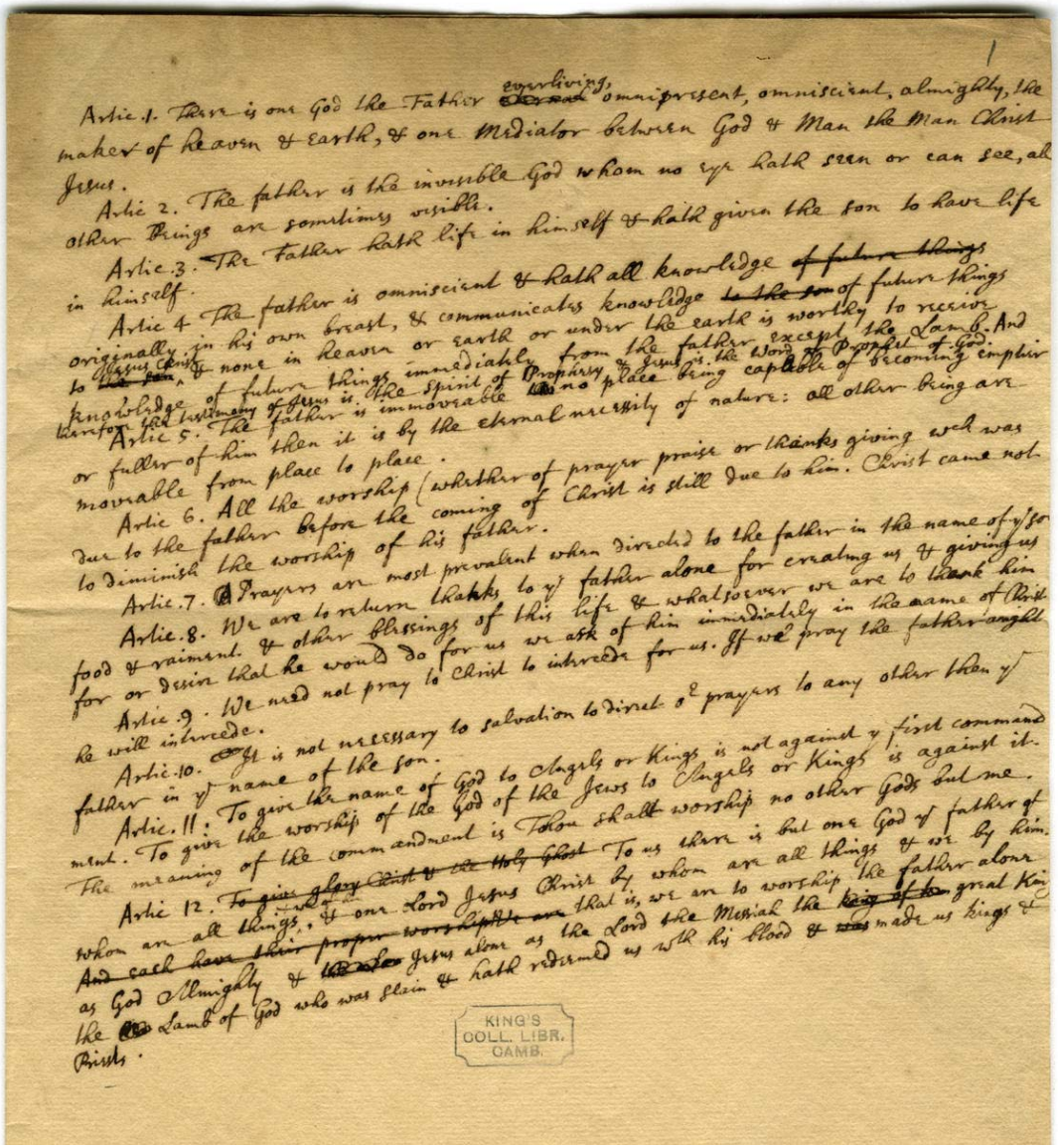
Word Spotting: Indexing Handwritten Manuscripts

Indexing George Washington's manuscripts

R. Manmatha

Multimedia Indexing and Retrieval Group
Center for Intelligent Information Retrieval
University of Massachusetts, Amherst
http://ciir.cs.umass.edu/research/wordspotting/

Isaac Newton Manuscript

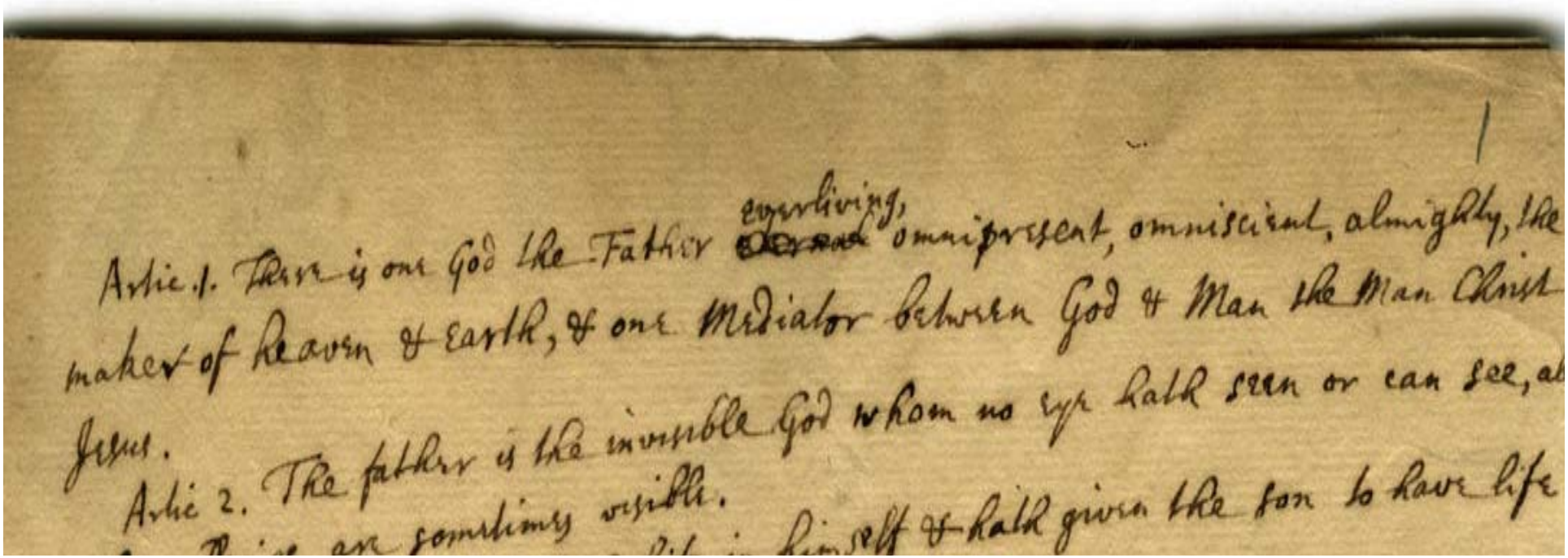


Newton's manuscripts are being scanned and will be annotated at a cost-40pounds/page

Note the varying orientation of the lines – causes complications

Courtesy: Stefan Rueger, Imperial College

Newton - Closeup



Artic 1. There is one God the Father everliving, omnipresent, omniscient, almighty, the maker of Heaven & Earth, & one Mediator between God & Man the Man Christ Jesus.

Artic 2. The father is the invisible God whom no eye hath seen or can see, al

are sometimes visible.

himself & hath given the son to have life

Newton was a religious man. This page lists his religious beliefs as a set of articles (axioms).

(Supposedly) a lot of his writing is on religion and alchemy.

Word Spotting: Indexing Handwritten Manuscripts.

- Index historical documents written by a single author
  - Make an index like one at the back of a printed book.
- Examples
  - Presidential papers at the Library of Congress.
  - Isaac Newton's manuscripts.
  - Margaret Sanger's correspondence at Smith and NYU.
- Variation in single author's writing is small.

- R. Manmatha and W. B. Croft, Wordspotting: Indexing Handwritten Manuscripts, in Intelligent Multimedia Information Retrieval, ed. Mark Maybury, AAAI/MIT Press, 1997.

An Example Manuscript (George Washington)

th
8. To Captain Robert Stewart,
at Winchester.

You are hereby required to take charge of the Recruits sent to Winchester by Captain Gist; whose Son you must order to proceed immediately and join his Father.

Captain Gist this day received one hundred pounds to recruit with, and the same Orders that were given to the other Officers on the 3. instant. &c.

G:W:

Alexandria: December th 8. 1755.

One Possible Approach

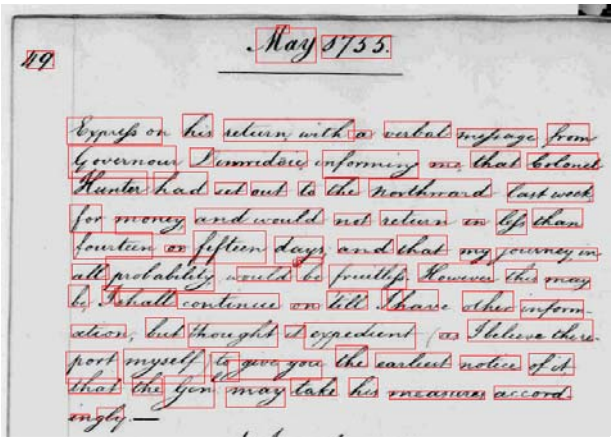
- Recognize words (e.g. optical character recognition – OCR).
- Use text indexing and retrieval.

- Handwriting recognition is a hard unsolved problem.

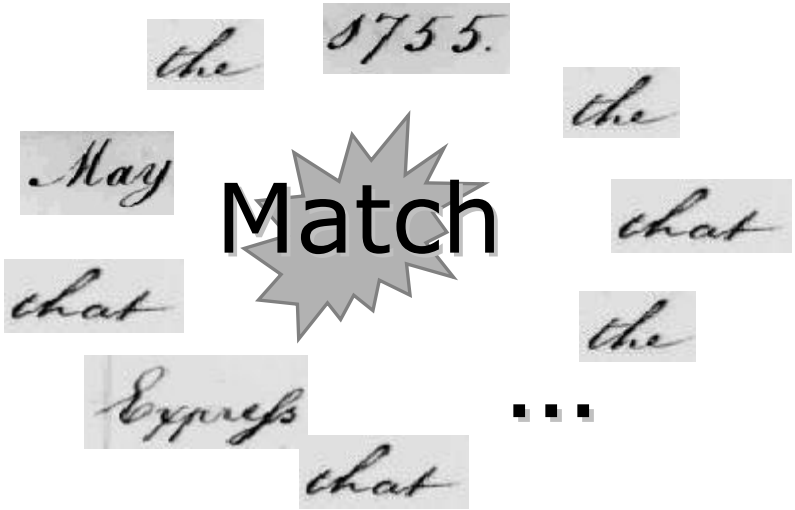
- *Handwritten manuscripts are noisy and the handwriting is variable. This makes it challenging.*

The Idea

Document



Word Image Extraction



Clustering

Tagging

Indexing

Index

...

➢that (p.12, p.45, ...)

...

➢the (p.12, p.34, ...)

...

People Working

- Current
  - Toni Rath, J. Jeon – Grad Students
  - A. Maguire, J. Rothfeder, K. Srivastav – Undergraduates
- Previously
  - Shaun Kane, Andrew Lehman, Elizabeth Partridge – REU's (site REU award)

  - Nitin Srimal – Grad Student.
  - Joshua Sarro, Eric Mulvihill and Liz Yon – Undergraduates
  - Fangfang Feng – staff programmer

Overview

- Scanning
  - Often have little control over this process.
- Segmentation
  - Developed a new scale-spaced segmentation algorithm – discussed last year.
- Word Similarity (Matching)
  - Preprocessing
  - Pruning
  - Matching/Clustering
- Indexing/Database Management
- User Interface
  - How does one map images to ASCII?
  - Alternatively what does a visual index look like?

Collection

- Have roughly 6,400 scanned pages of George Washington's manuscripts from the Library of Congress.
- Scanned in 8 bit graylevel at 300dpi.
- Scanned from microfilm
  - Quality not as good as scanning from original.
    - For example, boundary artifacts, noise etc.
  - Probably done for reasons of cost, fragility of manuscripts and security.

Compression Artifacts

Edge Image created from Jpeg compressed version

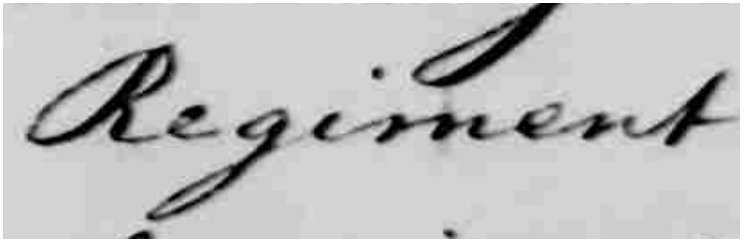
Edge Image created from lossless tiff compressed version

Moral: Scanning is important. Our work so far has been limited to jpeg images. In the process of getting lossless tiffs from the Library of Congress.

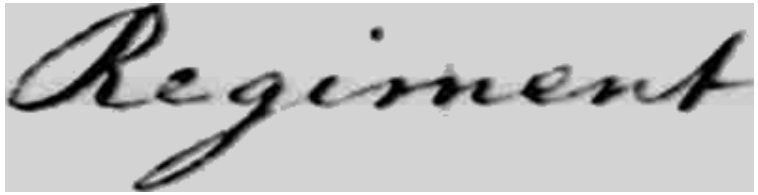
Word Similarity

- Pruning
  - Eliminate most matches using Area, Aspect Ratio and Descenders
  - Results
    - 87% possible matches pruned
    - 94% relevant matches retained

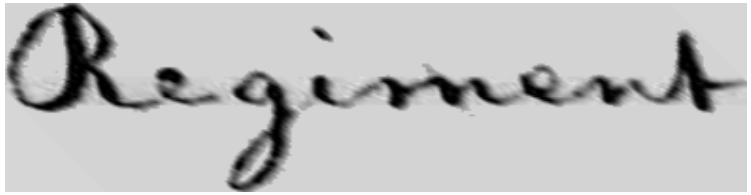
Preprocessing Example



Original



Cleaned and Cropped



Slant Corrected

Matching

- Template matching
  - SSD – Sum of Squared Differences
  - XOR
  - EDM [Danielsson 1980]
  - **AEDM** - Affine corrected EDM (align images, then EDM)
- Transformation recovery
  - SLH (affine transformation) [Scott & Longuet-Higgins 1991]
  - Shape Context (thin-plate spline transformation) [Belongie et al. 2000]
- Feature matching/alignment
  - **Dynamic Time Warping** for feature alignment and comparison (e.g. projection profiles)

Matching Considerations.

- George Washington manuscripts
  - Some variation in the way words are written.
  - *Preprocessing* is important.
  - *Representation* for matching is important.
  - *Alignment* is important.

  - Some standard techniques don't work so well on this problem.
    - Examples: SSD, Shape Context
  - Things which work reasonably well.
    - AEDM - EDM after an approximate affine alignment.
    - Dynamic Time Warping of projection profiles of words.
      - We think this can be improved even further.

AEDM

Align Images up to an affine transform. Then

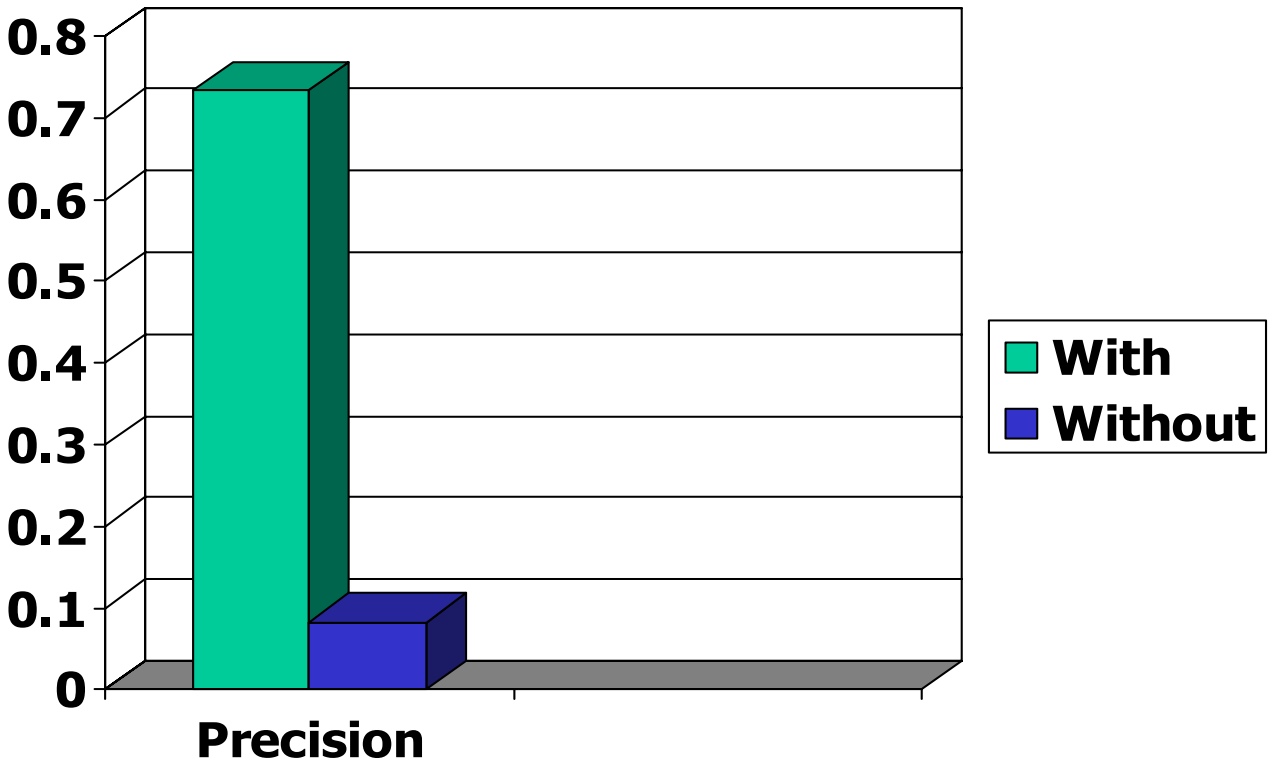
Image 1

Image 2

XOR

AEDM Error measure: sum over difference pixels in XOR image; pixels in “blobs” are weighted more heavily

Effects of Preprocessing

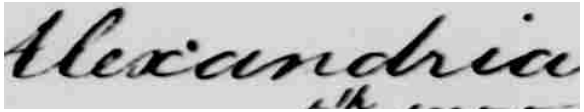


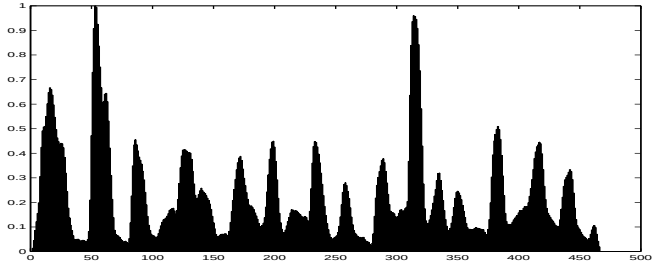
AEDM with and without preprocessing.

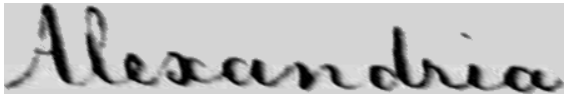
Dynamic Time Warping of Projection Profiles.

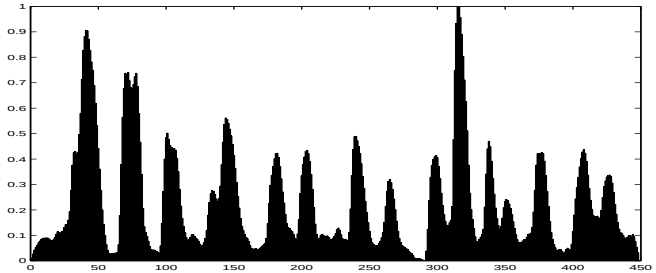
Projection Profiles

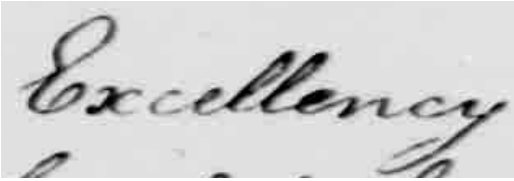
Original Images → Cleaned, deslanted, cropped → Projection profiles

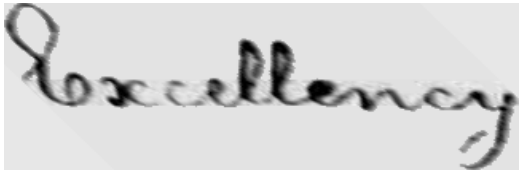


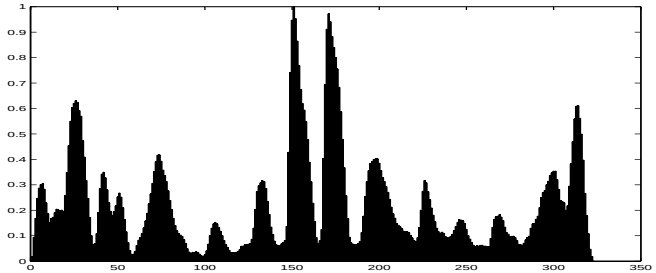








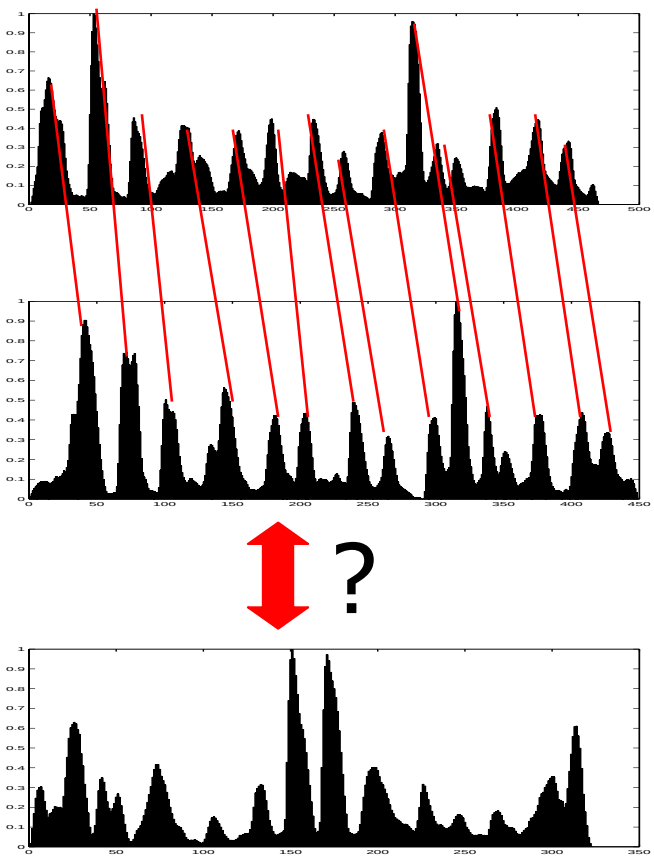




Similarity of Projection Profiles

What makes projection profiles similar?

Find correspondences in projection profiles



obvious correspon-dences

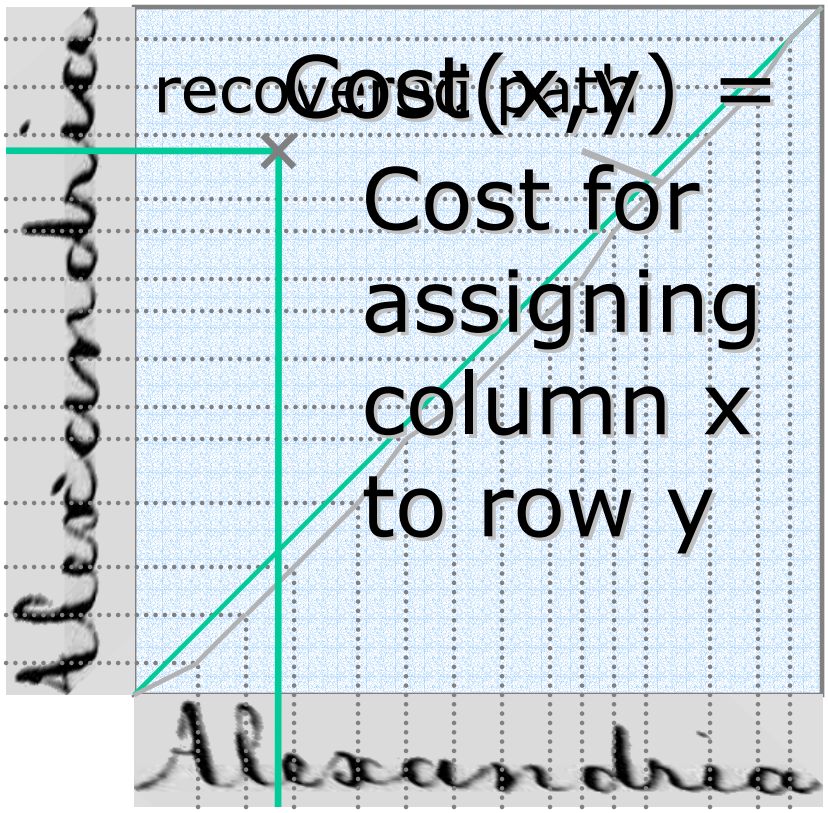
no obvious correspon-dences

Dynamic Time Warping can recover non-linear alignment between two time series

Projection Profiles are more stable.

Dynamic Time Warping

Find minimum-cost assignment path



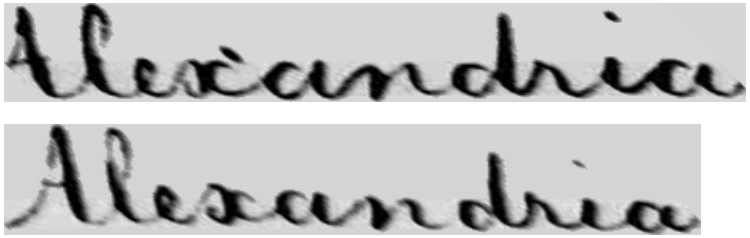
Generate match error (or score)

Use sum of assign-ment cost along recovered path

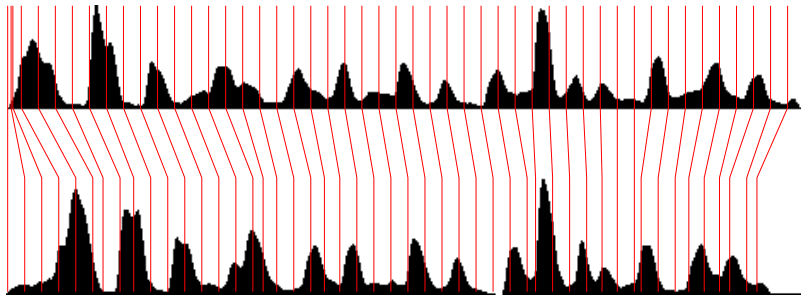
Use recovered path to align images for a subsequent matching algorithm

Using Alignment for Matching

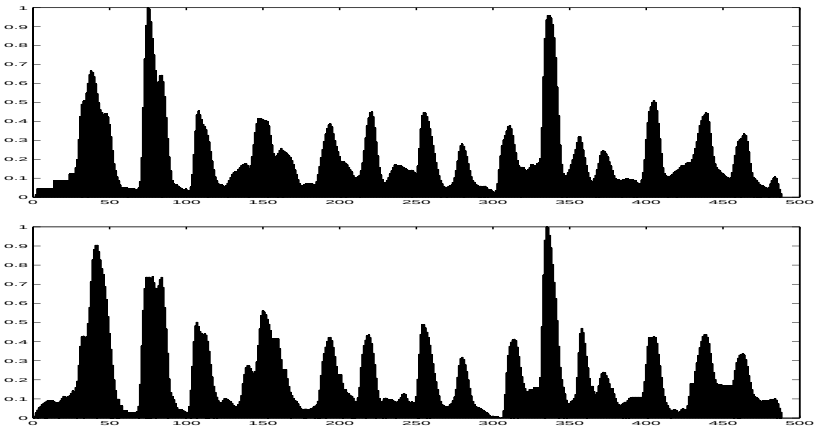
Original Images



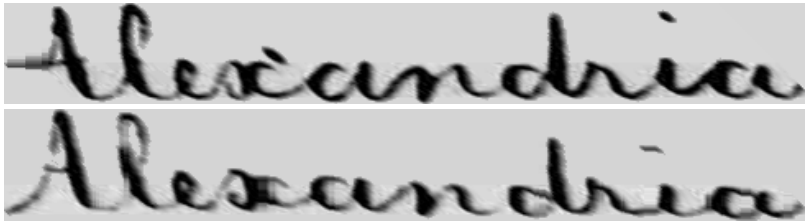
Recovered Alignment



Warped projection profiles



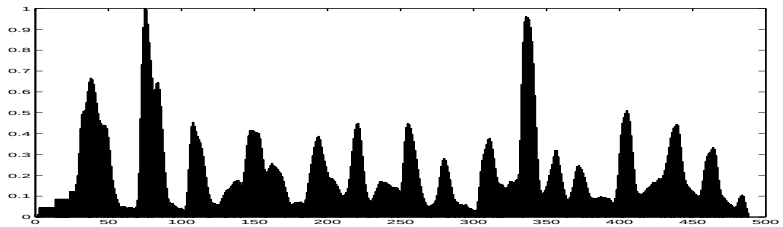
Warped images



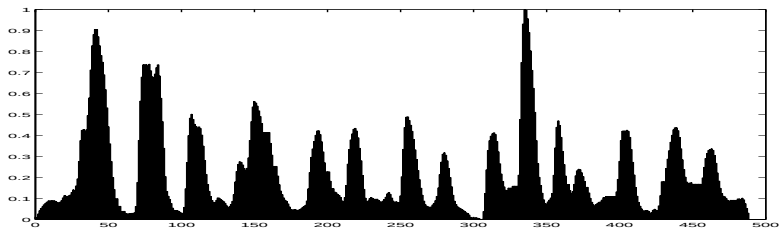
Computing Error Measures

Warped projection profiles

p(i)=

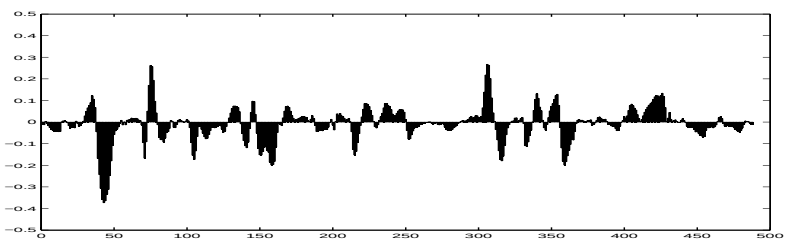


q(i)=



i=1,2,..,n

Difference d(i)=p(i)-q(i):



Error measures

1. Sum of abs. differences:

$$e = \frac{1}{n} \sum_{i \in \{1..n\}} | d(i) |$$

2. Euclidean distance

$$e = \frac{1}{n} \sum_{i \in \{1..n\}} d(i)^2$$

3. Kullback-Leibler div.:

$$e = \sum_{i \in \{1..n\}} p(i) * \log \frac{p(i)}{q(i)}$$

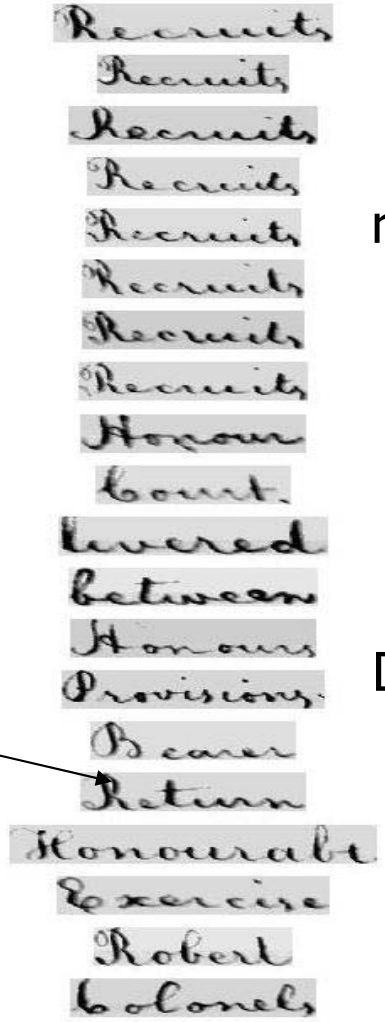
Retrieval Results

Alexandria

10/11 retrieved in top 10

8/8 retrieved in top 8

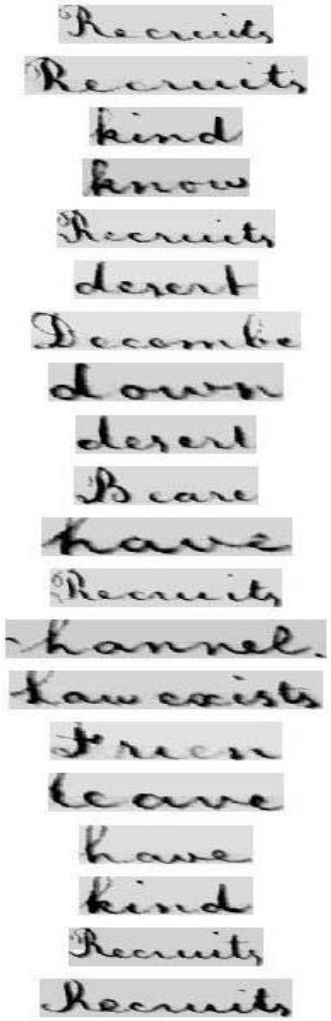
Recruits



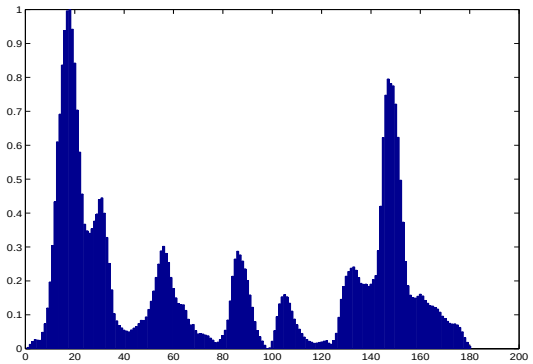
6/8 retrieved in top 20

Template used matters. Don't know why yet.

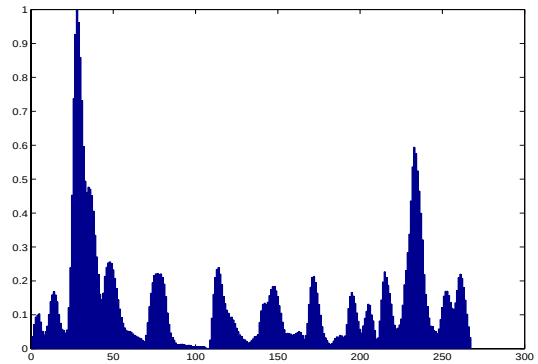
Template Dependence



Problem



Same word?

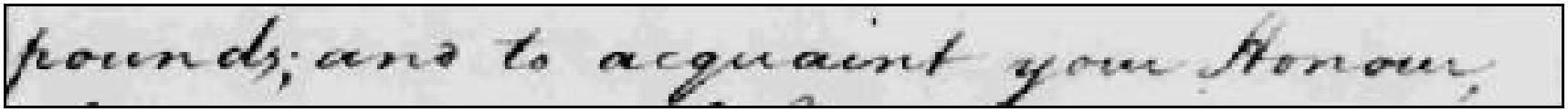


no!

Probably need additional features to distinguish words with similar projection profiles.

Note, we only use 1D projection profiles.

Variation between Pages

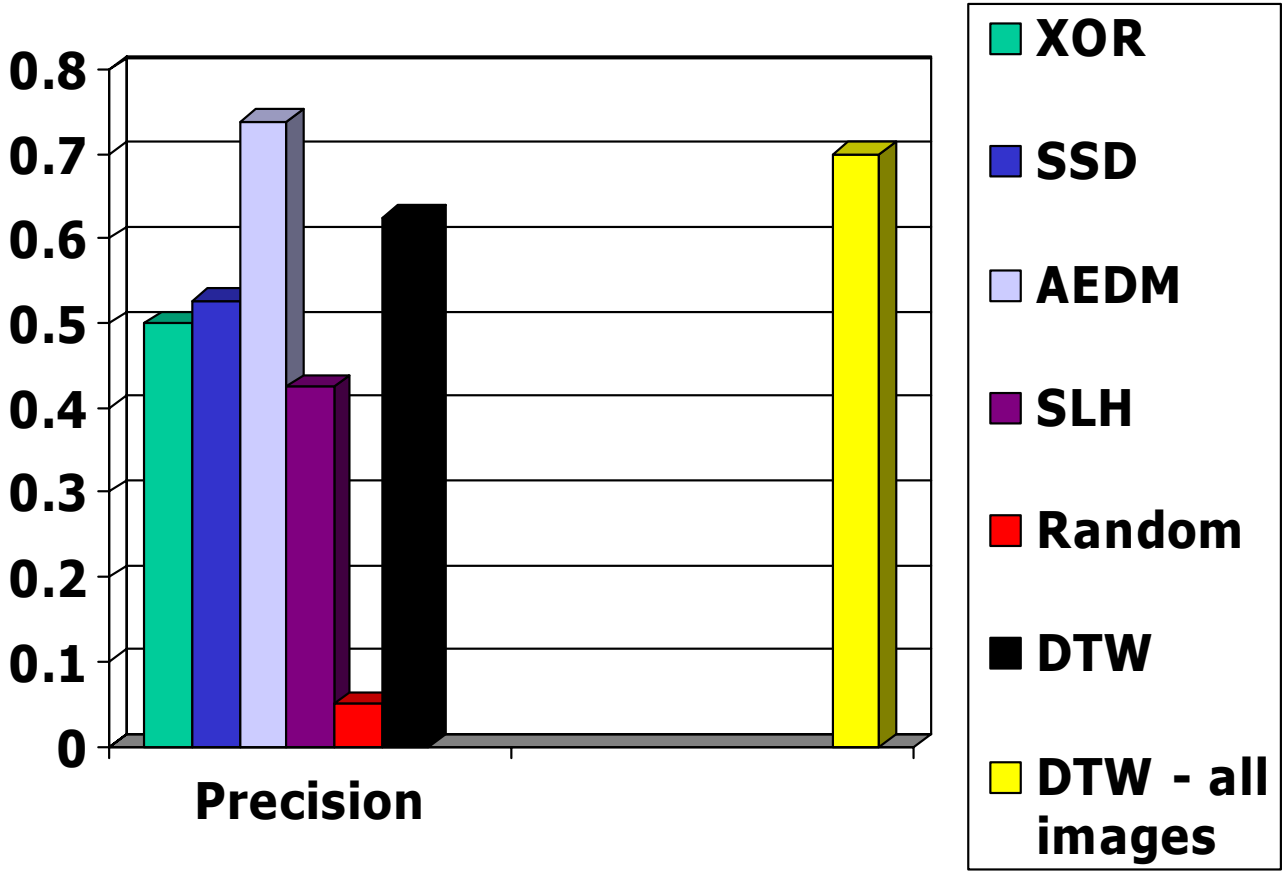


Example of cleaner page with less noise

Example of more difficult page. Faded Ink, Noise

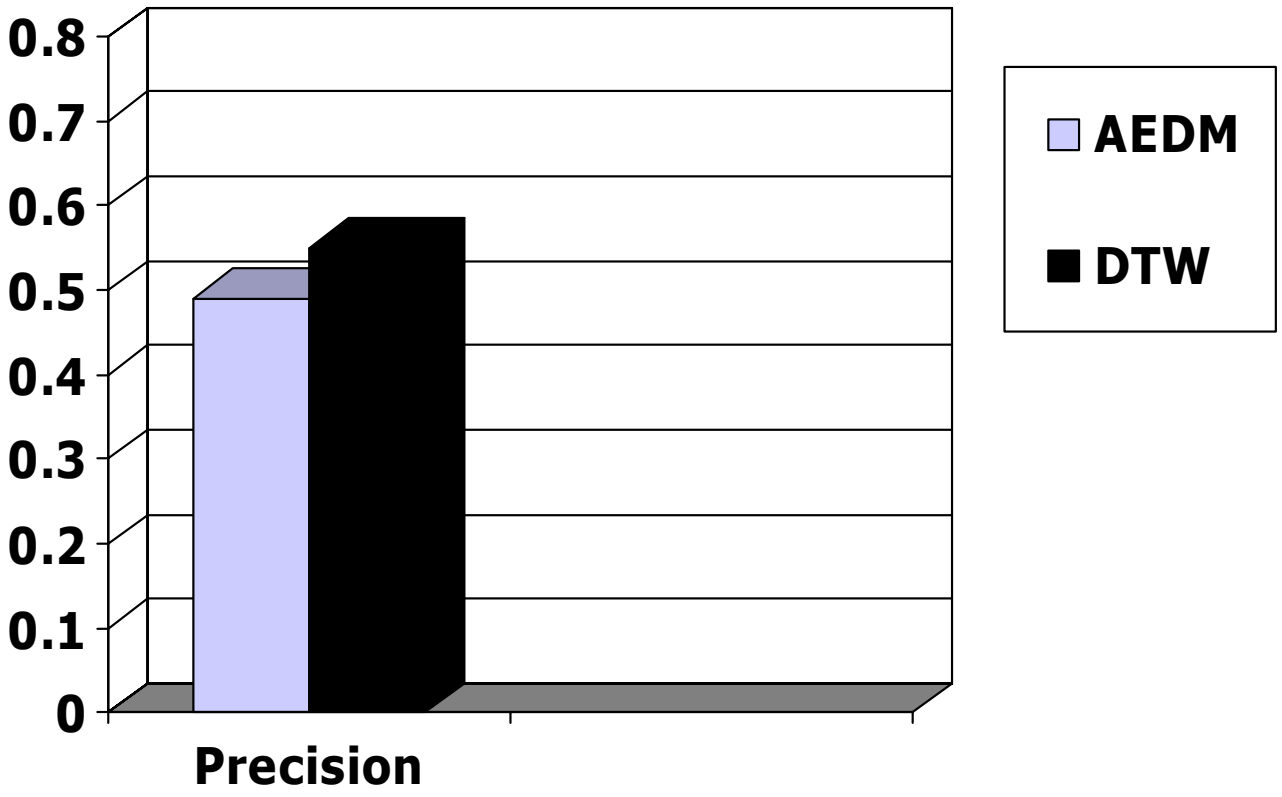
Comparison of Matching Algorithms

Test Set – 10 pages, 15 queries



Comparison of Matching Algorithms

Difficult Test Set – 10 pages



DTW - Works better on difficult test set

Computational

- Matching expensive operation – $O(n^2)$.
- In 6,400 images, roughly 1.8 million words, no. of possible matches ~ $2.5 * 10^{12}$.

- Our current pruning reduces this by a factor of 10.
- Clustering matches could also reduce this substantially.
- Highly parallelizable.

User Interface

- How does one map index images to ASCII?
  - Have people do this manually.
  - Read a few pages, recognize them using ASR.
    - Align these pages to the handwritten ones.
    - Currently working on this.
- Use a visual index.
  - People aren't used to a visual index.
  - How does one create such a visual index?

Demonstration

- See http://ciir.cs.umass.edu/research/wordspotting/index.html

Conclusions and Future Work

- Have a reasonable segmentation algorithm.
- Currently working on effective matching techniques.
  - Hard Problem.
  - Dynamic Time Warping of the right representation shows great promising.
    - DTW on additional features will probably improve performance
    - Computational Processing Issues.
- Database Issues.
- User Interface Issues.
- In principle, Word Spotting should work in other languages.