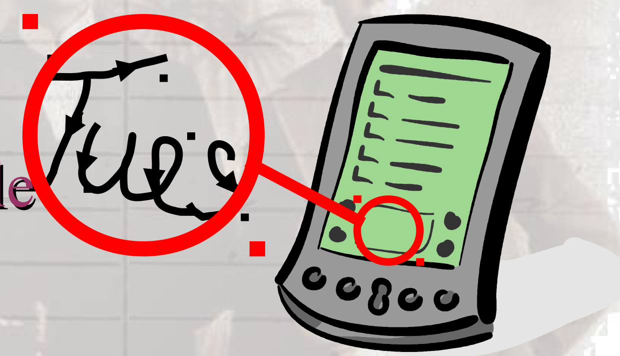
Handwriting Recognition for Genealogical Records

Luke Hutchison (lukeh@email.byu.edu)
[Advisor: Dr. Tom Sederberg]

Handwriting Recognition

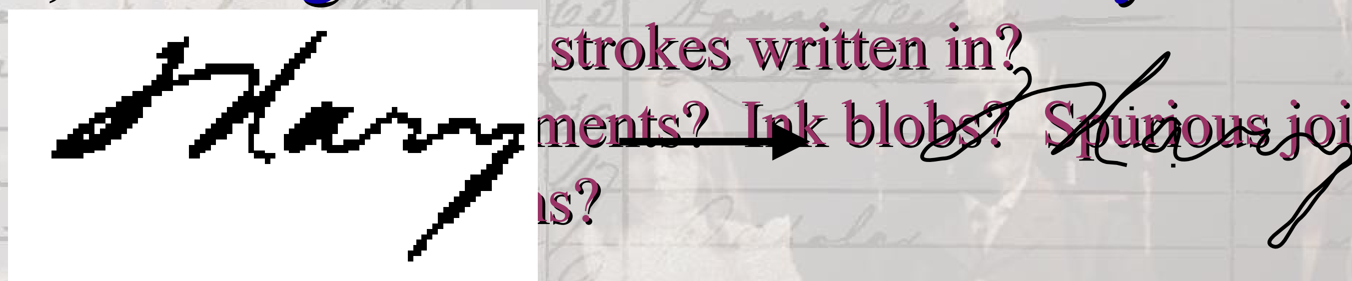
- Two different fields:
  - Online Handwriting Recognition
    - The writer's pen movements are captured
    - Velocity, acceleration, stroke order available
  - Offline Handwriting Recognition
    - Page was previously-written and scanned
    - Only pixel color information available
- Genealogical records are all offline
- Offline is harder (less information is available)



Handwriting Recognition

- Can we just convert offline data into (simulated) online data?

- Yes, although difficult to do reliably:
  - strokes written in?
  - ments? Ink blobs? Spurious joins between
    ns?



- Especially difficult with genealogical records

Handwriting Recognition

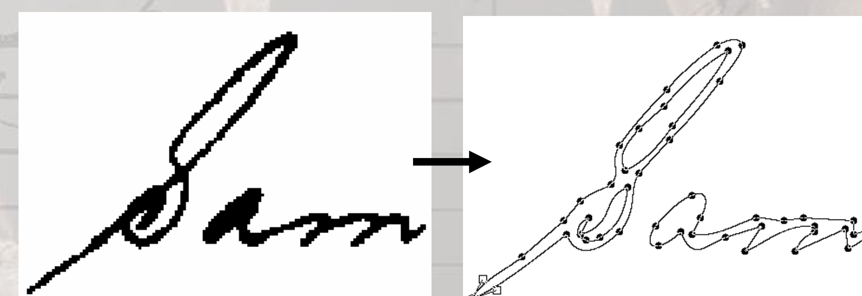
- A successful approach must combine results from analysis of different domains, and at different levels of abstraction, e.g.

- Discrete:
  - Stroke segmentation and ordering
  - Digraph frequency tables, lexicons

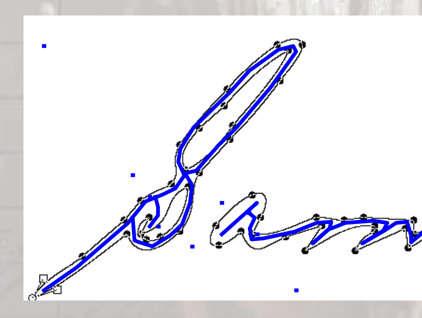
- Continuous:
  - Letter shape analysis and matching

Handwriting Recognition

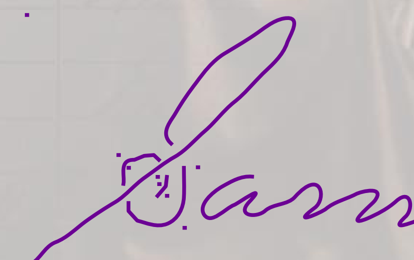
- An example of some common steps in the analysis process:
  - Contour extraction



  - Midline determination



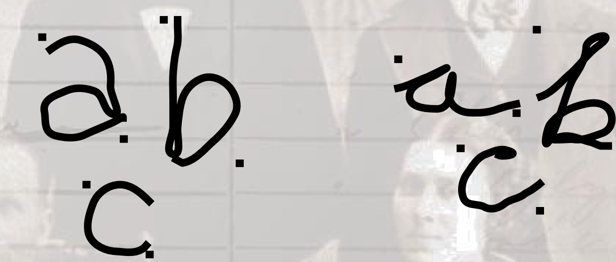
  - Stroke ordering



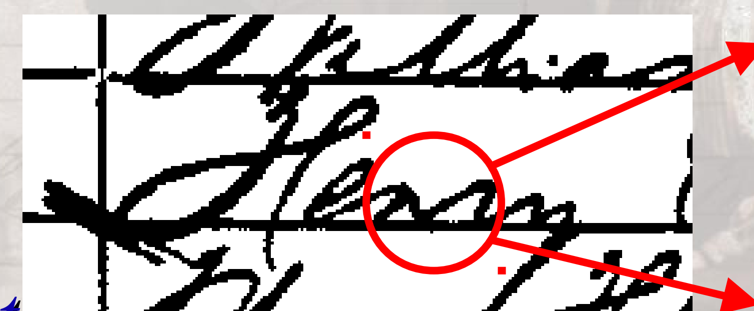
Handwriting Recognition

- An example of some steps in the recognition process:

  - ***Handwriting style clustering***



  - ***Letter recognition***



nr?

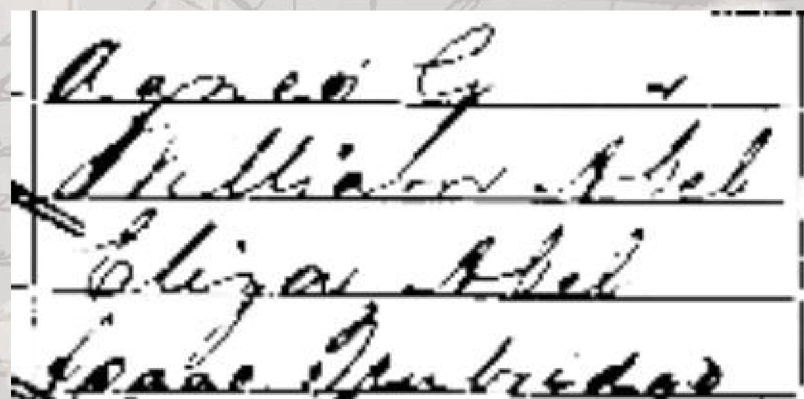
m?

  - ***Approximate string matching***

Smith
Smythe

HR for Genealogical Records

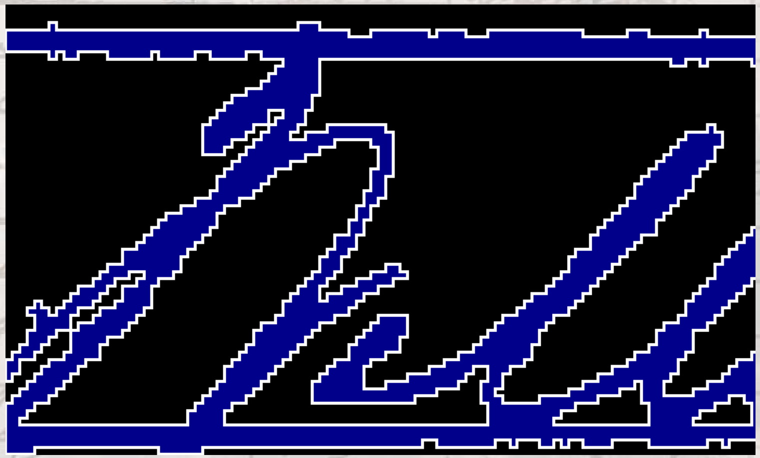
- Image quality is not always good with microfilms
  - Fading of documents / microfilm
  - Ink-well pens

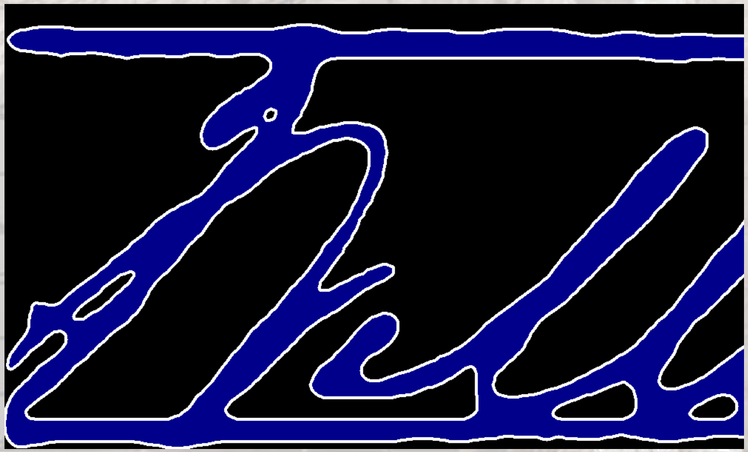


- But documents were usually written meticulously
  - Older handwriting more regular; simpler to match
  - Different approach required

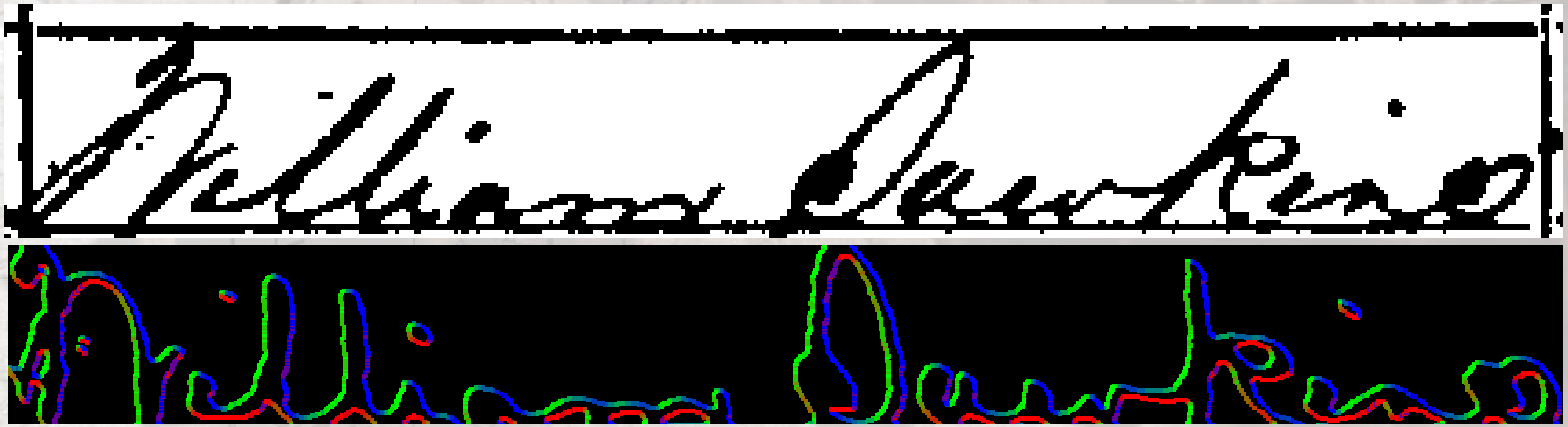
The Approach

- Outlines of word are traced and smoothed



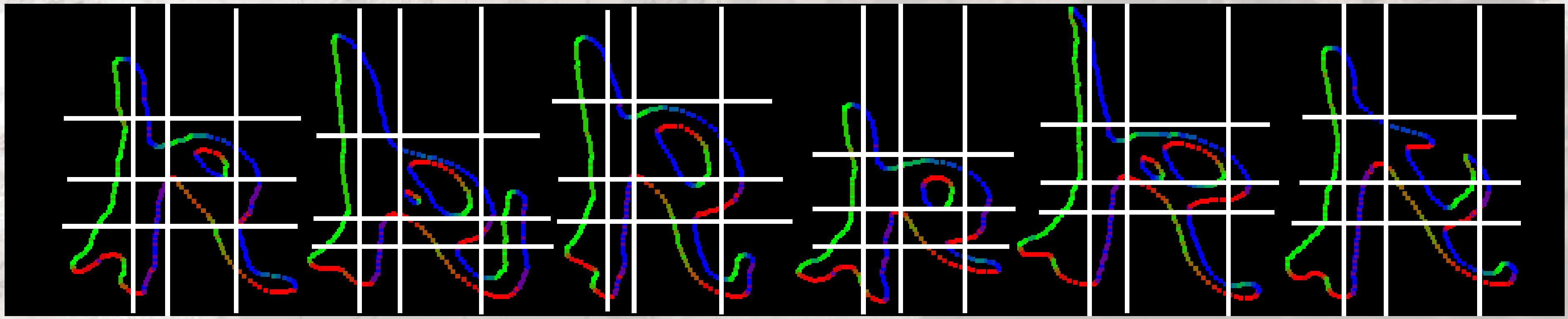


- Some common sources of variation (e.g. differences in slope) are automatically corrected for.



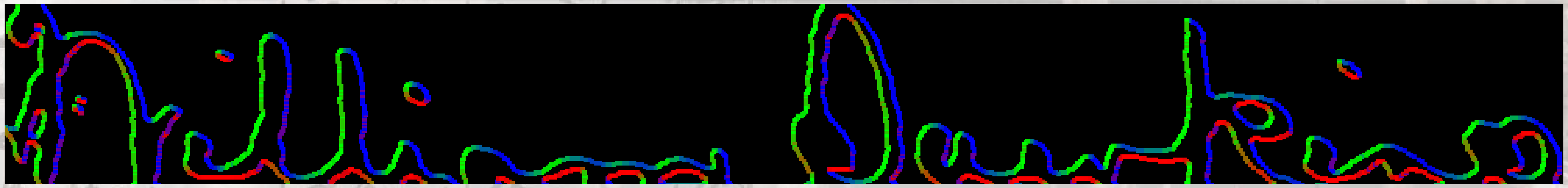
The Approach

- Robustly produce a characteristic “signature” for each letter



The Approach

- Find possible letter matches and determine possible readings (with accuracy of fit)



W i l l i a r w S u w k i n o

M n w r u s

i m a

J U o O t m

=> Williarw Suwkino (65%), ... , JiiUiom Oartums (1%)

The Approach

- Error Correction: Letter digraph frequencies
  - E _ 2.617%
  - E R 1.438%
  - N _ 1.280%
  - A N 1.276%
  - _ S 1.212%
  - O N 1.207%
  - I N 1.187%
  - E N 1.174%
  - [...]
  - A W 0.075%
  - N K 0.074%
  - T L 0.071%
  - [...]
  - U W 0.000%

Suwkino --> Sawkino

The Approach

- Error Correction: Name Lexicon

- Last names:
  - Smith 1.105%
  - Jones 0.817%
  - Williams 0.653%
  - Brown 0.371%
  - [...]
  - Sawkins 0.012%

- First Names:
  - James 1.615%
  - John 1.203%
  - Robert 1.022%
  - Michael 0.971%
  - William 0.954%

=> William Sawkins (95%)

Conclusions

- **[Work in progress]**
- **(Semi-) Automated extraction system could dramatically reduce extraction time**
- **[Demo: Concept search engine...]**