Week 4 Review

Zachary Daniels

Mentor: Haroon Idrees

Outline

- Major goals:
  - Get handwritten OCR's accuracy on par with (or near) machine-printed OCR
  - Solve synthetic puzzle
    - Edge matching algorithms
    - Oracle
- Side goal: Build oracle for real life puzzle

Mismatches and misalignments

- Vertical shifts have more impact than horizontal
- Incorrect alignment will almost always have at least a minor impact (but typically on text that is wrong to begin with)
- Overlapping characters will throw off recognition where it matters most

Handwriting Recognition

- How good is OCRopus out-of-the-box for handwriting recognition?
  - Started as a system for recognizing handwriting
    - Ability to recognize handwriting was later disabled and removed from the software
    - Main use now: high-volume document recognition (e.g. books)
- Tests for accuracy
  - Simple handwriting examples
  - On combined chads
  - On the puzzle solution

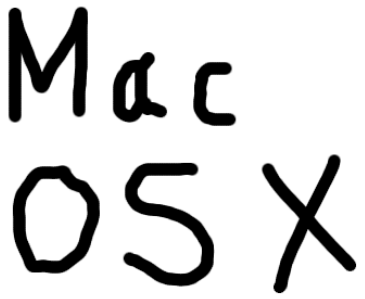
Handwriting Recognition

- ‘kdll’
- ‘tast’
- ‘las’



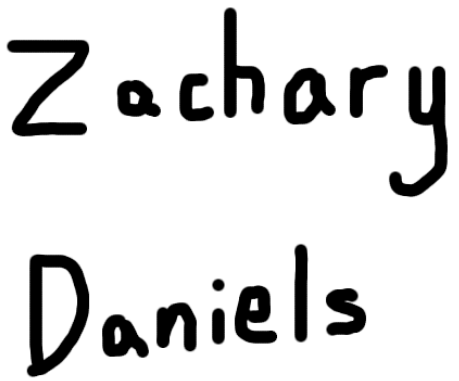
Handwriting Recognition

- ‘Mcc’
- ‘o5X’



Handwriting Recognition

- ‘2cboru’
- ‘Doelv’



Handwriting Recognition

- ‘haroon’
- ‘Tdrees’



Handwriting Recognition

- ‘UCF’
- ‘O)rlando’



Handwriting Recognition

- ‘Tks’
- ‘(5 o’
- ‘ie sh’



Handwriting Recognition

- ‘Lo sk’
- ‘ovef’
- ‘eberdl’



Handwriting Recognition

- ‘sj-/W’
- ‘-- , d’
- ‘icrl’



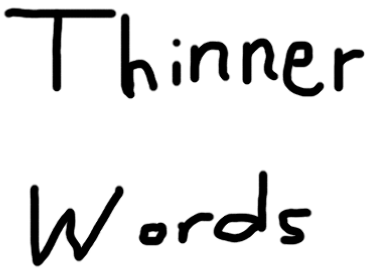
Handwriting Recognition

- ‘PdfLeb’
- ‘Writng’



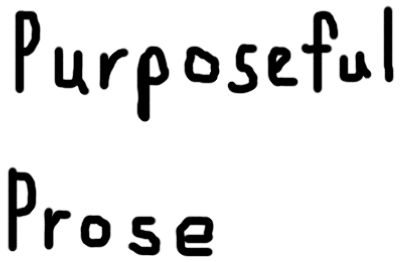
Handwriting Recognition

- ‘TLrner’
- ‘lyres’



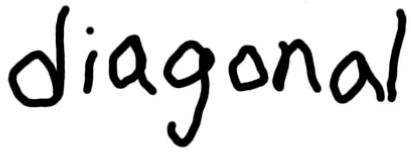
Handwriting Recognition

- ‘furpose`l’
- ‘Prose’



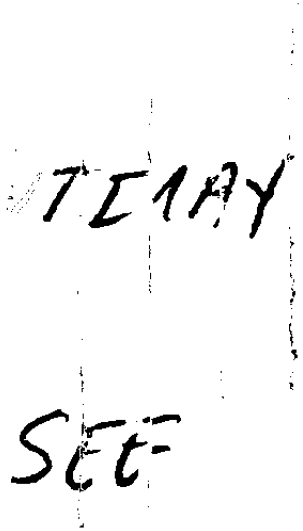
Handwriting Recognition

- ‘diA4onol’



Handwriting Recognition

- ‘7zcAy’
- ‘Sae’



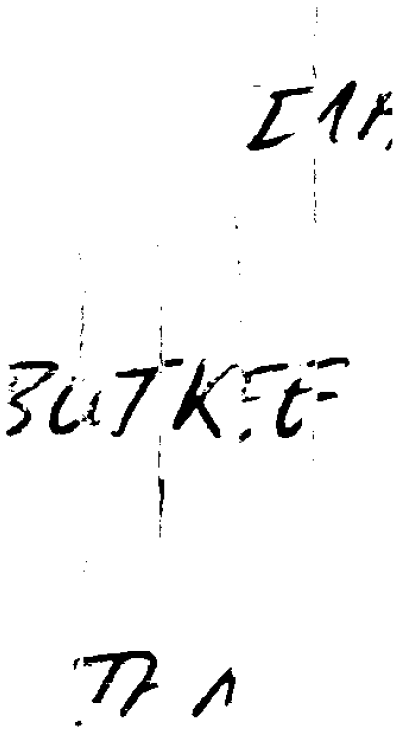
Handwriting Recognition

- ‘7 ` r5 4’
- ‘/ - / l’
- ‘,r`,--.r’
- ‘-, cA’
- ‘N F’
- ‘AA-t’



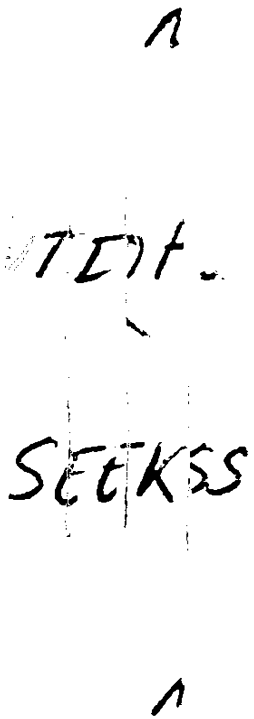
Handwriting Recognition

- ‘zrA’
- ‘s7Ae’
- ‘77’



Handwriting Recognition

- ‘AA’
- ‘7` r 4’
- ‘/ Z- / I’
- ‘A`-`Jre’
- ‘N F P’
- ‘-c ,AA A7’
- ‘I’



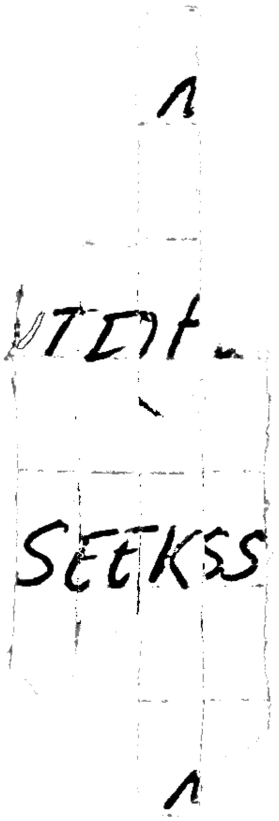
Handwriting Recognition

- ‘7z r’
- ‘SeTxss’



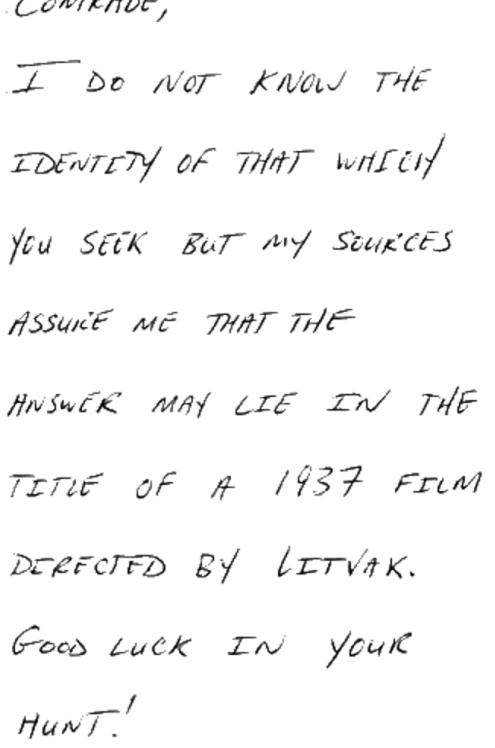
Handwriting Recognition

- ‘AA’
- ‘-m` `PA 4’
- ‘F/ z- t r w-’
- ‘r` -rirAA’
- ‘M A Pi A/ A’
- ‘w t, I ;tV A’



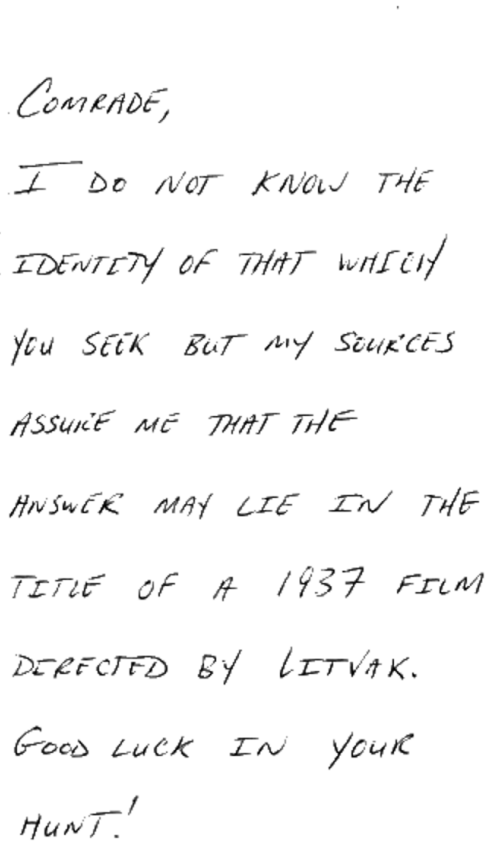
Handwriting Recognition

- ‘Coo-’
- ‘-4- oe Aor Aa./ 77F’
- ‘Tz7rr7y a T7 -oTo’
- ‘)ra SaX T7 Ay szarcrs’
- ‘Assr ar 2e 7)/e’
- ‘AwscX AsY czz` Z7r 7
- ‘T2re oF s /9A% Arw’
- ‘AzecrF2 2y xTaxr’
- ‘a=oo zocA ZA ywA’
- ‘raT``’



Tesseract Recognition

- Also fails miserably:
- . ($44!/70¢-',
- ....———-
- ...1- Do /I/of X'/WLJ 7"/F
- :mr~r;7y of 77/47‘ lv/1IL‘//
- /cu 5&7 5’a-7’ my 5‘oo/KCFJ
- /7550/¢'F Mt‘ 72/07’ 77%‘-
- ,7”-5-.:7<’ my 4:: 17¢ 7‘//E
- /‘Inf of-‘ /91 /K3? FILM
- Dner-crw 8/ lrv-/rm.
- éooo Luck rm‘ yam’
- _/
- /7'mv"/ .



Improving Handwritten OCR

1. Modify writer identification software to find most similar writers from dataset
2. Train OCRopus on data from most similar writers
3. Test models on puzzle text

Similar Writers

COMRADE,
I DO NOT KNOW THE
IDENTETY OF THAT WHICH

blotches. Hers was flushed and lovely. O.K. teacher!
hoping for the mark to rise, as has now happened.
Pugh. How could they be otherwise, when his
His ~~d~~only reference to Trelawny by name in the course of several
WERE NOT YET SUFFICIENTLY AT

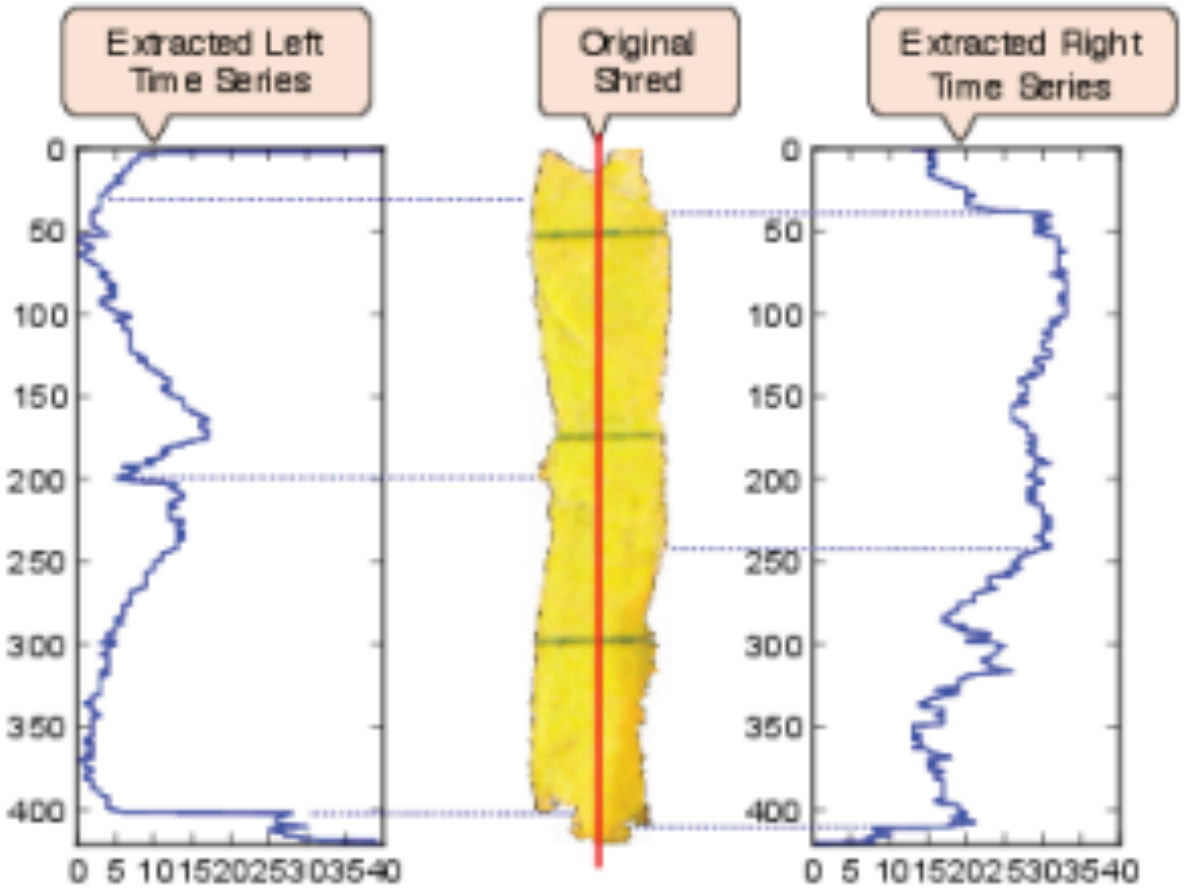
Results

- As bad, if not worse than the default model
- Only trained on most similar writer
  - Far too little data (30 lines)
  - Not enough variation in style
  - Puzzle written in all capital letters
    - Most lines in the dataset are mixed-case
- Possible improvements
  - Multiple writers
  - Mix in computer fonts
  - Use large amounts of data
    - Problem: Every line needs to be annotated
  - Try using puzzle 2 which has mixed-case writing

Edge Matching Algorithms

- Time Series
  - "Deshredder" results: 50% correct matches in top 20% output matches
    - "Deshredder" also used a time series base on color + brightness of edges which isn't applicable to the synthetic dataset
    - With just edge-shape: ~30-40%
- Unsupervised Feature Learning
  - Learns fragments of edges
  - Similar performance to time series
  - Might improve with better tuning

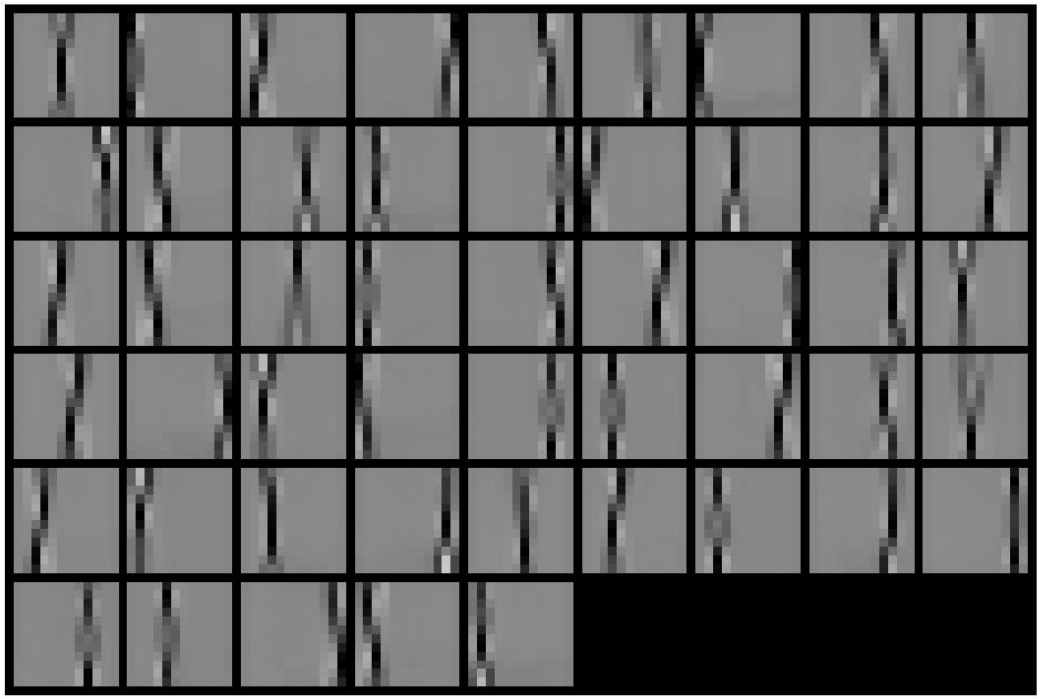
Time Series


Extracted Left Time Series
Original Shred
Extracted Right Time Series
0
50
100
150
200
250
300
350
400
0 5 10 15 20 25 30 35 40

Accuracy

- Time Series
  - At least 1 expected match in top 20% ranked matches: **39.44%**
  - At least 1 expected match in top 10% ranked matches: **18.33%**
  - At least 1 expected match in top 5% ranked matches: **8.33%**

UFL



Accuracy

- UFL
  - At least 1 expected match in top 20% ranked matches: **38.33%**
  - At least 1 expected match in top 10% ranked matches: **18.33%**
  - At least 1 expected match in top 5% ranked matches: **7.78%**

Future Work

1. Train OCRopus on better data
2. Finish building actual puzzle oracle
3. Improve synthetic oracle
   - Add character information
4. Experiment further with edge matching
   - Try different levels of smoothing on time series to see if accuracy improves
   - Try different tuning parameters on UFL
   - Try edge matching using features computed from Haroon's previous attempt at solving the puzzle
   - Apply edge matching to real chads
5. Implement OCR restriction algorithm