Semi-Automatic Reconstruction of Cross-Cut Shredded Documents

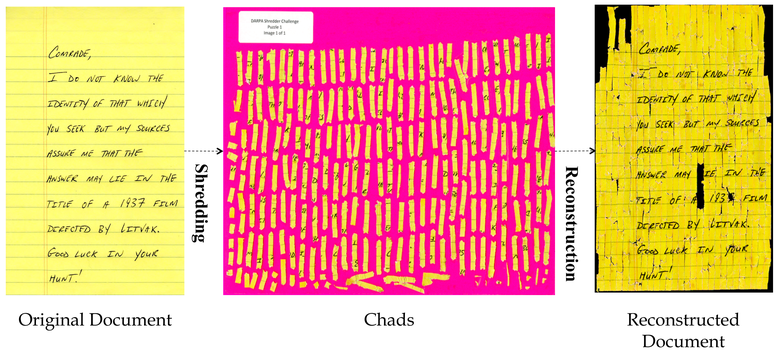
Zachary Daniels (zad309@lehigh.edu), Mentor: Haroon Idrees (haroon@eecs.ucf.edu)

University of Central Florida

UCF Center for Research in Computer Vision

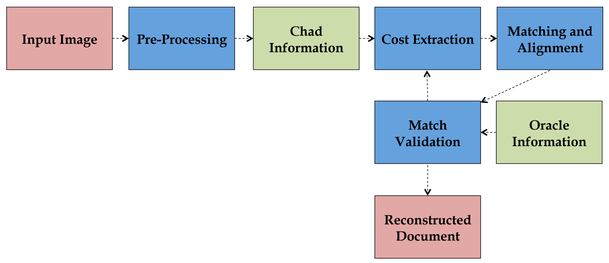
1 – Problem

- **Cross-Cut Document Reconstruction:**



Original Document | Chads | Reconstructed Document

- **Approach**
  - Build 'oracle': substitute for human input
  - Extract costs for chad matching and alignment
  - Combine chad costs and rank potential chad matches
  - Grow reconstruction, pruning bad matches using oracle
- **System Architecture**

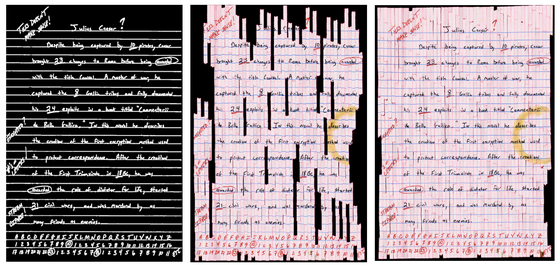


2 – Preprocessing

- Segment chads from background
- For each chad:
  - Orient
  - Find bounding box
  - Extract:
    - binary mask, derivative magnitude, derivative orientation, boundary, blurred boundary, Hough transform
  - Segment foreground, background, and lines
  - Cluster text and find top lines and baselines
  - Annotate character information
- Build bigram and trigram dictionaries

3 – Oracle

- Binarize, smooth, and clean original document
- For each chad:
  - Binarize and scale
  - Correlate across solution image
  - Find approximately correct location
- Correlate binary masks of remaining chads to fill gaps
- Adjust each chad for pixel-level precision



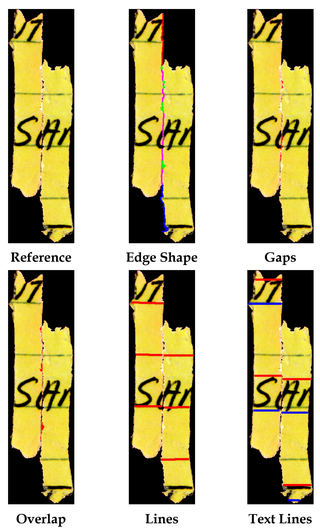
Binarized Solution | Produced by Correlation | Corrected

4 – Optical Character Recognition

- OCRopus Engine
  - Segments text lines
  - Binarizes text
  - Performs character recognition
- Accurate for chads with machine-printed text
  - Scanned pages from Google Books
  - Synthetically shredded with added noise and deformities
  - Each chad cleaned with Fred Weinhaus' textcleaner script for ImageMagick

5 – Costs

- Shape Correlation (C_S)
- Line Alignment (C_L)
- Overlap (C_O)
- Gap (C_G)
- Text Alignment (C_A)
- Histogram of Overlap Sizes (C_I)
- Histogram of Gap Sizes (C_H)
- Character Combination (C_C)
- Text Line Overlap (C_T)
- Text Line Alignment (C_X)
- Bigram Lookup (C_B)
- Trigram Lookup (C_R)



Reference | Edge Shape | Gaps

Overlap | Lines | Text Lines

6 – Reconstruction

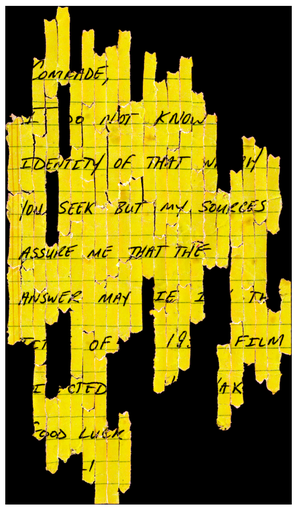
- **Pairwise Matching**
  - Calculate pairwise costs:
    - $C_1 = 0.01*C_S + 0.5*C_L + 0.0005*C_O + 0.0005*C_G + C_A$
  - Sort matches in ascending order
- **Tree-Based Expansion**
  - Find the top n pairwise matches where conditions hold:
    - One chad in match must appear in current solution
    - Overlap cost must remain below specified threshold
    - Match not discarded in past
  - Standardize costs between 0 and 1
  - Re-sort potential matches in ascending order:
    - $C_2 = 3*C_S + 2.5*C_O + 2.5*C_G + 1.5*C_I + 1.5*C_H + 0.5*C_L + 0.5*C_C + 1.5*C_T + 0.5*C_X + C_B + 0.5*C_R$
  - Find first correct match and discard all others

7 – Results

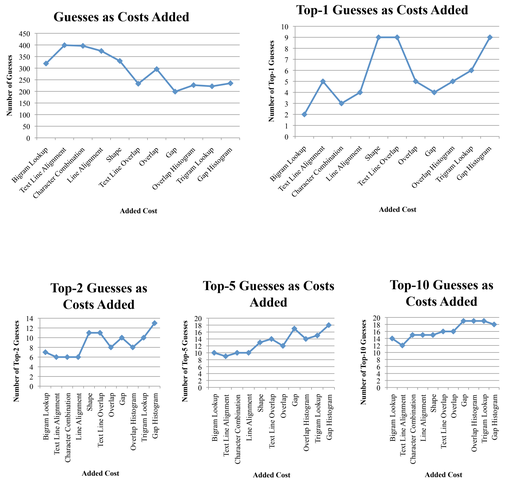
Guesses for Each Cost

Number of Chads	Shape	Line Alignment	Overlap	Gap	Overlap Histogram	Gap Histogram	Character Combo	Text Line Overlap	Text Line Alignment	Bigram Lookup	Trigram Lookup	Combined
5	17 / 3.4	39 / 7.8	29 / 5.8	37 / 7.4	27 / 5.4	37 / 7.4	71 / 14.2	37 / 7.4	31 / 6.2	47 / 9.4	210 / 42	9 / 1.8
10	105 / 10.5	228 / 22.8	220 / 22	197 / 19.7	242 / 24.2	185 / 18.5	210 / 21	140 / 14	90 / 9	95 / 9.5	354 / 35.4	18 / 1.8
15	213 / 14.2	281 / 18.7	320 / 21.3	430 / 28.7	385 / 25.7	863 / 57.3	231 / 15.4	267 / 17.8	140 / 9.3	151 / 10.1	502 / 33.5	75 / 5
20	257 / 12.85	374 / 18.7	454 / 22.7	671 / 33.55	611 / 30.55	1016 / 50.8	339 / 16.9	399 / 20	265 / 13.25	211 / 10.55	710 / 35.5	151 / 7.55
25	438 / 17.5	423 / 16.9	621 / 24.8	742 / 29.7	754 / 30.2	1258 / 50.3	411 / 16.4	471 / 18.8	383 / 15.3	320 / 12.8	802 / 32.1	235 / 9.4

*Number of Guesses / Mean Number of Guesses per Chad



82-Chad Reconstruction



8 – Future Work

- Color-Based Costs
- Adaptive Weighting of Costs
- Dictionary and Semantic Costs

9 – Acknowledgements

- This work is funded in part by NSF Grant # 1156990