Project Outline: Automatic Reconstruction of Shredded Documents

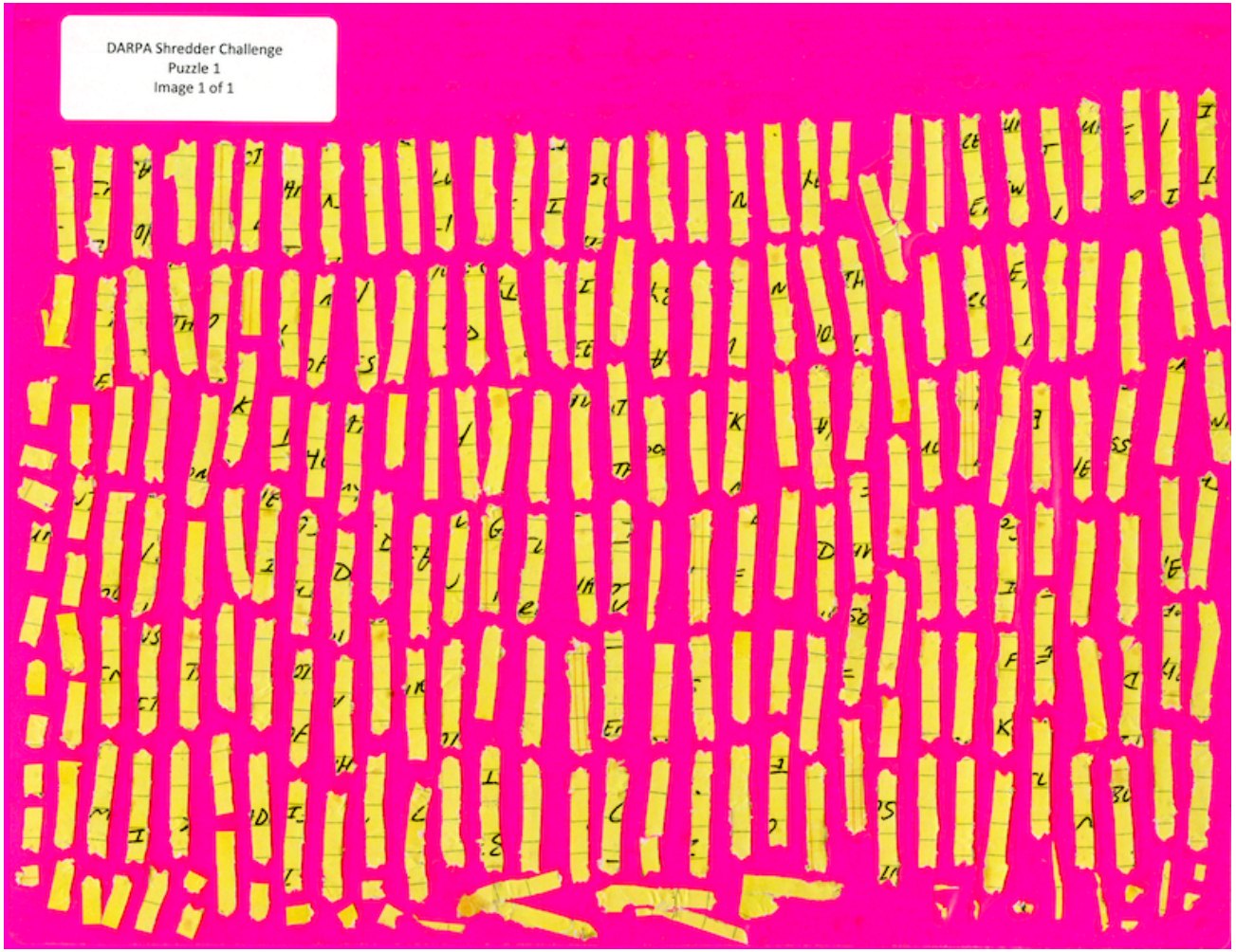
Student: Zachary Daniels

Mentor: Haroon Idrees

Problem

- Given pieces of a cross-cut shredded document, reconstruct the original document fully automatically.

Example



DARPA Shredder Challenge

- http://archive.darpa.mil/shredderchallenge/Default.aspx
- 5 "puzzles" of increasing difficulty
- Previous image was easiest puzzle
- 69 Teams competed
- Winning team was able to reconstruct 5 puzzles well enough to extract necessary information, but was far from perfect

Useful Reading

- Deever, Aaron, and Andrew Gallagher. "Semi-automatic assembly of real cross-cut shredded documents." Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012.
- Butler, Patrick, Prithwish Chakraborty, and Naren Ramakrishan. "The Deshredder: A visual analytic approach to reconstructing shredded documents." Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012.
- Zhang, Haoqi, John K. Lai, and Moritz Bächer. "Hallucination: A Mixed-Initiative Approach for Efficient Document Reconstruction." Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.
- Smith, Nathan. Automatic Character Identification for Shredded Document Reconstruction. Diss. Rochester Institute of Technology, 2013.
- Richter, Fabian, et al. "Learning to Reassemble Shredded Documents." (2013): 1-1.
- Perl, Johannes, et al. "Strip shredded document reconstruction using optical character recognition." Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on. IET, 2011.
- Lin, Huei-Yung, and Wen-Cheng Fan-Chiang. "Reconstruction of shredded document based on image feature matching." Expert Systems with Applications 39.3 (2012): 3324-3332.
- Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." Proceedings of the 25th international conference on Machine learning. ACM, 2008.
- Harding, Stephen M., W. Bruce Croft, and C. Weir. "Probabilistic retrieval of OCR degraded text using n-grams." Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 1997. 345-359.

Approach

- Combine edge matching with optical character recognition (OCR)
  - Check for ‘legal’ character combinations and words

Basic Steps of Document Reconstruction

- Preprocessing
  - Segmentation
  - Adjusting orientation
- Feature extraction
- Feature matching
- Alignment

Current State of The System

- Pairwise matching of shred pieces
  - Returns too many potential matches
- GUI for human validation

Next Steps

- Implement OCR to further restrict the solution space
  - First explorations:
    - Synthetic dataset using typed fonts instead of handwritten characters
    - Well-established OCR software: Tesseract
      - Written in C++; Matlab wrapper exists but might be of questionable quality
      - Additional advantage: trainable

Concerns

- Concerns:
  - Tesseract is very good at 'clean' scripts with disconnected characters and little variation between documents (i.e. fonts)
    - Handwriting often doesn't meet these requirements
  - Cross-cut shredding is more difficult to simulate than traditional strip shredding

Preprocessing

- Segmentation
  - Threshold on the background
  - Apply connected components algorithm
- Adjusting orientation
  - Potential Solutions:
    - Rotation based on principal components
    - Rotation based on minimum-area-rectangle
    - “Arrows and notches”

Feature Extraction

- Edge features
  - Potential solutions:
    - Traditional: SIFT-based, rule-alignment, color information, edge-based time series, etc.
      - A lot of literature on edge features
      - Might also benefit by examining content features
    - New: unsupervised feature learning
- Semantic approach
  - OCR validation
    - Simple dictionary
    - n-grams

Feature matching / Alignment

- Likely algorithms for feature matching:
  - Nearest Neighbor
  - Support Vector Machines
- A few potential algorithms exist for alignment

Next steps

- Create synthetic dataset to test existing OCR libraries
- Select dictionary (for word-lookup) or n-gram set
- Implement OCR on shreds
- Develop OCR-based shred-match validation algorithm
- Expand to full DARPA dataset

Thank you!

- Questions?