Existing Technologies Survey

Image Segmentation: from images to image blocks

- "Localizing and Segmenting Text in Images and Videos" (2002)
  - **Topics**: Detect, segment, recognize text in web pages
  - **Upshot**: newspapers are like webpages? Also, offers wealth of *references*!
- "Rapid Object Detection using a Boosted Cascade of Simple Features" (2001)
  - **Topics**: efficient location of a given pattern of light and dark rectangles within image.
  - **Upshot**: newspapers are segmented by whitespace and divider bars: light and dark rectangles.
- "A Hough Transform based Technique for Text Segmentation" (2010)
  - **Topics**: hough-transforms allow for rotation-robust segmentation of text into lines, words, characters.
  - **Upshot**: increase robustness against the typesetting imperfections of historic newspapers.
- **"Integrated algorithms for newspaper page decomposition and article tracking"** (1999)
  - **Topics:** segmenting newspaper page images into various objects(text, images and drawings, titles)
  - **Upshot:**
-

Optical Character Recognition: from image blocks to text

- "Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs" (2009)
  - **Topics**:
  - **Upshot**:

Natural Language Processing: from text to understanding

- "Using Information Extraction to Classify Newspapers Advertisements" (2000)
  - **Topics**:
  - **Upshot**:
- "A survey of text clustering algorithms"
  - **Topics**: group articles into different categories
  - I'm not sure if text clustering (group those articles into different categories) is too much for this project. This chapter describes many methods though.
    - Hey, good idea! Text clustering will make our output significantly more useful to the newspaper-philes who'll use our product. It might also help validate our model and make it more accurate (via sentence-level clustering: an image captioned with "moon landing"

more likely goes with the paragraph "...astronaut..." than with "...blue is the new magenta..."). Let's keep text clustering boiling in the backs of our heads until we've completed a basic version of the project? -sam

Newspaper Format/Layout and Metadata

- "Document Structure and Layout Analysis"
  -
- "A machine-learning approach for analyzing document layout structures with two reading orders"
  -
-