Google Newspaper Search – Image Processing and Analysis Pipeline

Krishnendu Chaudhury, Ankur Jain, Sriram Thirthala, Vivek Sahasranaman, Shobhit Saxena,
Selvam Mahalingam
Google Engineering
krish@google.com, jainaj@google.com, tvnsriram@google.com, svivek@google.com,
selvam@google.com

Abstract

The Google Newspaper Search program was launched on September 8, 2008[1]. In this paper, we outline the technology pieces underlying this large and complex project. We have created a production pipeline which takes newspaper microfilms as input and emits individual news articles as output. These articles are then indexed and added to the content base, so that they turn up in response to Google searches. Thus, in response to a Google query "Hitler death", we are able to show newspaper articles from the very day it was reported..

Non-uniform illumination, presence of significant noise, tears and scratches in the microfilm image, all pose special challenges for this project. The significant variation of layouts across newspapers and time eras, the variations in font sizes occurring in a single page (which confuses the OCR engine) compound the difficulties. The project is still going on after the initial launch was made (with about 15 million news articles).

1. Introduction

Google Newspaper Digitization, Indexing and Search program is an ambitious attempt to bring online a significant portion of human history, as reported at the time of its occurrence. Starting from archived microfilms corresponding to past newspaper editions, html news articles get generated which are indexed for subsequent search and retrieval. In this context, it is worth noting that in order to build a searchable index from archived images of newspaper pages, it is not enough to simply do OCR on the entire page and dump the resulting words in the index. The sheer variety of words and topics found on a newspaper page would confuse any system that attempts to rank and/or cluster them. Instead, it is desirable to segment the page into separate news articles and treat these articles (as opposed to the entire page) as individual items for indexing. Thus, article segmentation, extraction of individual articles from the page image is an important topic in this paper [3].

Another equally important topic is binding, which is the process of collecting pages from the same date of a given newspaper (edition) together. Binding allows us to tag each news article with its date of publication [2].

The authors would like to take this opportunity to thank *Dan Bloomberg, Adam Langley, Ray Smith* and *Luc Vincent* for their advice and support.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 outlines the algorithms and systems. Section 4 shows results.

.

2. Related Work

Baird [4] developed a system in which white space is covered greedily by rectangles until all text blocks are isolated. Like him, we are also motivated by the maxims, "Background is simpler than foreground", "white space is a layout delimiter" (we also add long vertical and horizontal lines to the list of layout delimiters). Breuel [5] also presents approaches for covering the background whitespace of documents in terms of maximal empty rectangles. Our approach however, does not depend on rectangular covers for the white space. Due to noise and non-uniform illumination on newspaper page images, white spaces detected are usually imperfect and rectangular cover based approaches fail.

In 2003, 2005, 2007 ICDAR held the page segmentation competition [10], [15], [16]. Notable entries there were the classifier based DAN system [11], the connected component based Oce system and the morphology based ISI system [12]. Antonocopoulos developed a background description based page segmentation approach [17]. We have been inspired by all these systems.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.272

IEEE computer society

Finally, the core image processing library used in the project is Leptonica [13].

3. Algorithms and System Descriptions

Fig. 1 shows the overall system architecture. The input to the system is a microfilm roll. By scanning it, we typically get a very wide image corresponding to about a month's worth of newspaper pages laid side by side in increasing order of date. This image is processed by the backend pipeline shown in Fig. 1.

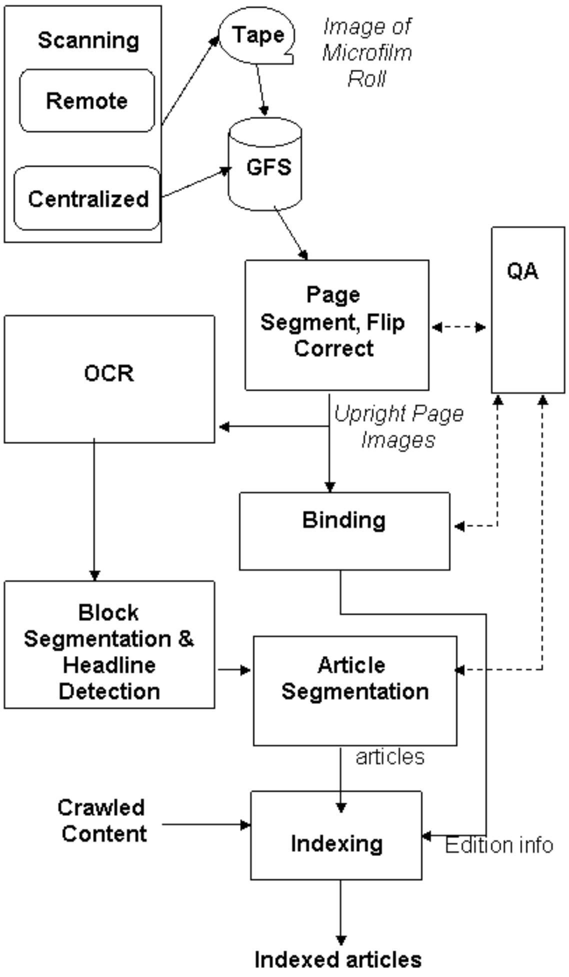


Fig. 1: System Architecture

Details appear in following sections.

3.1. Page Segmentation

This module extracts individual pages from the wide image corresponding to the entire microfilm roll. The wide image has newspaper pages (dark foreground on lighter background) separated by dark strips.

Consequently, our page segmenter essentially recognizes connected components of background color on the wide image. Then it eliminates the components that are too small and the remaining connected components are pages. Once pages are extracted, rest of the pipeline deals with pages only.

3.2. Flip Correction

Newspaper pages can and do get flipped (lateral inversion, 180 or 90 degrees rotation etc.) during the microfilming process. We have an automated system to fix this, using the fact that only the correct orientation would lead to valid, dictionary words from OCR. Since OCR is expensive, we prune the search space by utilizing the fact that newspaper blocks typically have uniform widths (up to some fuzz factor). Hence, we do a crude and fast block segmentation (which identifies blocks of foreground text) and compute a histogram of block widths. If the histogram lacks a sharp peak, we rotate the page image by 90 degrees. Subsequently, we do not need to explore the orthogonal (landscape) orientations. Even among the portrait orientations, we OCR only the three tallest blocks from the histogram peak, as they are most likely to be text blocks.

3.3. Binding

Binding refers to the process of collecting together the newspaper pages belonging to the same date (aka same edition). Now, in a typical microfilm, pages from a given edition appear contiguously and sequentially. Hence, if we identify all the front pages in a microfilm, binding effectively reduces to collecting together all pages from one front page up to, but not including, the next. Thus, the core task in binding is *front page identification*. For that, we manually obtain one sample/template front page image from every microfilm roll. Other front pages are obtained by matching against this template.

The matching is done via techniques for *object detection in cluttered environment*. In all the front pages of a given newspaper, the newspaper title (e.g., a stylized rendition of "Wall Street Journal") and perhaps some unique logo will appear. These are the objects we try to recognize in the presence of clutter (everything else on the front page is clutter). On each microfilm, one template front page is identified manually. The remaining front pages are obtained by comparing against this.

Object recognition is done in 2 steps:

1. **Feature Detection and Description**: Features are detected by convolving the image with the *Gabor wavelet* [14]
$$p_k(x) = \frac{|k|^2}{\sigma^2} e^{\frac{|k|^2}{\sigma^2}|x-x_0|^2} e^{ik\cdot\cdot(x-x_0)}$$ where amplitudes of responses yield components of descriptor vector.
2. **Identifying maximal set of consistent feature matches**: The maximal set of consistent feature matches is obtained via the RANSAC (*Random Sampling Consensus*) algorithm (feature matches are consistent if they subscribe to the *same affine transformation*).

3.4. Image Cleaning

Newspaper page images obtained from microfilms have non-uniform illumination and extremely high levels of noise. Background and foreground (text/pictures) gray levels vary significantly, from one portion of a page to another portion of the same page. Without cleaning, such images are unsuitable for display and/or OCR. Our image cleaning approach is based upon a novel image binarization technique. Obviously, global threshold based binarization is not suitable here. Our image binarizer is local in nature and is based on *morphological grayscale reconstruction* [7]. In the following discussion, we assume (without loss of generality) that the foreground is whiter than background.

Our approach is based on the assumption that there will be a minimum contrast between foreground and background gray levels. In other words, the foreground profiles will more or less look like a peak/dome above the background. Our fore ground detector is essentially an H-dome detector [13]. The entire process is described in FIGURE 2. One result is shown in FIGURE 3.

Once we have identified the foreground and background pixels from the binary image, we paint all background pixels with saturated white. We do *not* paint all the foreground pixels with saturated black, however, to avoid aliasing artifacts. Instead, foreground grey levels get mapped to one of 4 values at the dark end of the spectrum.

3.5. OCR

We use a third party OCR engine. Despite being one of the leading OCR engines of the world, it makes many mistakes on newspaper page images. This is due to the high noise level, non-uniform illumination, tears and scratches and extreme variations in font sizes on a newspaper page.

In particular, the OCR engine often mistakes large headlines on newspaper pages as pictures. To mitigate this, we OCR the page image, erase all detected text, scale the image down and re-OCR.

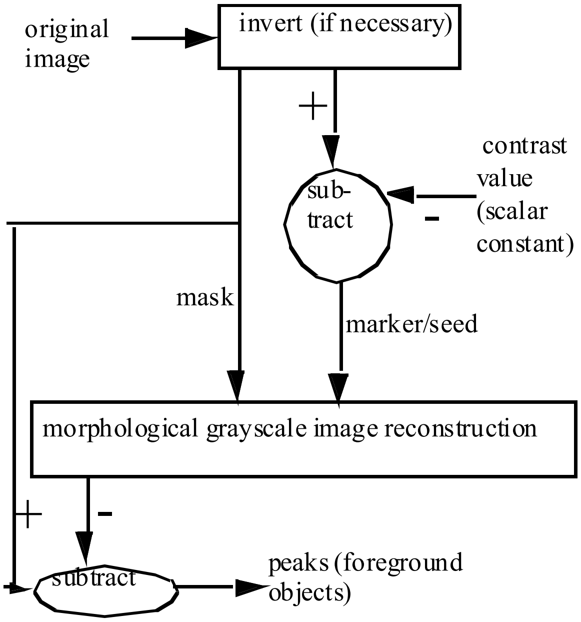
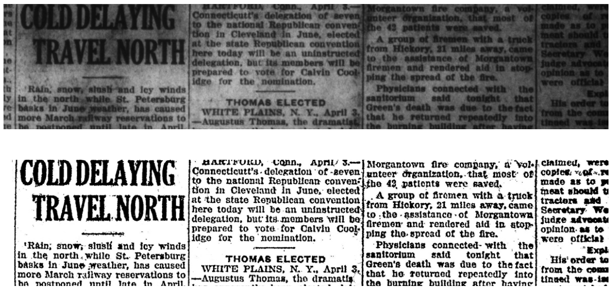

original image
invert (if necessary)
+
sub-tract
contrast value (scalar constant)
-
mask
marker/seed
morphological grayscale image reconstruction
+
-
subtract
peaks (foreground objects)

Fig. 2: Morphological Grayscale Reconstruction based Image Binarization



COLD DELAYING TRAVEL NORTH

'Rain, snow, slush and icy winds in the north while St. Petersburg basks in June weather, has caused more March railway reservations to be postponed until late in April.

HARTFORD, Conn., April 3.—Connecticutt's delegation of seven to the national Republican convention in Cleveland in June, elected at the state Republican convention here today will be an uninstructed delegation, but its members will be prepared to vote for Calvin Coolidge for the nomination.

THOMAS ELECTED

WHITE PLAINS, N. Y., April 3.—Augustus Thomas, the dramatist

Morgantown fire company, a volunteer organization, that most of the 42 patients were saved.

A group of firemen with a truck from Hickory, 21 miles away, came to the assistance of Morgantown firemen and rendered aid in stopping the spread of the fire.

Physicians connected with the sanitorium said tonight that Green's death was due to the fact that he returned repeatedly into the burning building after having

claimed, were copies of made as to ment should tractors and Secretary We judge advocate opinion as to were official

Expl

His order to from the com tinued was is

Fig. 3: Image Cleaning Result

3.6. Block Segmentation, Headline Detection and Article Segmentation

Our article segmentation involves the following steps:

1. **Block Segmentation**: Identify text blocks using gutters, lines on the page image. We use structuring elements like gutters and lines in the newspaper page for this purpose. A gutter

is a tall, narrow or short, wide strip of background separating blocks of text. Special image filters have been developed for gutter detection – they essentially compute the fraction of background pixels in the neighborhood to determine whether a given pixel belongs to gutter. Lines are detected in analogous fashion.

2. **Headline Detection** - Classify above blocks into headlines and body-text, using OCR reported font size and area-perimeter ratio of connected components as cue. Fig. 4 shows a result of block segmentation and headline detection.
3. **Binary Classifier**: Classify all neighboring body-text block pairs into two sets: (i) belonging to same article (ii) belonging to different article. We have experimented with the CART classifier [9] and a Rule Based classifier. Eventually the rule based classifier outperformed the CART based one and is currently deployed. It has two dominant rules:

   a**) Common Headline Rule**: Body text blocks under the same headline block belong to same article. This rule is extremely powerful and in many cases may be sufficient on its own (see Fig. 5 for instance).

   b) **Orphan Block Rule**: Orphan blocks are blocks with no headline above them. Examples of such blocks can be seen in 2nd, 3rd, 4th, 5th and 6th columns of Fig. 6. Given an orphan block directly below a line spanning multiple blocks or at the top of the page, we link it with the non-orphan block whose bottom is below the top margin of the orphan block *and* there is no other block between the two. Also, vertically overlapping orphan blocks belong to the same article.
4. **Transitive Closure** on the block pairs belonging to same article. Each closed set of body-text blocks constitute one individual article. Add the appropriate headline block to the set and we have the complete article.

4. Results

Fig. 5, 6 show some *article segmentation results*. Articles are color coded (headlines are shown with a deeper shade of same color as article). Overall article segmentation accuracy is ~90% (measured against manual ground truth). Overall OCR accuracy (in terms of fraction of dictionary words on page) is ~80%.

5. References

[1] P. Soni, “Bringing history online, one newspaper at a time”, http://googleblog.blogspot.com/2008/09/bringing-history-online-one-newspaper.html, Google Blog, Sept. 8, 2008.

[2] Sriram Thirthala, Krish Chaudhury, “Identifying Front Page in Media Material”, *pending Google patent application*, filed Aug 12, 2008.

[3] Ankur Jain, Vivek Sahasranaman, Shobhit Saxena, Krish Chaudhury, “Segmenting Printed Media into articles”, *pending Google patent application*, filed Aug 13, 2008.

[4] H.S. Baird, “Background Structure in Document Images”, *Document Image Analysis*, World Scientific, Singapore, 1994, pp. 17-34.

[5] Thomas Breuel, “Two Geometric Algorithms for Layout Analysis”, *Proceedings of the workshop on Document Analysis Systems*, Princeton, NJ, USA, 2002, pp. 188-199.

[6] Thomas Breuel, “Robust least-square baseline finding using branch and bound algorithm”, *Proceedings of the SPIE*, 2002.

[7] Luc Vincent, “Morphological Grayscale Reconstruction in Image Analysis: Application and Efficient Algorithm”, *IEEE Trans. On Image Processing*, vol. 2, No. 2, April 1993, pp. 176-201.

[8] Hartley, R., Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, 2003.

[9] Bishop, C., *Pattern Recognition and Machine Learning*, Springer, 2006.

[10] A. Anotonacopoulos, G. Gatos, D. Karatzas, “ICDAR 2003 page segmentation competition”, *proc. of seventh intl. Conf. On Document Image Analysis and Recognition*, ICDAR, Edinburgh, Scotland, 2003.

[11] Cinque., S. Levialdi, A. Malizia, F. Rosa, “DAN: an automatic segmentation and classification engine for paper documents”, *proc. of fifth IAPR intl. Workshop On Document Analysis Systems*, Princeton, NJ, USA, Aug. 2002, pp. 491-502.

[12] A. Das, S. Chowdhuri, B. Chanda, “A complete system for document image segmentation”, *proc. natl. workshop on computer vision, graphics and image processing (WVGIP)*, Madurai, India, Feb. 2002, pp. 9-16.

[13] Bloomberg, Dan., *Leptonica: An open source C library for efficient image processing, analysis and operation*, http://code.google.com/p/leptonica/.

[14] Ulrich Buddameyer, Hartmut Neven, "Systems and Method for Descriptor Vector Computation", *pending Google patent application*.

[15] A. Anotonacopoulos, G. Gatos, D. Bridson, "ICDAR 2005 page segmentation competition", *proc. of eighth intl. Conf. On Document Image Analysis and Recognition*, ICDAR, Seoul, South Korea, 2005, pp. 75-79.

[16] A. Anotonacopoulos, G. Gatos, D. Bridson, "ICDAR 2007 page segmentation competition", *proc. of nineth intl. Conf. On Document Image Analysis and Recognition*, ICDAR, Curitiba, Brazil, 2007, pp. 1279-1283.

[17] A. Anotonacopoulos "Page Segmentation Using the Description of the Background", *Computer Vision and Image Understanding*, vol. 70, No. 3, 1998, pp. 350-369.

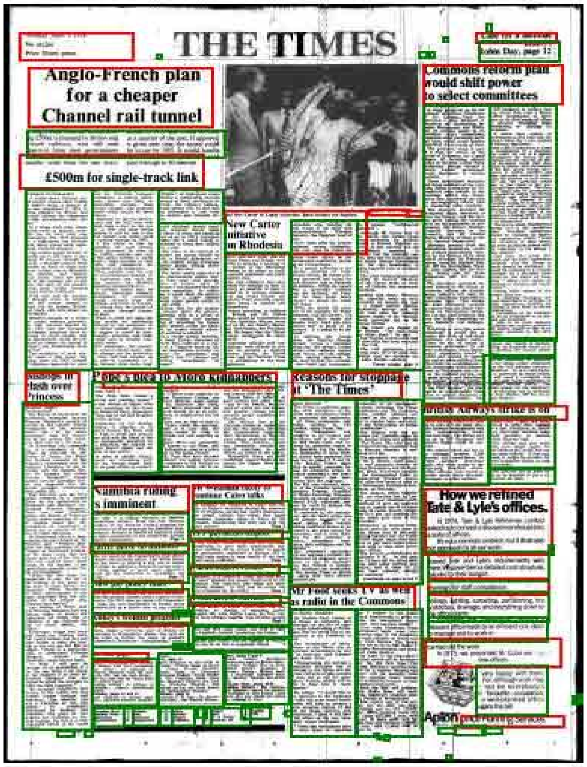

THE TIMES
Anglo-French plan for a cheaper Channel rail tunnel
£500m for single-track link
Commons reform plan would shift power to select committees
New Carter initiative on Rhodesia
Pope's plea to Moro kidnappers
Reasons for stoppage at 'The Times'
British Airways strike is on
Namibia ruling is imminent
How we refined Tate & Lyle's offices.

Fig. 4: Block Segmentation and Headline Detection Result (green = body-text, red = headline)

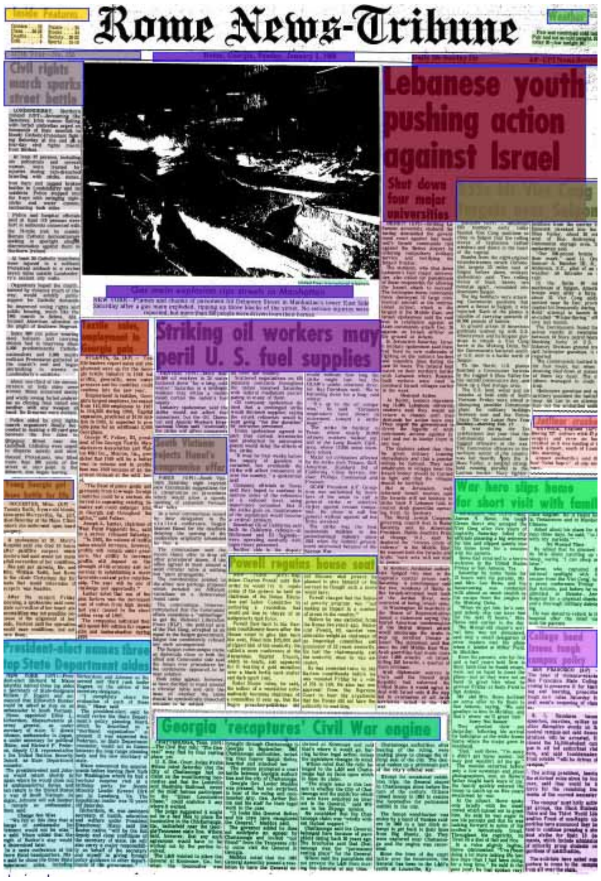

Rome News-Tribune
Civil rights march sparks street battle
Lebanese youth pushing action against Israel
Shut down four major universities
Striking oil workers may peril U. S. fuel supplies
Powell regains house seat
Georgia 'recaptures' Civil War engine
President-elect names three top State Department aides
War hero slips home for short visit with family

Fig. 5: Article Segmentation result (common headline rule)

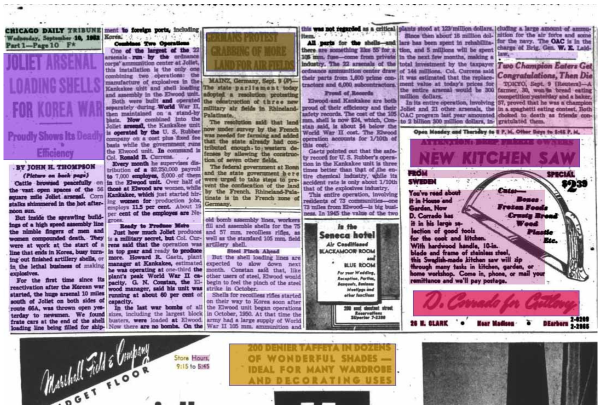

CHICAGO DAILY TRIBUNE
JOLIET ARSENAL LOADING SHELLS FOR KOREA WAR
Proudly Shows Its Deadly Efficiency
GERMANS PROTEST GRABBING OF MORE LAND FOR AIR FIELDS
Two Champion Eaters Get Congratulations, Then Die
NEW KITCHEN SAW
D. Corrado for Cutlery
Marshall Field & Company
200 DENIER TAFFETA IN DOZENS OF WONDERFUL SHADES — IDEAL FOR MANY WARDROBE AND DECORATING USES

Fig. 6: Article Segmentation result (Orphan Block rule – purple article)