



Document Segmentation with Applications to the Book Publishing Industry

Dan Chevion, Ehud D. Karnin, Asaf Tzadok,
IBM Labs in Haifa

October 14, 2004



Contents

- ◇ Introduction
- ◇ Stages of the process
- ◇ Pictorial description of the process
- ◇ Image segmentation
 - ◇ Detection
 - ◇ Continuous tone / line-art separation
 - ◇ De-screening and compression
- ◇ Text segmentation
 - ◇ Detection
 - ◇ Skew finding
 - ◇ Enhancement and compression
- ◇ Line art extraction, processing and compression
- ◇ Experimental results
- ◇ Manual editing
- ◇ Summary



Introduction

- ◆ In many scenarios documents are available in paper form only, and there is a business need to digitize them.
 - ◆ Books publishers need to reprint out of print titles
 - ◆ High volume scanners generally support a maximum resolution of 300dpi, and use 8 bit gray-scale mode to optimize image quality
- ◆ Direct print of the scanned images would result in a poor printout
 - ◆ Digital printers generate 600 dpi bi-level image by halftoning the gray scale input: the text will not be sharp, and the image region will show Moiré patterns
- ◆ The proposed solution: Separate the different regions, so that they would undergo different image processing methods, targeted at alleviating their specific potential artifacts



The Stages in Document Handling

- ◇ Paper scan
- ◇ Preprocessing
- ◇ Automatic page segmentation
- ◇ Potential manual editing
- ◇ Quality enhancement (per region)
- ◇ Compression
- ◇ Assembling



Pictorial Description of the Segmentation of Scanned Pages

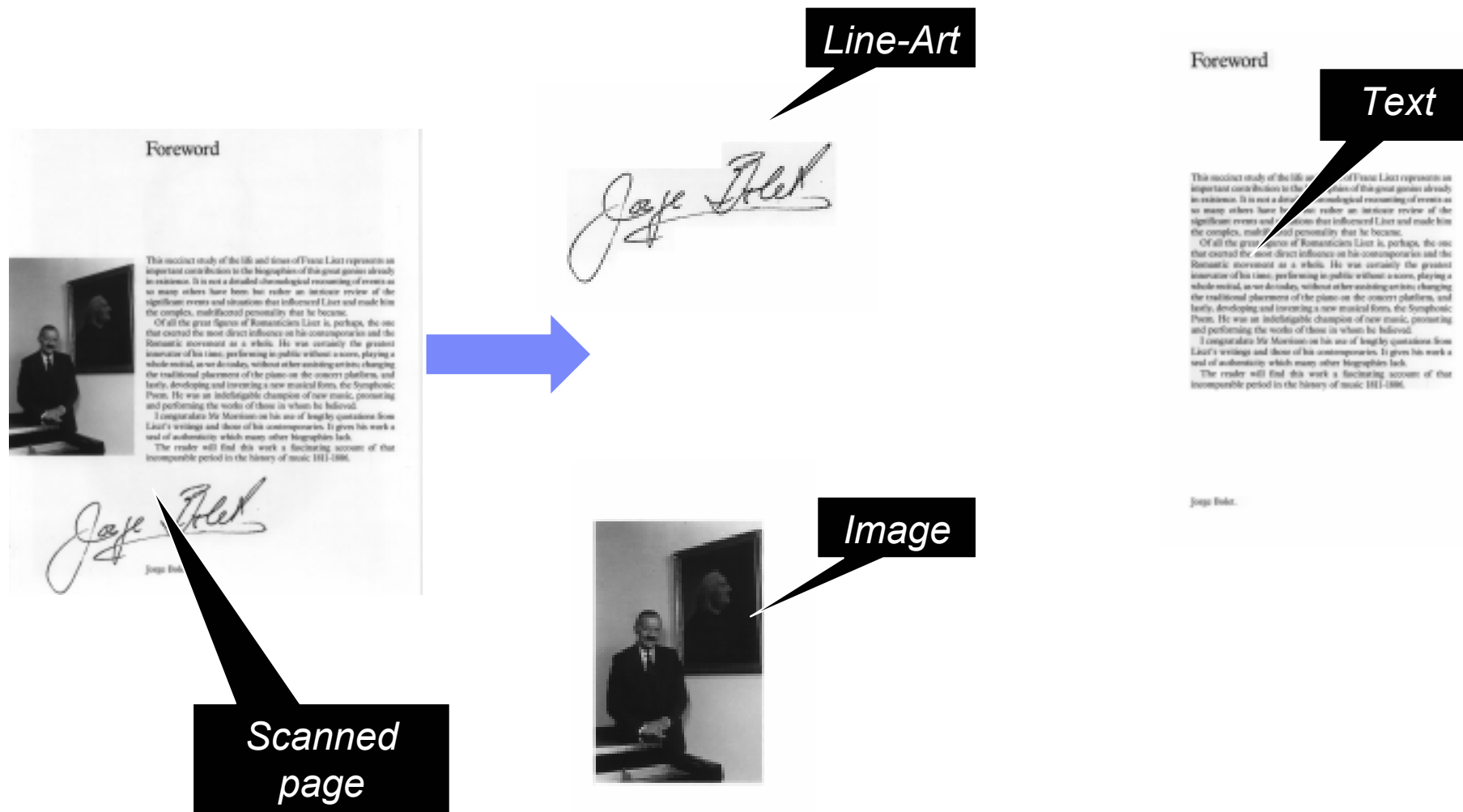




Image Segmentation

- ◆ Images are best detected at low resolution, being relatively large objects
 - ◆ However, high resolution refinement is needed to avoid blockiness
- ◆ Hence we use a linked pyramid structure as a framework for an iterative process
 - ◆ Pioneered by Burt and Rosenfeld, 1981
- ◆ The ground level is at the original resolution
 - ◆ A measure assigned to each pixel - the number of different colors in its 4 x 4 vicinity
- ◆ The pyramid is built bottom up in an iterative manner:
 1. Father's value in layer k is the average of its 4x4 sons in layer $k-1$
 2. For each son keep the link only to the closest father
 3. Re-compute father's value by averaging the remaining sons
 4. Project father's value to his sons (top down)
 5. Iterate on steps 2 - 4

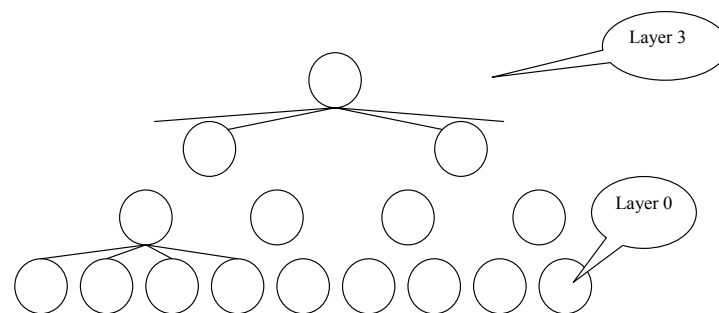




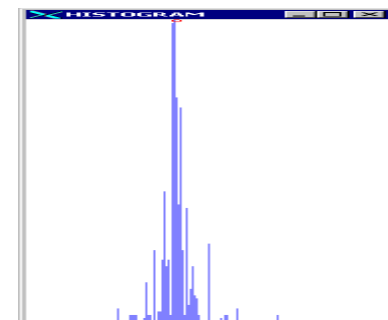
Image Segmentation (continued)

- ◆ Apply a threshold to the ground level map (resulting from the iterative process) to detect image regions
 - ◆ Color or grey levels are varying in image areas, hence they exceed the threshold
 - ◆ Line art, and certainly text, tend to have a single color across a large area.
- ◆ Use connected component approach to further analyze the resulting binary mask
 - ◆ Connect the basic components to large compound objects
 - ◆ Study the topology to better classify as image or line-art: Typically line-art are skeleton-like while image regions are fat bodies with potentially small holes
- ◆ Image regions are de-screened, and JPEG compressed.



Text Segmentation

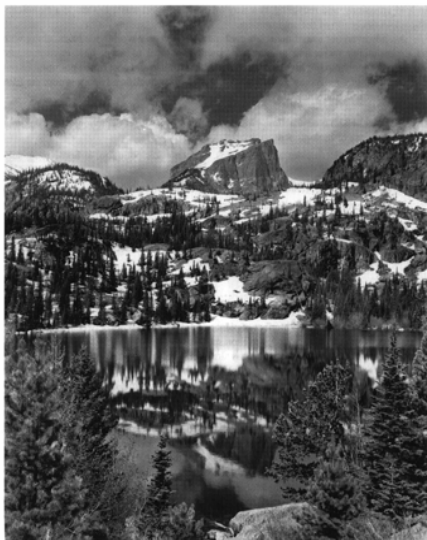
- ◆ Text is extracted in a positive manner in a bottom-up approach
 - ◆ Adjacent pixels that exceed a certain level of darkness are grouped to a component
 - ◆ Several features are assigned to each component, including a set of vectors pointing to neighboring components
 - ◆ An histogram of the angles of these vectors is generated, and its peak detects the skew
 - ◆ Links in the skew direction correspond to lines, with words characterized by shorter links between its members
 - ◆ Paragraph structure is revealed by going perpendicular
 - ◆ The combination of directional information with other layout properties (e.g., adjacent characters have similar heights) detect the text regions.
- ◆ Text areas are up-scaled to 600 dpi, binarized, and compressed by G4
- ◆ The remaining areas are considered line-art
 - ◆ They are not necessarily solid black, hence halftoned by error diffusion, and compressed using LZ or JBIG





Experimental Results: original scans

Part III: Ecology 123

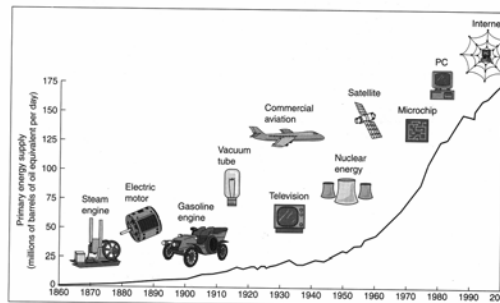


The power and resources of our delicate environment. Photo courtesy of Union Pacific Railroad.

When the environment is finally forced to file for bankruptcy because its resource base has been polluted, degraded, dissipated, and irretrievably compromised, the economy will go down with it.

TIMOTHY WIRTH, U.S. DEPARTMENT OF STATE

Energy



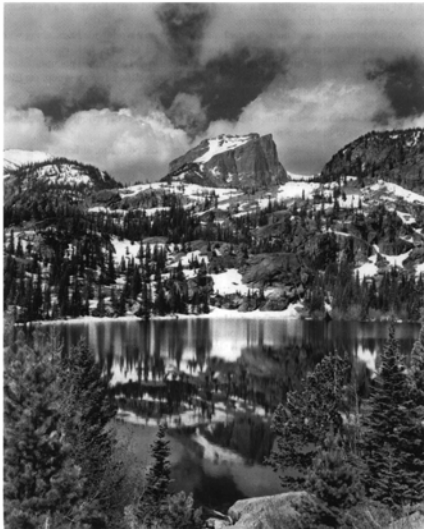
- 5 **Energy**
Ian Barbour
- 6 **Beyond the Gulf Crisis: An Energy Strategy for the '90s**
Christopher Flavin
- 7 **The Immortal Waste**
Ahmed S. Khan and Barbara A Eichler
- 8 **Hydroelectric Power**
Ahmed S. Khan and Barbara A Eichler
- 9 **Solar Energy**
Ahmed S. Khan and Barbara A. Eichler
- 10 **Wind Power**
Ahmed S. Khan and Barbara A. Eichler
- 11 **Wind Power: Small, But Growing Fast**
Christopher Flavin
- 12 **The Electric Car Arrives—Again**
Seth Dunn

43



Experimental Results: segmentation of page 1

Part III: Ecology 123



The power and resources of our delicate environment. Photo courtesy of Union Pacific Railroad.

When the environment is finally forced to file for bankruptcy because its resource base has been polluted, degraded, dissipated, and irretrievably compromised, the economy will go down with it.

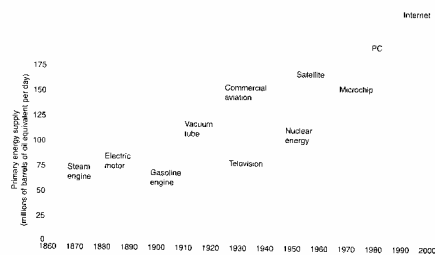
TIMOTHY WIRTH U.S. DEPARTMENT OF STATE



Experimental Results: segmentation of page 2

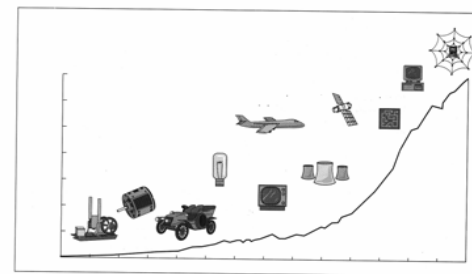


Energy



- | | |
|--|---|
| 5 <i>Energy</i>
Ian Barbour | 9 <i>Solar Energy</i>
Ahmed S. Khan and Barbara A. Eichler |
| 6 <i>Beyond the Gulf Crisis: An Energy Strategy for the '90s</i>
Christopher Flavin | 10 <i>Wind Power</i>
Ahmed S. Khan and Barbara A. Eichler |
| 7 <i>The Immortal Waste</i>
Ahmed S. Khan and Barbara A. Eichler | 11 <i>Wind Power: Small, But Growing Fast</i>
Christopher Flavin |
| 8 <i>Hydroelectric Power</i>
Ahmed S. Khan and Barbara A. Eichler | 12 <i>The Electric Car Arrives Again</i>
Seith Dunn |

43



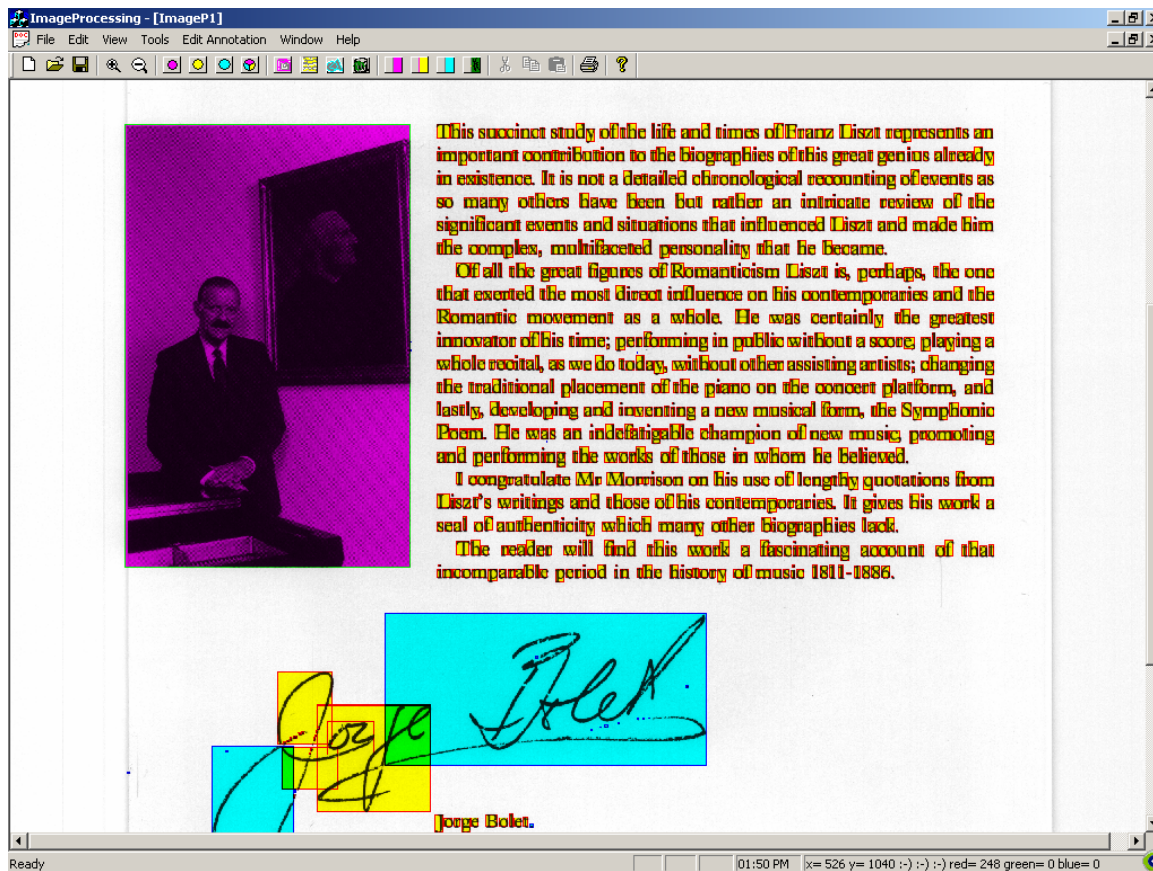


Manual Editing

- ◆ Manual editing may be applied to correct misclassifications and to override the decision of the automatic system
- ◆ A Windows/ MFC environment
 - ◆ Powered by our own image processing functions together with an image processing library from Visionary Solutions, Inc.
- ◆ Connected-component based segmentation editor
 - ◆ Colorful “segmentation view” on top of the original document (yellow for text, magenta for image, cyan for line-art)
 - ◆ Selection of a region – no need for exact mouse dragging, as the region is built up by connected components
 - ◆ Change the category of the selected region (re-color the region), e.g., from text to line-art



Manual Editing - example





Summary

- ◆ The components of the segmented page are reassembled into a single PDF file
 - ◆ The size is much smaller than the original scan (as the compression algorithm is matched to the component)
 - ◆ Yields efficient storage
 - ◆ Printed pages are of superior quality compared to a straightforward print of the scan