4. Images of Pages

4. Image of pages

4.1 Scanning pages

4.2 Image conversion

4.3 Indexing image of pages

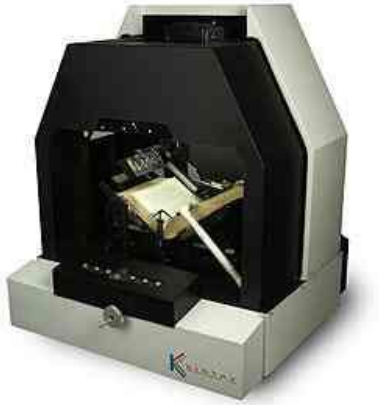
4.4 Shared text image/system

4.5 Large scale projects

4.1 Scanning of Pages

- Most economical way for converting analog material into machine-readable form (many times cheaper than keying) is *scanning*
- Different types of scanning machines exist (page turning scanning machines good for library use)

Digitizing Line
About $300,000.



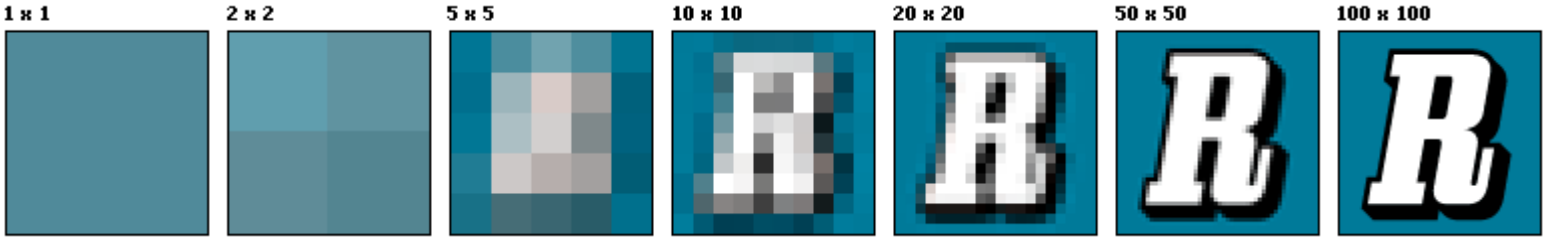
Kirtas Bookscan
About $150,000.

- Microfilm scanners – for scanning legacy microfilms

4.1 Quality of scanning

- The readability of the text
  - ✓ resolution – dots per inch

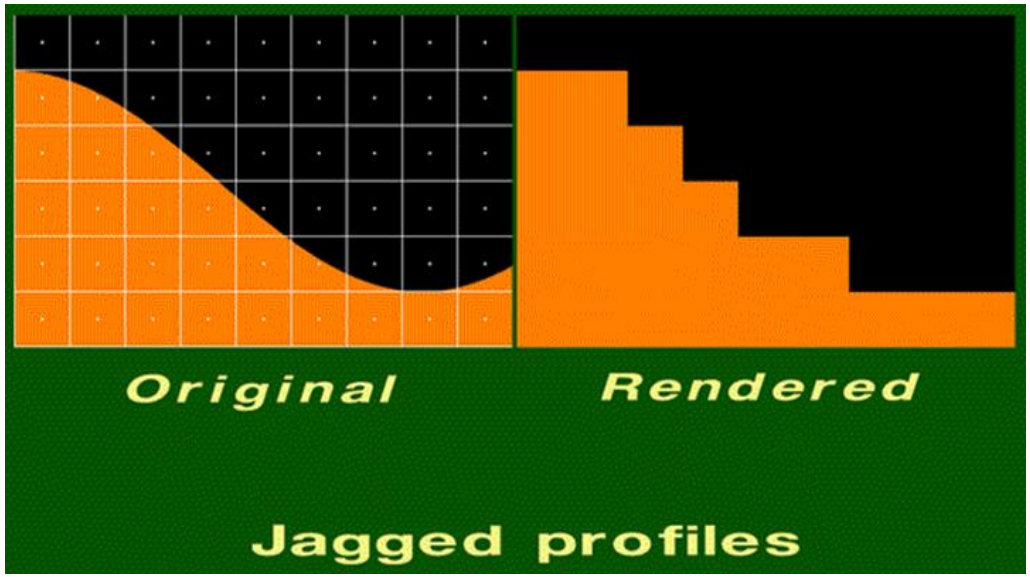
  Scanning at 118 dots/centimeter (300 dpi) is adequate for conversion to digital text output, but for archival reproduction of rare, elaborate or illustrated books, much higher resolution is used.



  - ✓ type size

4.1 Improve the quality of scanning - antialiasing

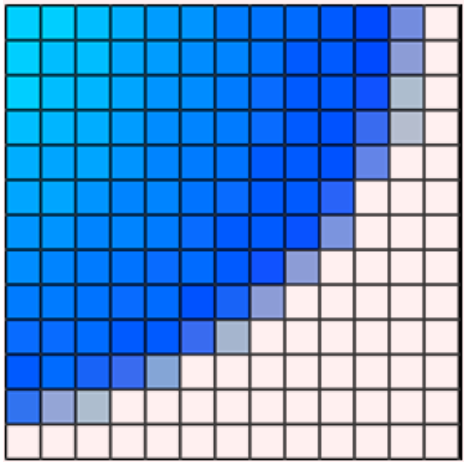
- Jagged profiles - distortion artifacts known as aliasing when representing a high-resolution image at a lower resolution.



- Antialiasing techniques - fooling the eyes that such jagged edges look more smooth.

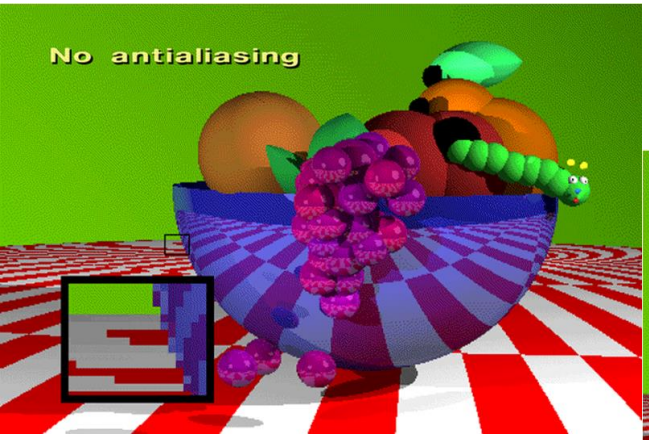
4.1 Improve the quality of scanning -antialiasing

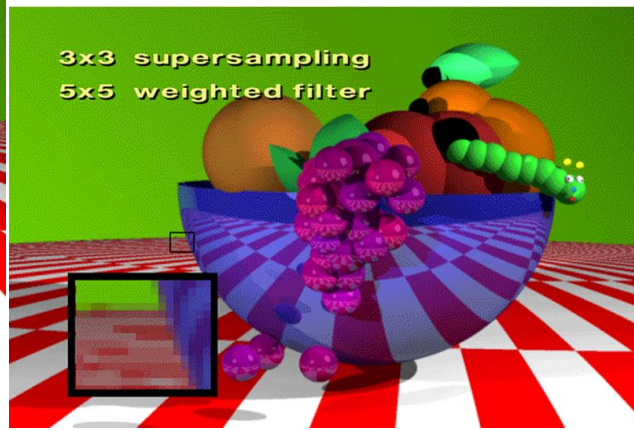
- **Pre-filtering** - treat a pixel as an area, and compute pixel grey scale based on the overlap of the object within a pixel's area



4.1 Improve the quality of scanning -antialiasing

- **Post-filtering** (super sampling) - increasing the frequency of the sampling grid and then averaging the results down





4.1 Image of Pages

- Advantage:
  - familiarity and easy creation
- Disadvantage:
  - bigger in size (2,000 bytes text vs. 30,000 bytes image)
  - less adaptable (font, size, etc.)
  - copy and paste text difficult (if not impossible)
  - separate index or **Optical Character Recognition** file needs to be developed for searching

4.1 Images of pages

4.1 Scanning pages

4.2 Image conversion

4.3 Indexing image of pages

4.4 Shared text image/system

4.5 Large scale projects

4.1 Image conversion

Optical Character Recognition

- technology that can convert different types of documents, such as scanned paper documents, files or images captured by a digital camera into editable and searchable data.

- Not 100% accurate
- OCR results have been improving – the accuracy can be over 99%
- Some information retrieval algorithms are very resistant to OCR errors

4.1 Image formats

- Image file size
  - number of pixels (rows * columns)
  - color depth (8-bit pixel (1 byte) stores 256 colors, a 24-bit pixel (3 bytes) stores 16 million colors - > true color)
- Image compression to fit in the standard - can be lossless or lossy:
  - **Lossless compression** algorithms reduce file size without losing image quality, When image quality is valued above file size, lossless algorithms are typically chosen.
  - **Lossy compression** algorithms take advantage of the inherent limitations of the human eye and discard invisible information. Most lossy compression algorithms allow for variable quality levels (compression) and as these levels are increased, file size is reduced.
- Various computer operating systems have their own preferred standard image representation.

4.1 Image of pages

4.1 Indexing images of pages

Question: How to index a set of document images?

- Manual approach
  - Write text description of each picture and index the text, index to the table of contents level
- Automatic approach
  - Use OCR
- Find an index made for other reasons

  e.g. Adonis CD-ROMs, IEEE journals in CD-ROM format

4.1 Image of pages

4.1 Shared Text/Image Systems

Page segmentation is often used for scanned pages which have a mixture of text, tables, equations, figures, and schemes.

CORE (Chemical Online Retrieval Experiment project) developed automated methods to distinguish the above mentioned five different components from the scanned pages

- deskewing pages
- equations and tables – in the database derived from keying
- figures – caption word “figure”
- regularity of lines (bit density plot) used to separate text from the illustrations

4.1 CORE- correlation function

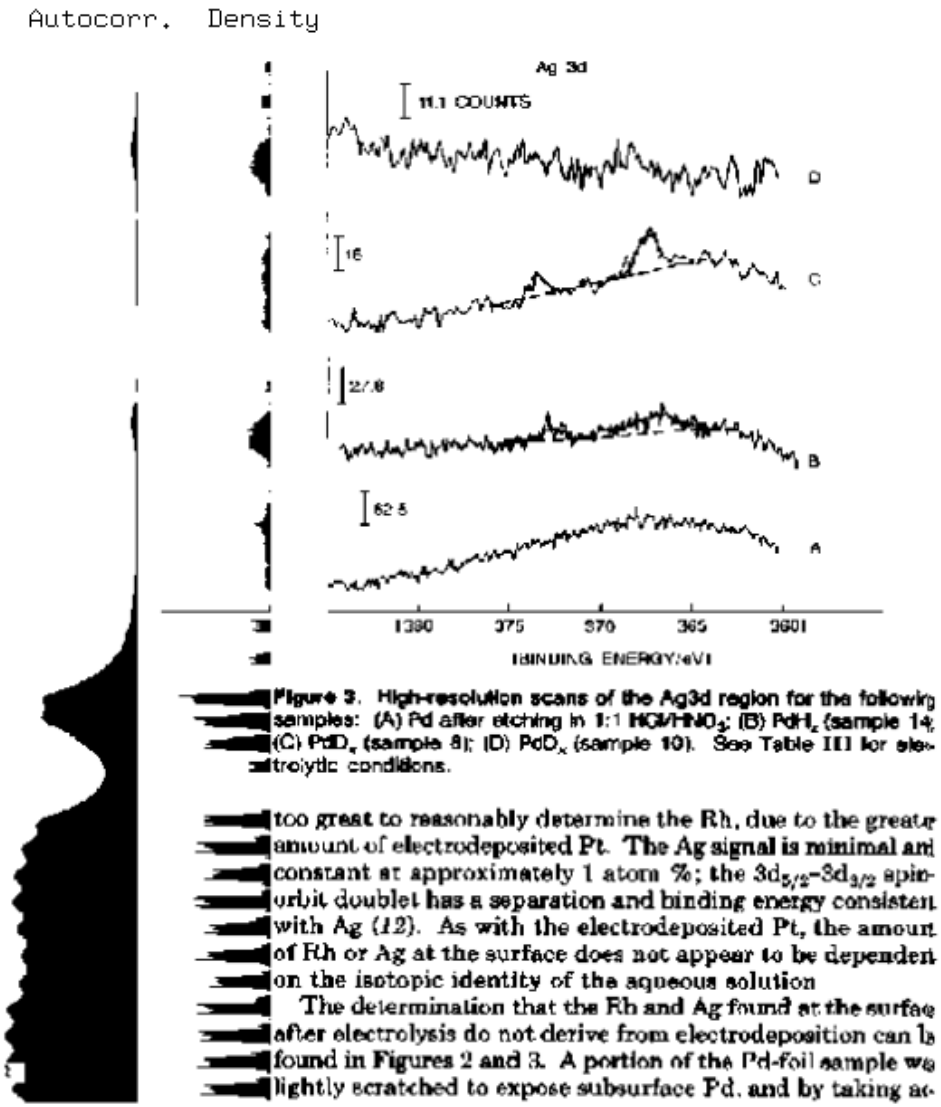


Figure 3. High-resolution scans of the Ag3d region for the following samples: (A) Pd after etching in 1:1 HCl/HNO_3; (B) PdH_x (sample 14; (C) PdD_x (sample 8); (D) PdD_x (sample 10). See Table III for electrolytic conditions.

too great to reasonably determine the Rh, due to the greater amount of electrodeposited Pt. The Ag signal is minimal and constant at approximately 1 atom %; the $3d_{5/2}$–$3d_{3/2}$ spin-orbit doublet has a separation and binding energy consistent with Ag (*12*). As with the electrodeposited Pt, the amount of Rh or Ag at the surface does not appear to be dependent on the isotopic identity of the aqueous solution

The determination that the Rh and Ag found at the surface after electrolysis do not derive from electrodeposition can be found in Figures 2 and 3. A portion of the Pd-foil sample was lightly scratched to expose subsurface Pd, and by taking ac-

4.1 Image of pages

4.1 Large scale digitization projects

- **Thesaurus Linguae Graecae project**

  Started 1970s, convert to machine readable the entire corpus of classical Geek literature, around 300 volumes, ancient Greek not included

- **Gallica collection** (Bibliothèque nationale de France)

  100,000 French books in image format

- **Questia, Netlibrary, and Ebrary** (commercial projects)

  100,000 books

- **Million Book Project**

  cooperation between Indian universities and Chinese Academy of Science, 1 million books free to read on the internet

4.1 Million Book project

Issues:

Supply of books: people are afraid to lend (with reason)

Copyright law: most recent material not available

Access to results: poor bandwidth to India and China

Coordination: not really adequate even within one country

Cataloging: need to train in OCLC

Progress and cost:

On average, 2000 books were scanned per day

On average, $3 per book

What we do not yet know : the effect that such a library will have

Learning Goals

- Understand page images as one important representation of documents
- Overview of important quality parameters of page images
- Basic understanding of image indexing methods (Recognition of text and image/figure/table elements)