The course so far…

- Reminder: All course materials are online: http://www-inst.eecs.berkeley.edu/~cs294-9/
- Overview of the DIA Research Field
- Some applications (Postal Addresses, Checks):
  - Ad hoc engineering
  - Complex / fragile / no effective models
- Research Objectives: more systematic modeling, design

DIA relies on several prerequisite engineering feats

- Converting paper media (physical) to electronic data (digital)
- Storage and retrieval of large quantities of digital images
- Agreed upon standards for representation of recognized results

A Potpourri of Topics

- Scanners
- Storage Formats for images
- Storage Formats for results
**Image Capture Devices**

- Film Cameras, then scanning?
- Direct Digital Cameras (still, video)
- Flatbed scanners (CCD)
- Drum Scanners (photomultiplier)
- Overhead Scanners
- Handheld Scanners (pen, array)
- Accessories (sheet feeders, networks, disks)

**Film cameras**

- Conventional lens/film optimizes for
  - Color rendition
  - Speed latitude
  - Storage before/after exposure
  - Cost
  - Sharpness (not same as resolution)
- Specialized cameras/film are used for making printing plates but...

**Film to Digital**

- Negatives or slides can be scanned
  - Kodak Photo CD
  - After-processing SOHO (e.g. Nikon Coolscan)
  - Professional (usually drum scanner)
- Expensive, slow, tedious, offline
- Very high quality drum scan possible

**Digital Cameras**

- Expensive CCD (needs 2-D sensor)
- Optics optimized for distance
- Color
- On-line memory and batteries dominate costs
Flat-bed scanners

- Prices from $50 / $300 / $3000
- Sometimes bogus comparisons
- Resolution from 200dpi to 2400dpi or more
- ‘Interpolated’
- (Bits depth per pixel, 1, 8, 24, 30, 32, 36, 48)
- Dynamic Range (2 – 3.8)
- Speed, feeder capacity, etc
- Transfer rate
- Accuracy of color
- Bundled software (photoshop lite, OCR...)

Flat-bed scanners

- Mostly standard construction
  - Array of ccds/light moves down paper
  - Optics, light stability, mechanics, interfaces vary
- Compare to hand-held: alignment speed
  (How does Capshare work, anyway!)

Observations re: resolution...

- FAX is hard (100x200dpi)
- Many optimized for about 300x300dpi
- Higher res. (600x600) increase costs; may improve results
- 1200x1200 seems to be overkill

OCR requirements: bit depth?

- Bit-depth 1 (for text), but who decides if gray=white or gray=black?
- Improved adaptive thresholding can be a selling point for a scanner
- Reading gray-scale (a burden for storage and software) may help
  - HP Capshare allows 1bit or 4bit b&w
  - Mixed text & photos benefit
What does the scanner see?

The scanner apertures

Sampling frequency vs pattern; see readings for Fourier sampling

How much resolution to find an edge?

Do you exactly care?

How about gray scale?

Not so different, if we threshold at 0.50

Why threshold at 50%?

- We made that up. How do we find an appropriate parameter?
- (tangent: Choosing the right values for some of hundreds of parameters can significantly affect performance of commercial OCR. Far too many mysteries)
Global Thresholding by Histogram

- # of pixels
- Amount of ink/pixel

Other global measures

- 1st or 2nd derivative of histogram

Varying threshold on a gray-level image

Adaptive thresholding

The black printing on line 1 is lighter than the background on line 2
Pretty good thresholding algorithms can often be done in hardware in parallel:

- Speed
- Improved image quality at the source (less noise to transmit, etc.)
- Plausibly modeled mathematically
- Maybe other heuristic processing tossed in as well: toss out black scanning margins (scanning small papers or photos)

Image storage

- Too many file formats
- UNIX `convert` utility mentions these...

---

<table>
<thead>
<tr>
<th>File Format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMP</td>
<td>Microsoft Windows bitmap image file.</td>
</tr>
<tr>
<td>CMYK</td>
<td>Raw cyan, magenta, yellow, and black bytes.</td>
</tr>
<tr>
<td>DCX</td>
<td>Zsoft IBM PC multi-page Paintbrush file.</td>
</tr>
<tr>
<td>DIB</td>
<td>Microsoft Windows bitmap image file.</td>
</tr>
<tr>
<td>EPSS</td>
<td>Adobe Encapsulated PostScript Interchange format.</td>
</tr>
<tr>
<td>FAX</td>
<td>FAX (Group 3).</td>
</tr>
<tr>
<td>FITS</td>
<td>Flexible Image Transport System.</td>
</tr>
<tr>
<td>GIF</td>
<td>CompuServe Graphics image file.</td>
</tr>
<tr>
<td>GIF87</td>
<td>CompuServe Graphics image file (version 87a).</td>
</tr>
<tr>
<td>GRAY</td>
<td>Raw gray bytes.</td>
</tr>
<tr>
<td>HDF</td>
<td>Hierarchical Data Format.</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language.</td>
</tr>
<tr>
<td>HISTOGRAM</td>
<td>Histogram.</td>
</tr>
<tr>
<td>JBIG</td>
<td>Joint Bi-level Image Experts Group file interchange format.</td>
</tr>
<tr>
<td>JPEG</td>
<td>Joint Photographic Experts Group file interchange format.</td>
</tr>
<tr>
<td>MAP</td>
<td>Red, green, and blue colormap bytes followed by the image colormap indexes.</td>
</tr>
<tr>
<td>MATTE</td>
<td>Raw matte bytes.</td>
</tr>
<tr>
<td>MIF</td>
<td>Magic image file format.</td>
</tr>
<tr>
<td>PNG</td>
<td>Portable Network Graphics.</td>
</tr>
<tr>
<td>PICT</td>
<td>Apple Macintosh QuickDraw/PICT file.</td>
</tr>
<tr>
<td>PDF</td>
<td>Portable Document Format.</td>
</tr>
<tr>
<td>PICT</td>
<td>Apple Macintosh QuickDraw/PICT file.</td>
</tr>
<tr>
<td>PNM</td>
<td>Portable bitmap.</td>
</tr>
<tr>
<td>PS</td>
<td>Adobe PostScript file.</td>
</tr>
<tr>
<td>RAD</td>
<td>Radiance image format.</td>
</tr>
<tr>
<td>RGBA</td>
<td>Raw red, green, blue, and blue bytes.</td>
</tr>
<tr>
<td>RLA</td>
<td>Alias/Wavefront image file; read only.</td>
</tr>
<tr>
<td>RLE</td>
<td>Utah Run-length encoded image file; read only.</td>
</tr>
<tr>
<td>SUN</td>
<td>SUN Rasterfile.</td>
</tr>
<tr>
<td>TEXT</td>
<td>raw text file; read only.</td>
</tr>
<tr>
<td>TGA</td>
<td>Travevision Targa image file.</td>
</tr>
</tbody>
</table>
TIFF  Tagged Image File Format.
TILE tile image with a texture.
UYVY 16bit/pixel interleaved YUV (e.g. used by AccomWSD).
VICAR read only.
VIFF  Khoros Visualization image file.
XBM X11 bitmap file.
XPM X11 pixmap file.
XWD  X Window System window dump image file.
YUV  CCIR 601 4:2:2 file.

How to choose a format?

- Storage cost per pixel (disk space, transmission)
- Encode/decode cost of compression
  - Offline, online, 1-D, 2-D, 3-D (time)
- Versatility, extensibility
- Robustness (error sensitivity)
- Incremental (page at a time..)

How to choose a format?

- Programming ease
- Machine independence, standardization
- Vendor support
- Popularity
- Proprietary (+ or -)

Why TIFF

- Tagged image file format
- Tags can be added: standard grows
  - Old programs may not work with new tags
  - New programs should work with old tags
- Raster based / matches scanner output
- Wrapper around other encodings, compressions (LZW, CCITfax3,4, JPEG ..)
- Multiple images per file
- FREE LIBRARIES FOR UNIX, WINDOWS...
Extras in TIFF

- Lots of features we don’t use
  - Color spaces (RGB, pseudocolor, CMYK, CIELab…)
  - Arbitrary bits/pixel (we use 1 !)
- Developed by Aldus & Microsoft; owned by Adobe
- See the unofficial TIFF home page

Restrictions on TIFF

- No native provisions for storing vector graphics, text annotations
- File based: offsets for headers.
- Limit of 4 gigabytes of (compressed) data
- Some programs don’t implement it right
  - E.g. assume byte order
- Extensions: “XIFF”

Compression: Can we do better?

- Yes: 2-D image coding / JBIG
  - DigiPaper/ Huttenlocher/Xerox
  - CPC explanation is pretty good.
  - DJVu (ATT/Lizardtech) http://www.djvu.com/
  - Adobe Capture
- More work to compress, decompress
- Claimed factors of 5:1 over CCITTfax4

How much randomness is there in a (compressed) doc?

- Look for 2-d patterns (AKA characters or even words
- Computed on-line in a stream or batch
- Separate out background colors/textures
- Allow for some loss (how much, a parameter)
- Deal with small differences cheaply
Compression ratios CCIT test
page 1.  188:10:8:4:1

6 page document, compressed, shown as bits

aside: JSTOR application
- On-line journals OCR + images
- Needs special (but free) viewer
- CPC compression engine not free
- OCR is not visible except in abstracts
- (getting the OCR right is done via hand correction)
- http://www.jstor.org/

Other advantages
- Faster download and rendering
- Viewing can begin before the whole file is downloaded
- Browser plug-ins available
NEC CiteSeer

- ResearchIndex provides autonomous citation for PS and PDF research articles on the WWW
- Citations are cross-linked
- Full-text indexing
- Page images provided
- Source code available for non-commercial use

Berkeley’s digital library project

- Multivalent documents: new document model: extensible, distributed
- OCR + image + ...

So why do we also use PDF?

- Common viewing/printing interface
- Supported by WWW browsers
- Alternative for HP Capshare
- Supported in printer hardware

What does PDF look like...

```
%PDF-1.1
<< /Type /Catalog /Pages 2 0 R >>
endobj
3 0 obj
<< /Type /Page /Parent 2 0 R /MediaBox [ 0 0 162 323 ] /Contents 4 0 R /Resources << /ProcSet [ /PDF /ImageB /ImageC /ImageI ] /XObject << /Im005 5 0 R >> >> >>
endobj
4 0 obj
<< /Length 29 >> stream
162 0 0 323 0 0 cm
/Im005 Do
endstream
endobj
5 0 obj
<< /Type /XObject /Subtype /Image /Name /Im005 /Filter /CCITTFaxDecode /DecodeParms << /K -1 /Columns 672 >> /Width 672 /Height 1344 /BitsPerComponent 1 /ColorSpace /DeviceGray /Length 6 0 R >> stream
```
What does TIFF look like? (OD)

A BIG JUMP to the end of the task

How do we represent answers?

• Ideal: Whatever signal produced the image on the paper (absent any noise)
• Plausible: Enough of a signal to produce the same image on the paper, but with more semantic content than a bit map
• Reality: An approximation that could (perhaps after some editing) be use for some well-defined purpose

Will ASCII do it all?

• Hardly. The discussion of UNICODE shows there is sometimes a very indirect connection between glyphs and characters.
• E.g. fi vs ğ
• The different glyphs for the same character depending on context (mid- versus beginning of word)
Will UNICODE do it

- Character encoding (up to 32 bits): sounds like plenty!
- Yet, does not describe attributes like point size, bold/italic/compressed …
- Does not describe FONT (like Arial, this font, or Times Roman, this.)
- Does not describe structures or semantics such as “author” or “title”

No. not even for text, but it is a start

- What about math?
  - Syntactic Math {various…}
  - Semantic Math {???
- Other “media” …e.g.
- What about printed music?

Music Recognition

- Idea: scan scores & do stuff
- Convert to MIDI (to play)
- Convert to NIFF (notation interchange file format appropriate for composition/correction programs etc.)
- Possible paradigm for other special areas.

Examples (Musitek)

- Scanned image
- MIDI (?)
- NIFF/LIME
Semantic interpretations

We will return to this issue

- If the world becomes web centric, maybe the solution will be found in that direction.
- What does it REALLY mean to read and represent a text… If we understand an image of text, does that mean we can generate a translation to another language “transpose to the key of German”?