

State of Art of Document Image Processing

George Nagy
Rensselaer Polytechnic Institute

Not as a stranger ...

Reader, PhD theses:

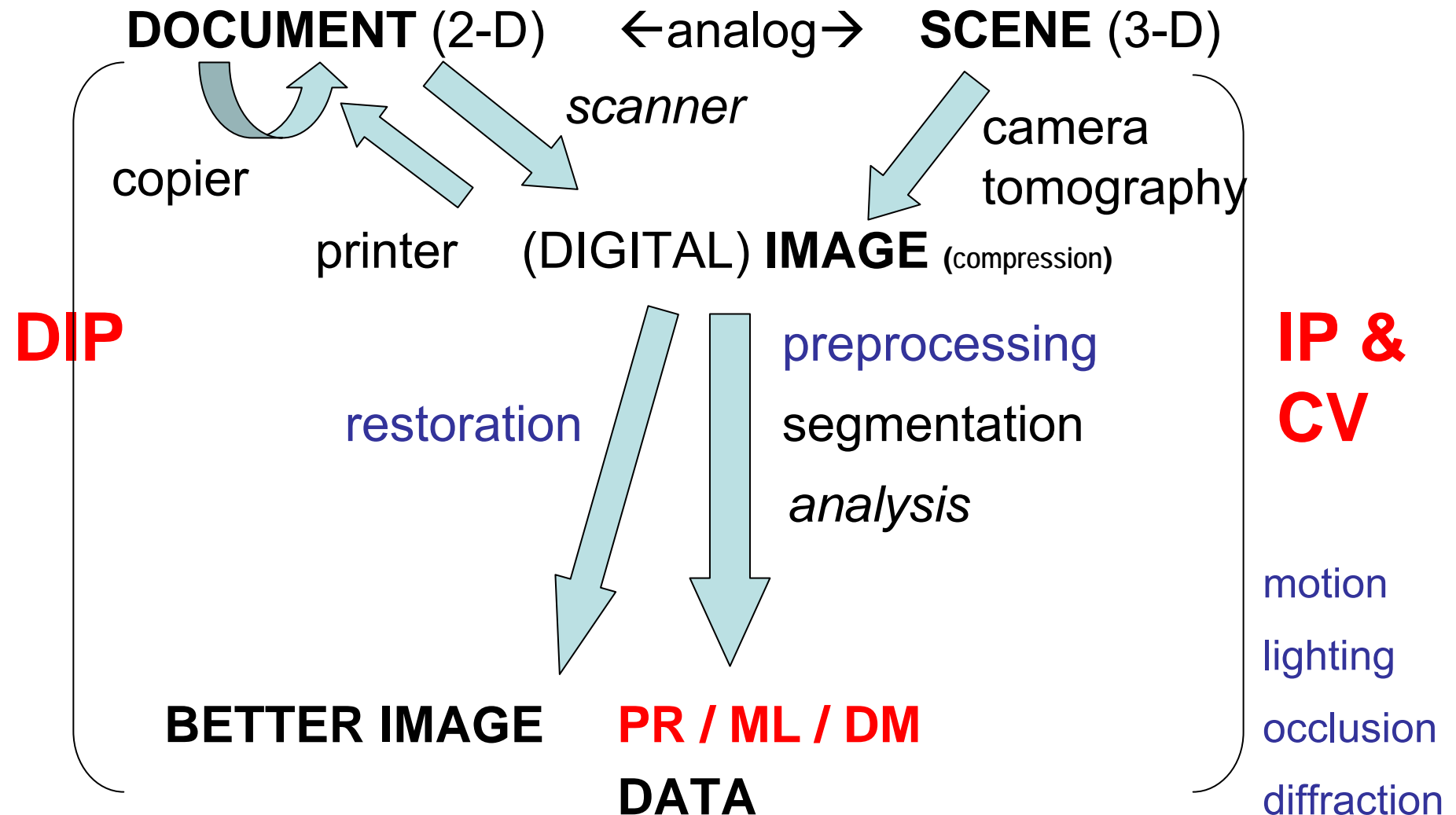
Nagaraja, G.	IISc	1975
Bansal, V.	IIT Kanpur	1997
Pal, U.	ISI	1997
Murali, S.	Mysore	2002

India-US NSF Research Collaboration 1989

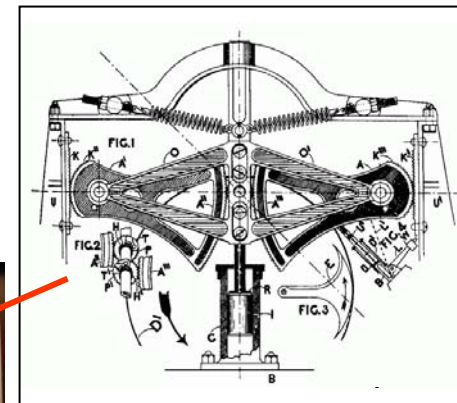
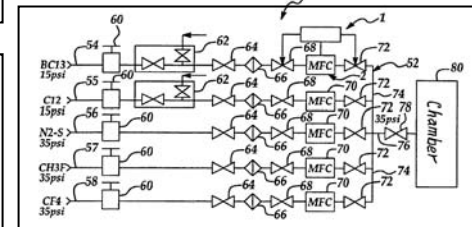
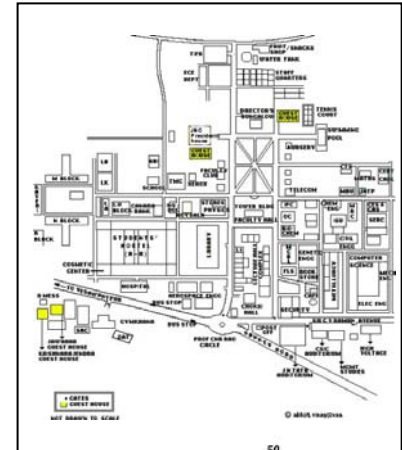
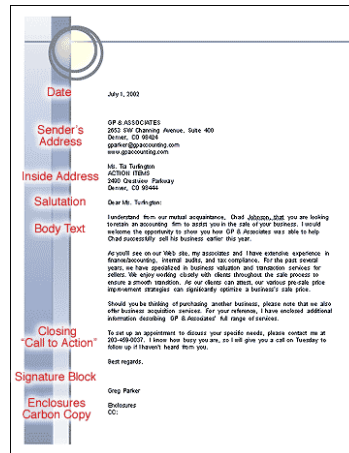
ICDAR Bangalore, GREC Jaipur 1999

PhD students at UNL & RPI: Wagle*, Mehta*, Viswanathan, Maulik*,
Mukherjee, Narendra*, Sarkar, Veeramachaneni, Joshi, Andra,
+ *many* MS students, supervisory committees, friends and colleagues

DIP in context



Documents: archives, newspapers, magazines, books, letters, engineering drawings, diagrams, maps, sheet music,



5/15/2008



SSDIP, GN, Bangalore



Goal of DIA depends on document type

Document type

DIA Target

plain text

correct word order for OCR

illustrated text

reading order, links to illustrations

structured text

compilable or executable form

 envelope, letter

 routing information

 directory, TOC

 name-attribute pairs

 business form

 links to database, add tags

schematic diagram

net list or graph

engineering drawing

current CAD format

map

GIS representation

music score

MIDI representation

table

layout-independent descriptor

Caveat Emptor:

DIA now largely retrospective

- Most documents we wish to keep now produced digitally: digital books, journals, newspapers, letters, drawings, forms (like tax returns and Indian visa applications). But ...
- Digital version is not always available: personal DIA.
- Original software or digital medium is not always available: conversion of CAD drawings, tech journals, census data.
- Many pre-1980 documents remain to be converted, some of actual value (utility drawings, cadastres), and many historical artifacts.
- In the US, there is continuing interest in security applications.

OUTLINE

- The evolution of documents
- Advances in document image capture
- Document image analysis
- Challenges
- To read further

Conclusions

Ancient manuscripts still require manual keying.

Automated OCR of **original plain text documents** in common scripts usable for most purposes.

Conversion of **illustrated text, tables, graphs, line drawings, maps** requires interaction.

Form recognition helped by context from database, but requires occasional intervention and confirmation. Better tools are needed.

Ditto for **line drawings and maps**.

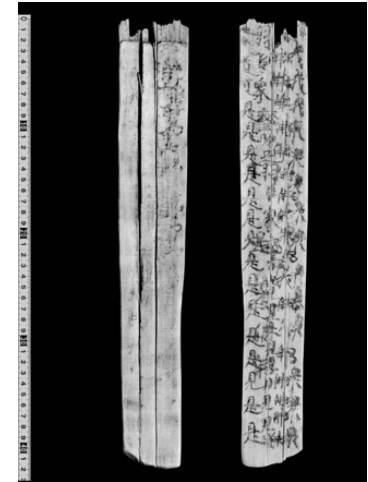
We are about to witness the convergence of **digital libraries** and the **semantic web**.

OUTLINE

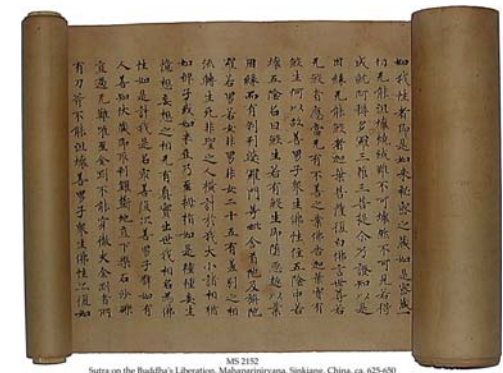
- The evolution of documents
- Advances in document image capture
- Document image analysis
- Challenges
- To read further

Before the printing press

Stone, papyri, silk, reeds, wood, paper, ...



www.city.niigata.jp/.../about/history/mokkan.jp



<http://www.pbs.org/wgbh/nova/archimedes/images/manu-herculaneum.jpg>

<http://www.trin.cam.ac.uk/show.php?imgid=215>

5/15/2008

SSDIP, GN, Bangalore

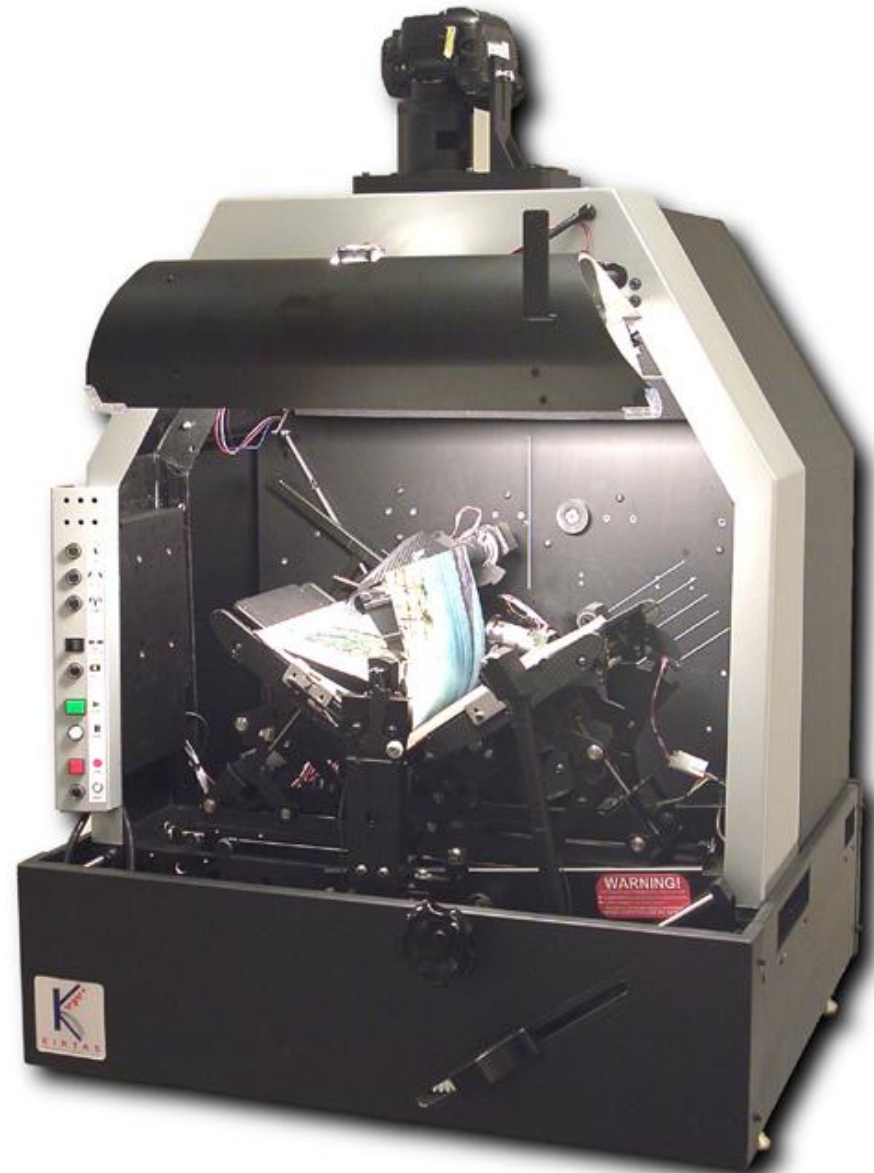
10

Preservation

In the last decade,
tremendous *global*
interest.

Digitization for wide access
(& sequestration of
originals!)

Requires non-contact
imaging, and *fast*,
interactive image
enhancement, annotation,
and **indexing.**



Large-scale conversion projects underway in many countries

Sanskrit Documents

॥ श्रीः ॥

न चोरहार्यं न च राजहार्यं न भ्रातृभाज्यं न च भारकारि ।
व्यये कृते वर्धते एव नित्यं विद्याधनं सर्वधनप्रधानम् ॥

[CLICK FOR MEANING]

ॐ namo namah . . Salutations . . ॐ

Welcome to the compilation of [Sanskrit Documents](#) in Devanagari display and transliteration format. In addition to the sanskrit texts, you will find here various tools for learning Sanskrit such as the [Online Sanskrit Dictionary](#), [Sanskrit Tutorials](#), [Sanskrit Pronunciation guides](#), and software for [learning Sanskrit and producing documents in Devanagari & Roman formats](#), and much more. You can generate display in Devanagari or other scripts using the [web-interface](#).

We encourage you to [participate](#) in these efforts by reviewing, encoding, spreading the word, and implementing available resources to make Sanskrit learning easier. Answers to frequently asked questions are given in [FAQ](#). A live Sanskrit links page is maintained at [sanskritlinks](#). We need volunteers to proofread the texts available in pending sections [1](#), [2](#), and [3](#).

Join the email-based lively discussions in the [Sanskrit Mailing List](#) or Sanskrit digest for discussions on the Sanskrit language and literature. Mails are archived, but you have to subscribe to browse them.

For scholarly promotion of Sanskrit literature studies, we have also compiled a long list of [Sanskrit documents available elsewhere](#), [hundreds of scanned books](#), and [audio files](#) on the internet.

Visit the non-profit [sites](#) such as Complete Narayaneeyam, Surasa.net, and SETU hosted on this site.

Please fill in our [GuestBook/Feedback](#) and browse through the Site using the [Site Contents Map](#) or the navigation bars on top and bottom.

15th-20th C: Printed Documents

Documents contain symbols.

They require different techniques from natural pictures (photos).

Documents, intended for ease of human reading, have
high contrast (are **bilevel** or decomposable into bilevel layers);
isothetic (**rectilinear**) layout;
limited symbol vocabulary;
significant language and application **context**;
isogeny (common source) wrt author, printer, copier, scanner.

Most DIP applications are batch oriented:
they require fast, repetitive processing of similar images.

Large-scale DIA applications

- Post ↓ (email, courier, barcode, RFID)
- Forms ↓ (direct data entry through web forms)
- Books ↑ (© ?)
- Engineering Drawings ↓ (mainly utilities)
- Maps ↓ (satellite remote sensing, GPS)
- Technical journals ↓ (nearly done)
- Historical documents ↑ (small budget)

21st C: Electronic Documents

- **PDF** files (based on compressed Postscript)
(often easier to analyze after rendering!)
- **HTML** files
tags often used in arbitrary ways: e.g., <table> for layout
- **XML** tags require Document Type Definition
- **Nested** pages are troublesome (Document Object Model)
- **Dynamic** documents!

Syntax → Semantics

- Domain-specific ontologies for the semantic web
Dublin Core Metadata
Resource Description Framework
OWL Web Ontology Language

OUTLINE

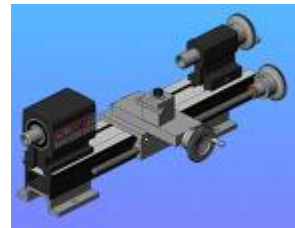
- The evolution of documents
- Advances in document image capture
- Document image analysis
- Challenges
- To read further

Information lost during
capture cannot be
recovered!



Indian scanner pioneer

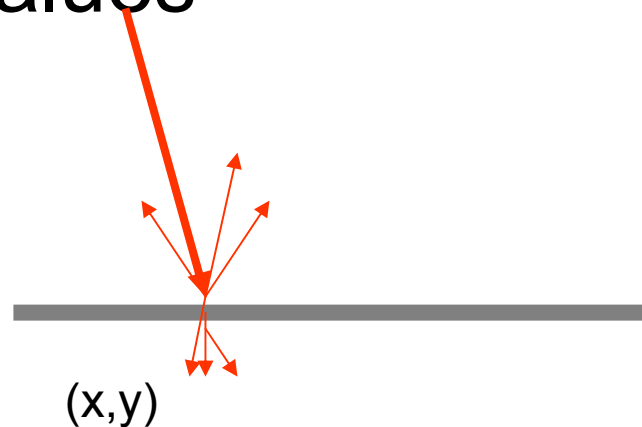
Dr. Deekshatulu was a Visiting Scientist at the **IBM Watson Research Centre**, York Town Heights, New York, and at the **Environment Research Institute of Michigan** during 1971-72 on Digital Image Processing and Remote Sensing. He designed and fabricated for the first time in India, **Grey scale and color Drum Scanners** for computer picture processing which won him an NRDC Award.



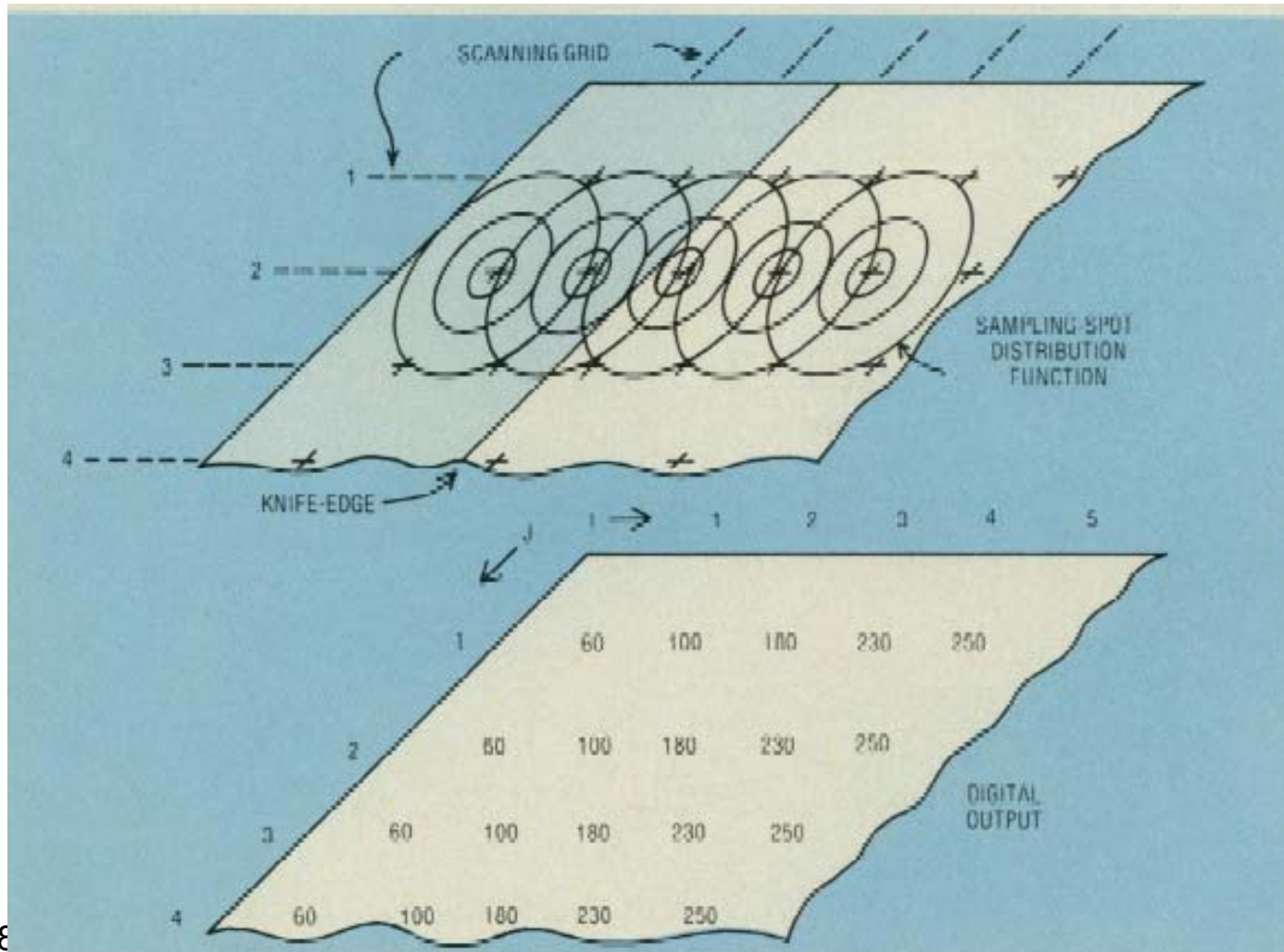


SCANNERS and DIGITAL CAMERAS

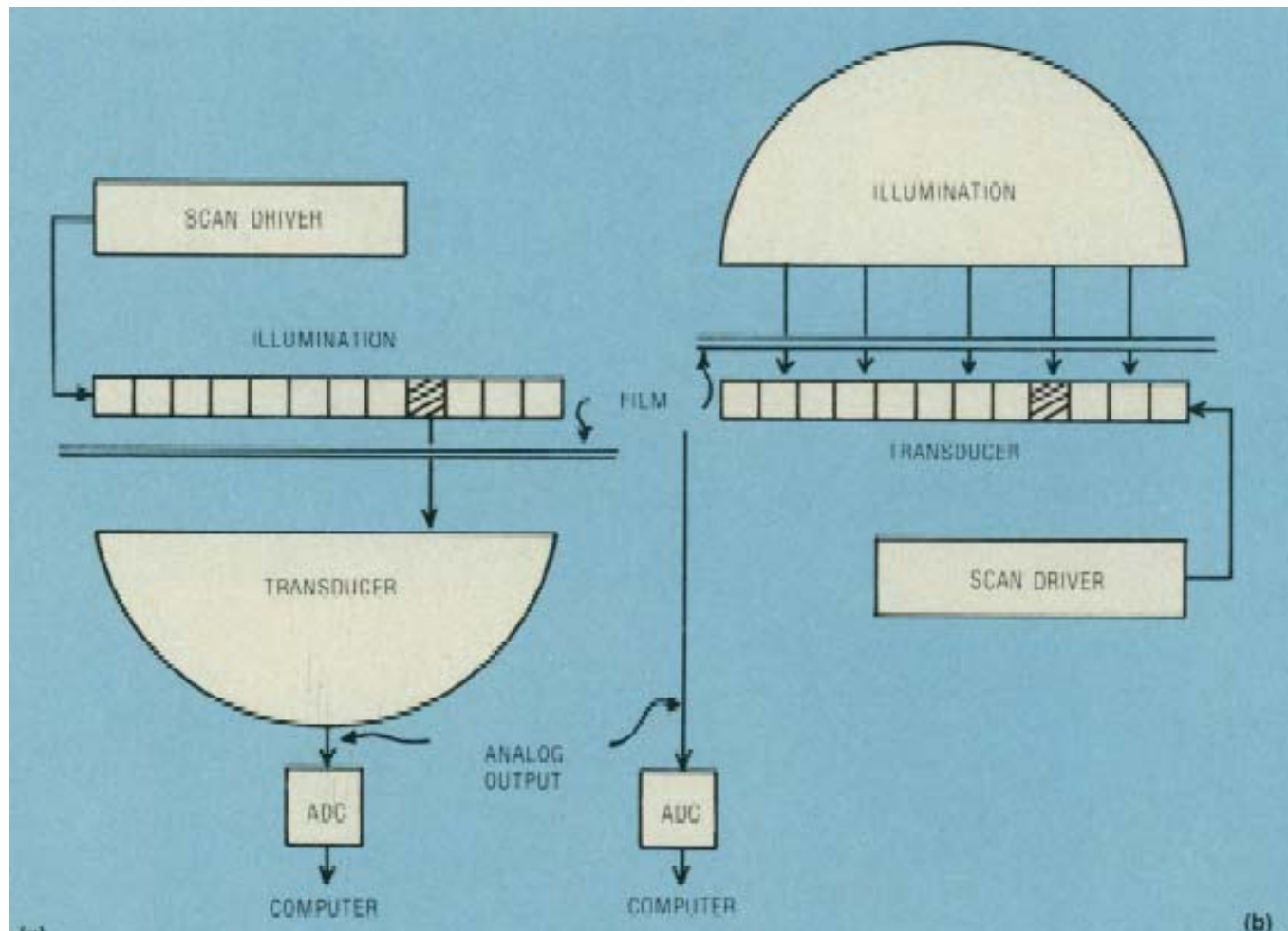
Convert reflectance (or transparency)
to pixel values



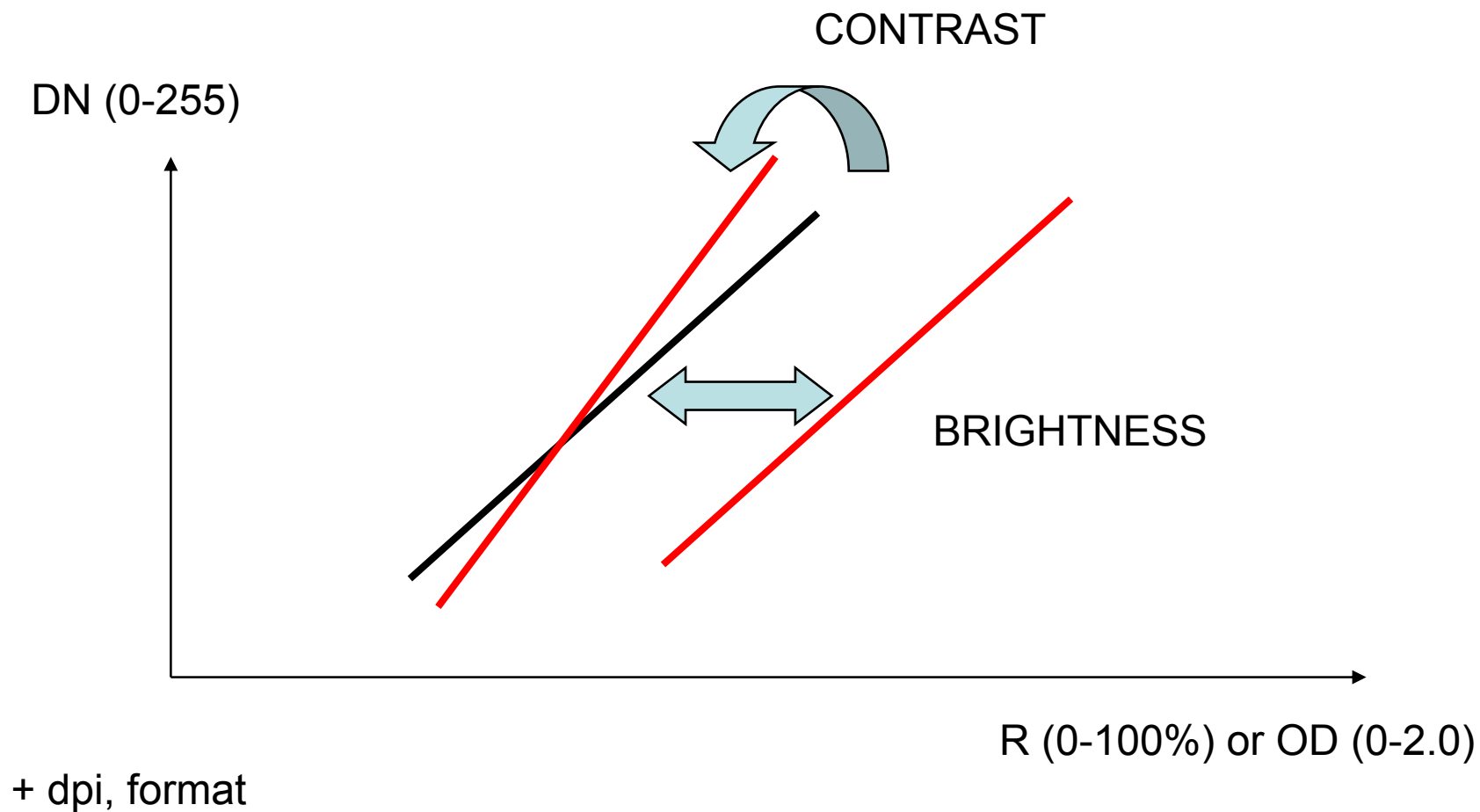
Point Spread Function (PSF)



Flying spot / flying aperture



Photometric scanner controls



Important Scanner parameters

- **Point spread function** (PSF) diameter and shape
- Spatial sampling rate (nominal / physical)
- Photometric transfer function (reflection density to gray levels)
- **Photometric uniformity** across the page
- Horizontal vs. vertical, and left vs. right uniformity
- **Geometric linearity**
- Color response (for maps, drop-out forms, magazines)
- Repeatability (temperature, aging light, scan start/stop)
- Cosmetic functions (crop, straighten)
- Document format support (including compression)
- Speed and data transfer rate (simplex/duplex)
- Digitally stored scan parameters
- **Support for calibration**

Calibration

Compensates for

Non-uniform **illumination**

(variation usually smooth)

Non-uniform **sensor sensitivity**

(row or array sensors)

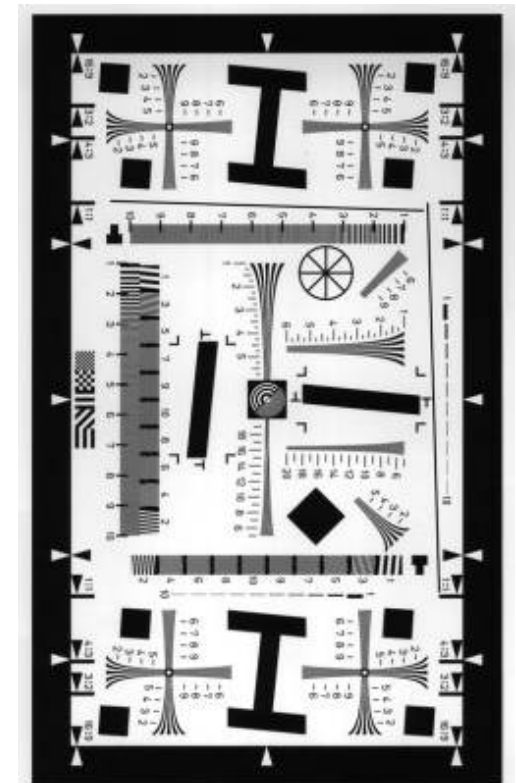
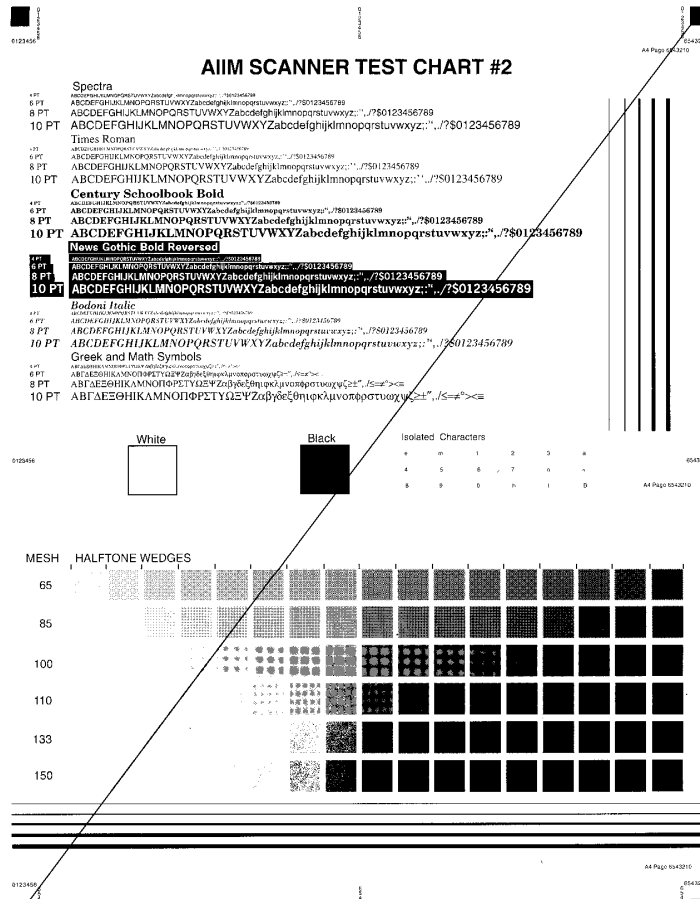
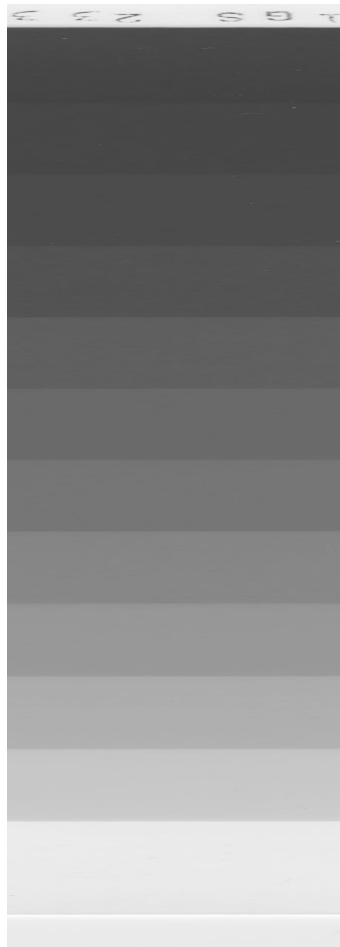
Geometric **distortion**

(including page transport if any)

Can be done within or outside the device

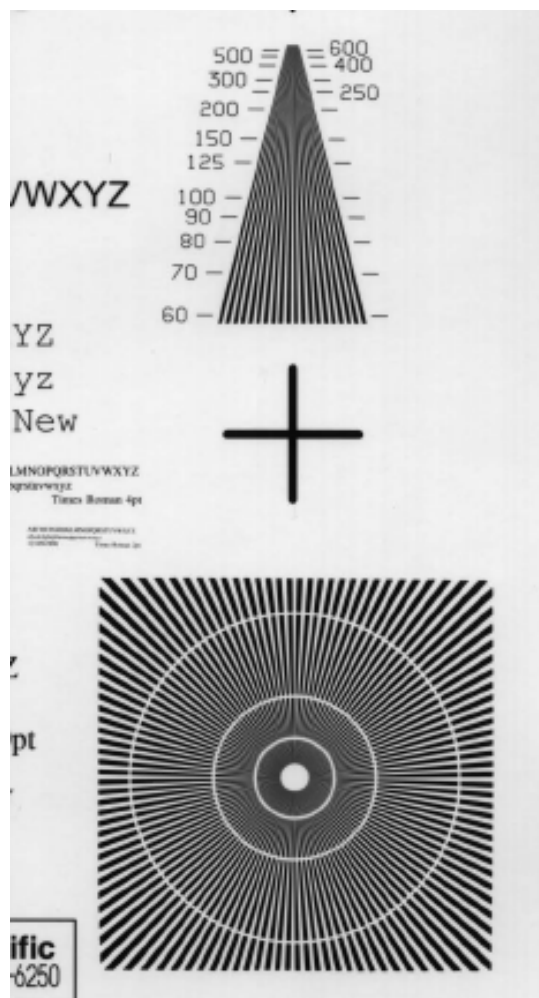
Test targets should be scanned with every batch

Test Charts



CALIBRATION CHART

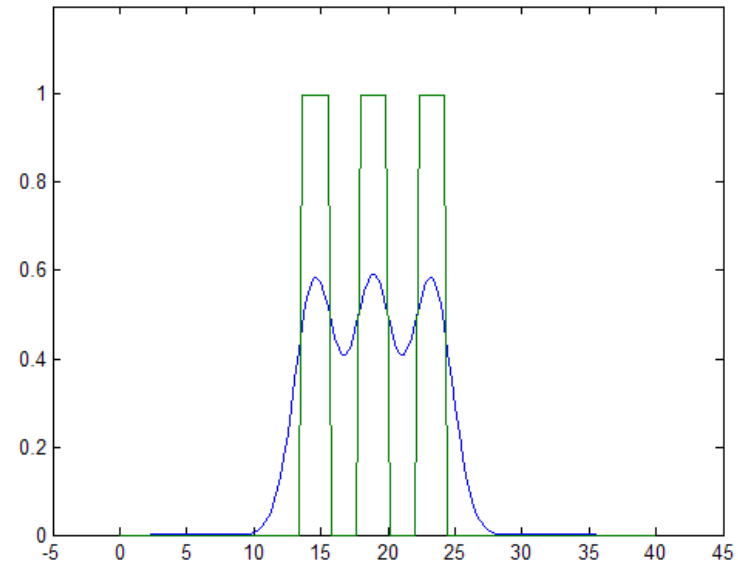
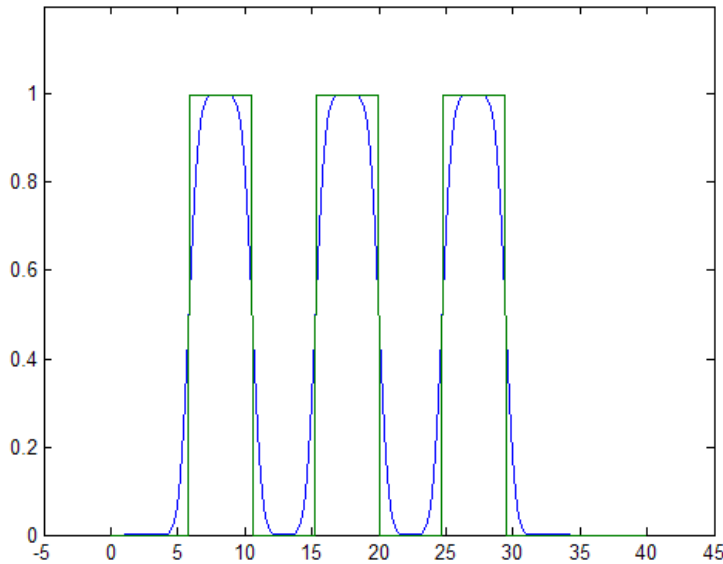
gray scale
spatial sampling
resolution
geometric linearity



SSDir, GN, Bangalore

Modulation Transfer Function (MTF)

(sometimes sinusoidal rather than bar patterns are used)



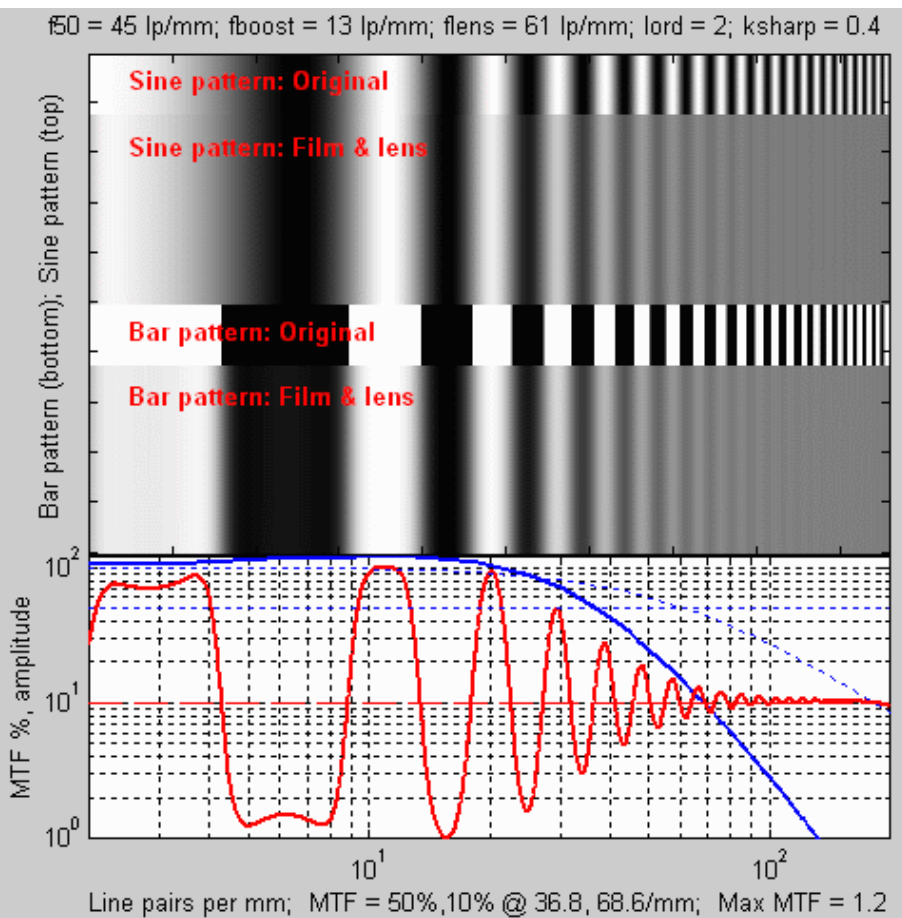
- **Modulation = (Max – Min) / (Max + Min)**
- MTF is the ratio of the modulation to the modulation at $f=0$ lpm or cpm as a function of spatial frequency.
- Horizontal and vertical MTF are measured separately.
- The limiting resolution is the lpm where the modulation is 5% of the maximum modulation.

OTF, MTF (=SRF), PSF

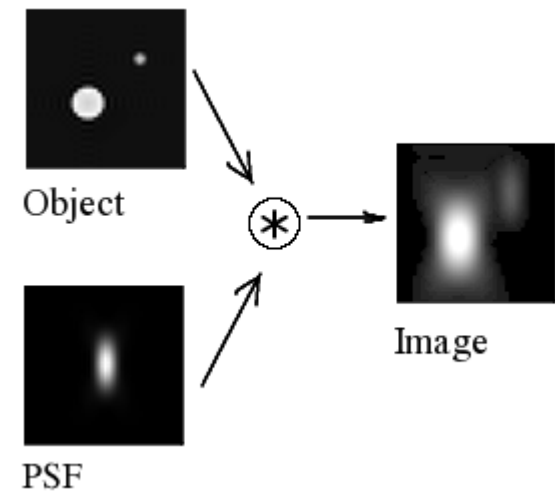
The optical transfer function is the Fourier Transform of the point spread function.

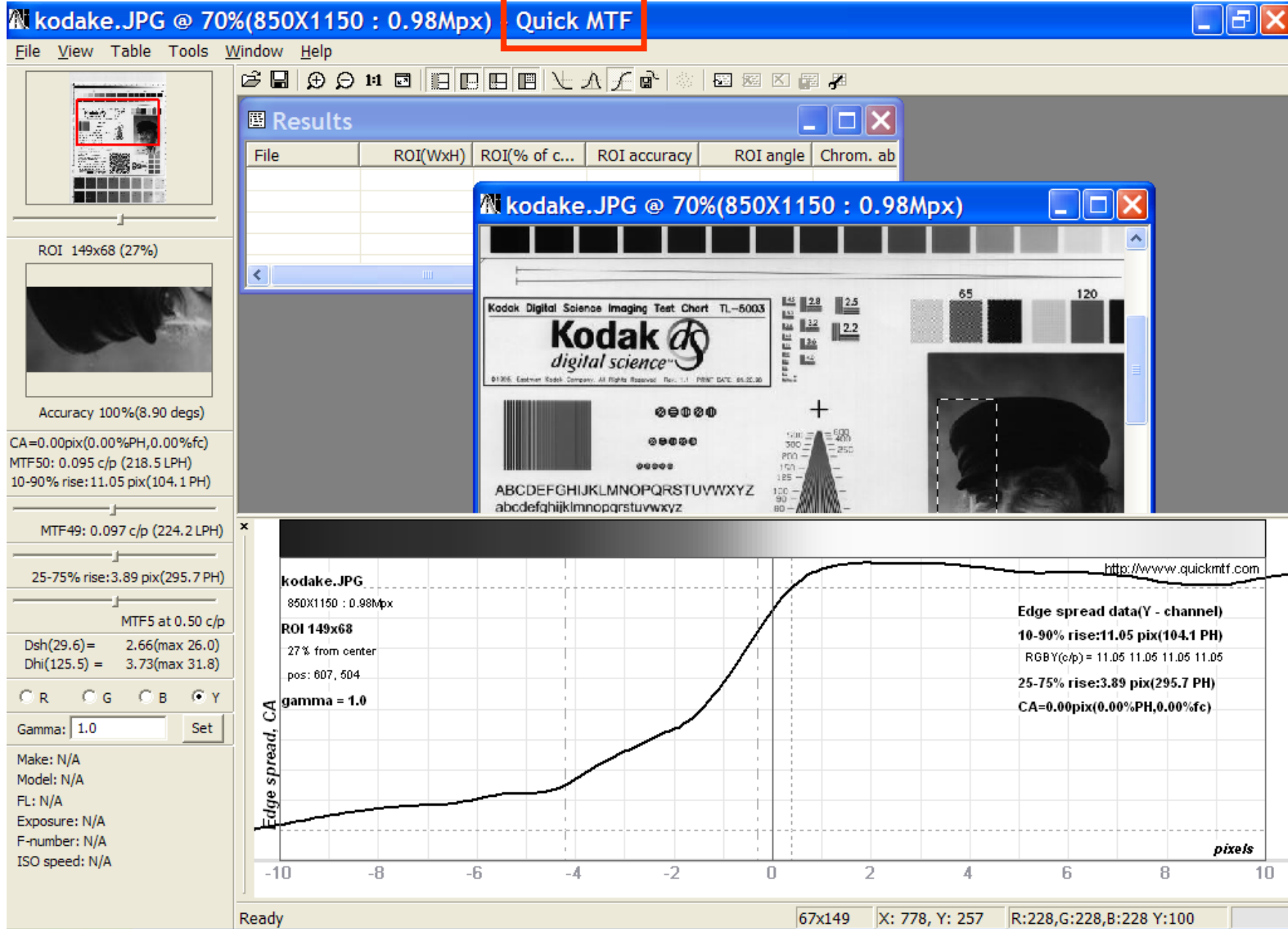
The real component (modulus) of the OTF is the MTF; its phase is the PTF.

$OTF(f_x, f_y) = MTF(f_x, f_y) * PTF(f_x, f_y)$,
where f_x, f_y are spatial angular frequencies.



The ideal MTF is the Fourier Transform of a delta function.





Acronymous Quantitative Scanner/Camera Characterization

OTF Optical Transfer Function

MTF Modulation Transfer Function

PTF Phase Transfer Function

SRF Spatial (frequency) Response Function

PSF Point Spread Function

ESF Edge Spread Function

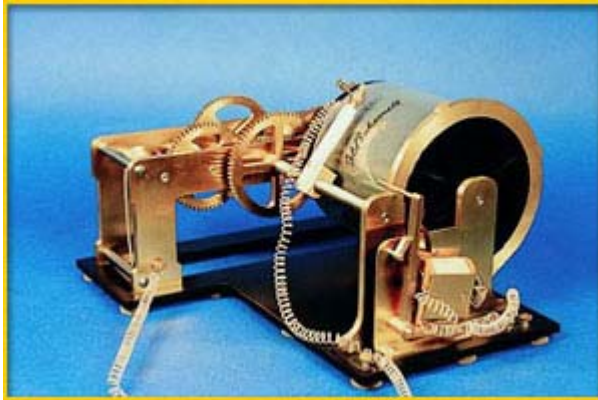
lpm lines per mm **cpm** cycles per mm **lppm** line pairs per mm

lph, pph lines per picture height, pixels per picture height

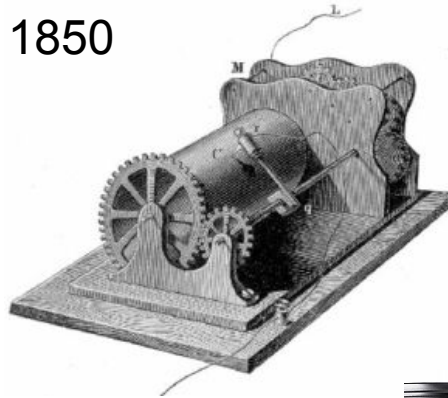
OD Optical Density, **DN** Digital Number, **GL** Gray Level

SLR Single Lens Reflex, **ISO** International Standards Organization

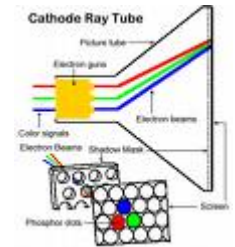
Progress in image digitization



1850



1960



1980



2008



The-Digital-Picture.com Reviews



2010?



5/15/2008

SSDIP, GN, Bangalore

Current status

- Normal reading material is digitized at 200-400 dpi.
- There is a spatial sampling rate / gray-scale trade-off, but most OCR software is still bilevel.
- Equipment must be matched to document quality, but scanner cost is now longer a major factor.
- Consumer cameras match CCD scanner psf, and almost match geometric linearity.
- Cell-phone type cameras will soon offer ubiquitous document digitization.
- Most applications require human oversight of digitization.

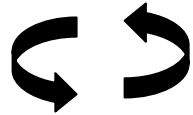
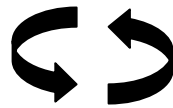
OUTLINE

- The evolution of documents
- Advances in document image capture
- Document image analysis
- Challenges
- To read further

Bottom-up DIA

- Preprocessing (pixels)
- Glyph segmentation (primitives)
- Text recognition (OCR) (structures)
- Page layout analysis (document)
- Indexing and IR (corpus)

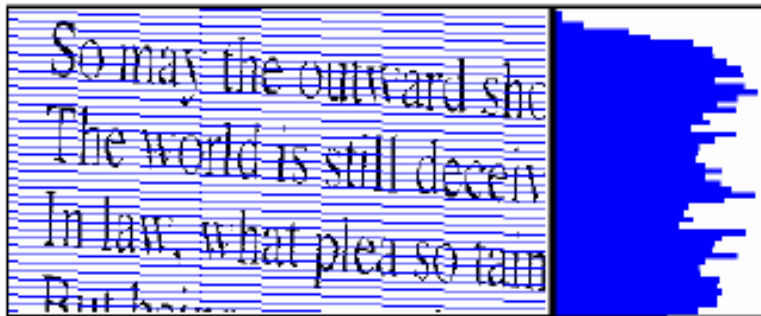
Preprocessing

- Recover scan parameters (dpi, psf, gamma, color)
- Recover batch application data
- Decompress (or analyze compressed representation)
- Filter noise (but keep periods and dots on the i and j!)
- Binarize (global / local / interactive) (??) 
- Detect and remove (??) skew
- Character segmentation / line-art vectorization 
- Script, orientation, language, font recognition

Preprocessing often removes useful information.

Devise distortion-invariant analysis procedures instead!

Skew detection & Character Segmentation



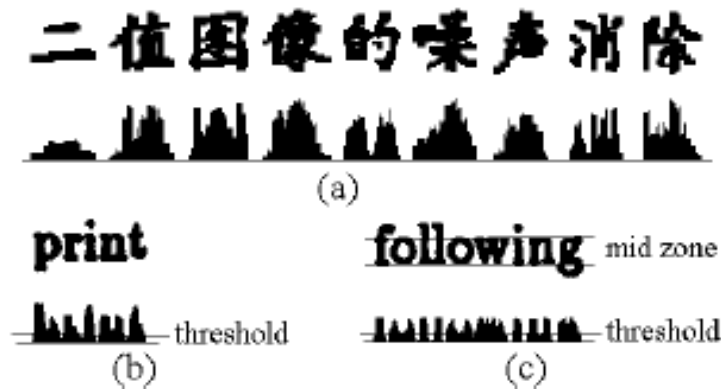
So may th
The world
In law, wh

So may th
The world
In law, wh

from Cheriet et al

Glyph segmentation

- Projections
- Connected component analysis
- Morphological operators (open, close)
- Templates



“Physical” layout analysis

- Assemble or subdivide rectangular regions
- X-Y trees (successive H & V cuts)
- DID (side-bearing model)
- Columns, paragraph blocks, illustration blocks, word blocks, character blocks (?)

Evaluation difficult without downstream processing!

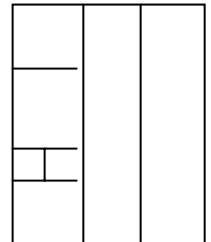
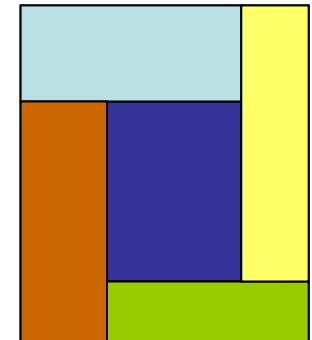
Some examples of generic typesetting knowledge for text set in derivatives of the Latin alphabet:

- Printed lines are parallel and roughly horizontal.
- The baselines of characters are aligned.
- Each line of text is set in a single point-size.
- Ascenders, descenders, and capitals have consistent heights.
- Serifs are aligned.
- Typefaces (including variants italic or bold) don't change within words.
- Within a line of text, word spaces are larger than character spaces.
- The baselines of text in a paragraph are spaced uniformly.
- Each paragraph is left-justified or right-justified (or both), with special provisions for the first and last line of a paragraph.
- Paragraphs are separated by wider spaces than lines within a paragraph, or by indentation.
- Illustrations are confined to rectangular frames.
- In multi-column formats, the columns are of the same width.

X-Y TREE

Advanced Character Recognition 6610
(invited)
George Nagy
Rensselaer Polytechnic Institute, Troy, NY, USA

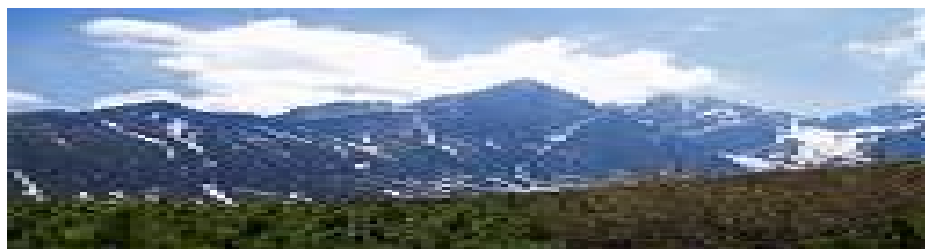
CATALOG DESCRIPTION	SYLLABUS
<p>ECSE 6610 - Advanced Character Recognition. <i>Principles and practice of the recognition of isolated or connected typeset, hand-printed, and cursive characters. Review of optical digitization, supervised and unsupervised estimation of classifier parameters, bias and variance, expectation maximization, the curse of dimensionality. Advanced classification techniques including classifier combinations, support vector machines, hidden Markov methods, styles, language context, adaptation, segmentation-free classifiers, indirect symbolic correlation. Prereq: ECSE 2610, Probability, Linear Algebra. Spring term annually.</i></p> <p>ECSE-6610 FIRST DAY HANDOUT</p> <p>Instructor: Prof. George Nagy Office hours: After class in the bar Email: nagy@ecse.rpi.edu</p> <p>Text: S. V. Rice, G. Nagy, T. A. Nartker Optical Character Recognition: An Illustrated Guide to the Frontier [RNN 99]</p> <p>Reference texts (on reserve at Folsom Libe): Duda, Hart, & Stork, Wiley 2001 [DHS 01] Mitchell, McGraw-Hill 1997 [MT 97] Nadler & Smith, Wiley 1993 [NS 93] Schürmann, Wiley 1996 [SJ 96] Theodoridis & Koutroumbas, Acad.1999 [TK 99]</p> <p>For additional sources, see the Text and the Bibliography.</p>	<p>1. Review: Intro to OCR (ECSE 2610)</p> <p>Preprocessing: Scanner calibration, correction of scan distortions; noise removal; text-figure separation; skew correction; gray-scale and color, text layout extraction (column, line, and word segmentation) [NG 00]. Character image defect models [KBH 94] Recovery of scanner distortions [BE 00]. Help Session: Wed. 6 pm Prof. E. Barney Smith.</p> <p>Features: Reflectance, geometric, & topological invariants [FG 60, SM 61] Features as weak classifiers [KE 00] N-tuples and feature selection [JN 95, JDM 00, JKNS 96]</p> <p>Resource person: Dr. D-M Jung, Yahoo!</p> <p>Static single-pattern classifiers: <i>Bayes:</i> Single & Multimodal, Linear, Quadratic, Gaussian and Bilevel [DHS 01, LKF 01] <i>Neural Networks:</i> Backprop, LVQ, RBF [BC 95] <i>Support Vector Machine</i> [VV 98] <i>Nearest Neighbors</i> [DHS 01] <i>Decision Trees and Forests</i> [AGW 07, TH 98]</p> <p>Classifier training: Sample size and dimensionality [RJ 91] Bias and variance [GBD 92] Bagging, Boosting, Random Subspaces [JDM 00] Clustering [TK 99] Expectation Maximization [DLR 77, RW 84]</p>



Reading order?

Holnap este talan
hazajovok, de lehet,
hogy nem tudok.

Mikor vacsorazunk?
Szeretnek enni egy
burgonyat.



Miert akarod hogy
okvetlenul jojjek?
Talan lesz vendeg?
Biztos sokaig tart.

Elo kell nekem
kesziteni ket
kulombozo elo-
adast.

“Logical” layout analysis

Identify *domain-specific* components

e.g. title, author, affiliation, page number,...

or wire, component, connector, label,...

or destination name, street, number, city, county, ...

or clef, note, duration mark, ...

- DID
- Page grammars

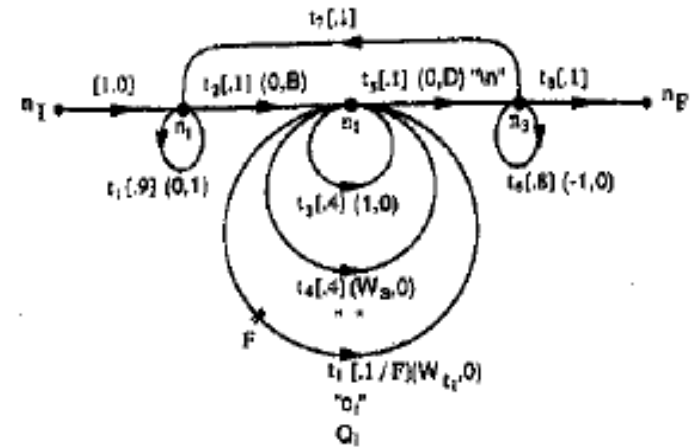
Some examples of publication-specific rules (for articles in IEEE –PAMI):

- **Title-lines** are set in 21/23-point roman bold.
- There are at most 4 lines in the title.
- **Bylines** follow the title and are separated by 17-point leading.
- Bylines are set in 10/12-point roman all-caps.
- **Text paragraphs** are indented, except the first, which begins with a 26-point drop-cap.
- The **page numbers** are set flush with the margin and alternate from left to right.
- **Footnotes** are set 6/7 point, numbered with leading superscripts, and separated from the narrative by at least 4-point leading.

DID

Gary Kopec and Phil Chou

Communication theory framework for document recognition



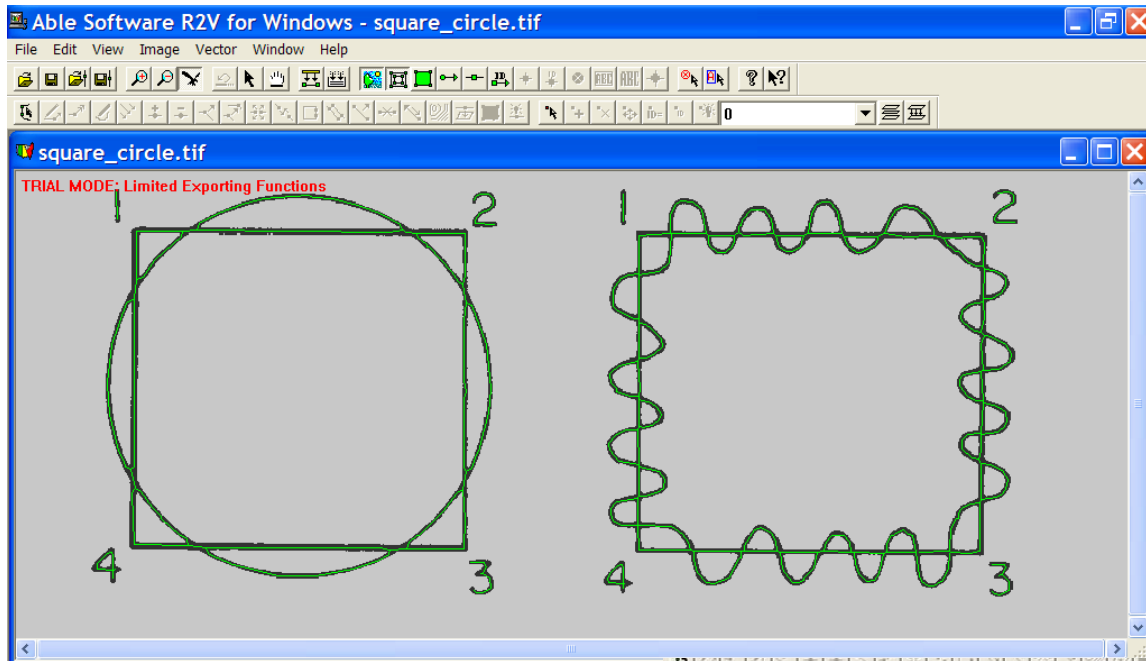
Document Image Decoding:

Whole page recognition with stochastic attributed context-free grammars.

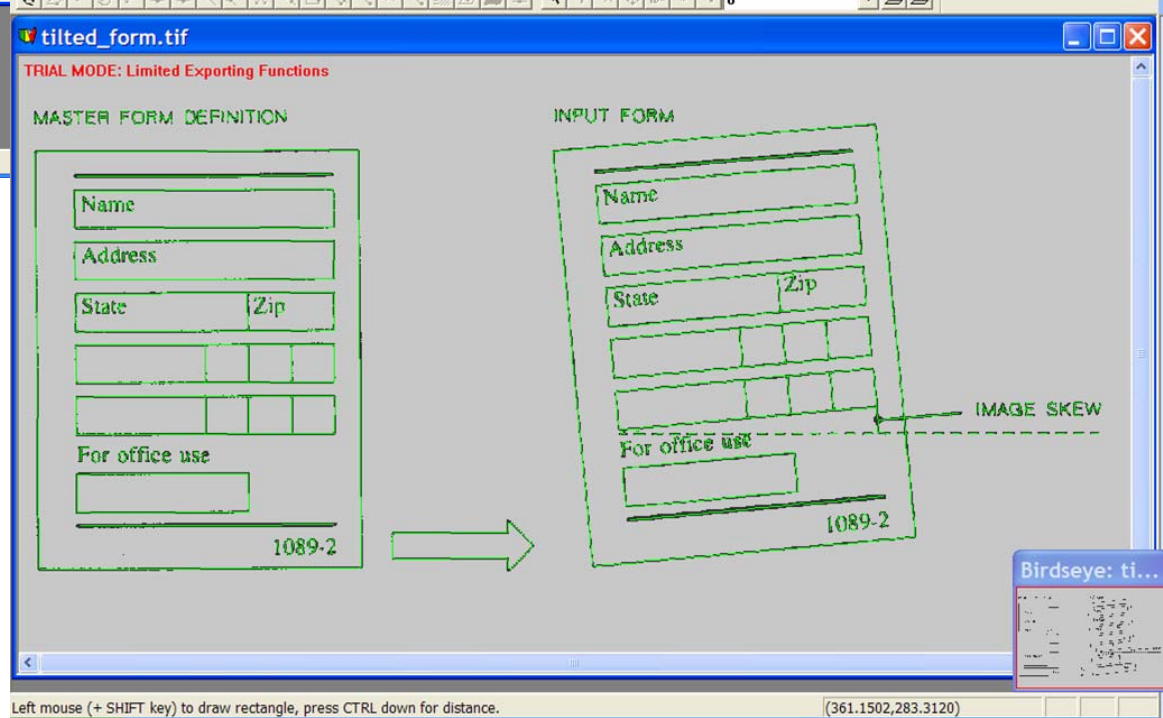
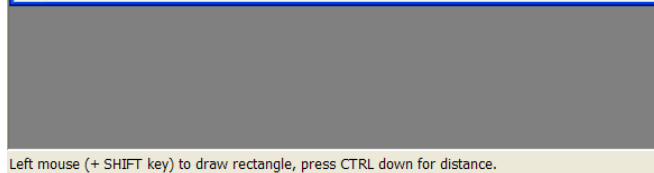
Based on earlier work on a text-image editor,
Image EMACS

Line-drawing and Map conversion

- Always interactive
- Layers sometimes available as hardcopy
- Auto-vectorization superimposed on bitmap
- Approved segments change color
- Label entry facilitated by *grouping labels*
- Loss of context (partial display) disorients
- Software costs vary over two orders of magnitude



R2V



5/15/2008

Other DIP applications

- Document authentication:
watermarks and signatures
- Duplicate detection
- Redaction recovery
- Access control (CAPTCHAs)
- Text in photos and video
- Specialized scripts (music, chess, ...)

OUTLINE

- The evolution of documents
- Advances in document image capture
- Document image analysis
- **Challenges** (take the baton and run!)
- To read further

Tech-text

- Convert formulas and equations to executable form
- Tables: recover header to content-cell relations
- Illustrations: categorize (graph, photo, chart) and extract text for downstream use
- OCR: beyond within-class style
- **Don't waste operator interventions!**
- Web documents for stuffing databases: “curators” are expensive and have a limited attention span (e.g. biology, sociology, marketing, intelligence)

Bureaucratic forms

- Currently one-form at a time: should exploit information from forms processed earlier, and from forms arriving **later!**
- Accumulate layout/classification data from all processed forms, right or wrong: the current data stream is more representative than your training set.
- **Don't waste operator interventions!**

Historical documents

- Combine interactive processing with ground-truthing
- Scan it right the first time, and keep calibration information for reprocessing with future algorithms.
- Instead of attempting to clean up the document, seek distortion-invariant algorithms.
- Don't waste operator interventions!

Evaluation

- Evaluate intermediate stages, e.g., text/photo/line-art segmentation (pixel level? rectangles? overlaps? GT?)
- Several orders of magnitude discrepancy between test data used for IR (e.g. TREC) and test data used for DIA (e.g. UW)
- **Metrics for interactive processing** (because it won't disappear anytime soon)
- Same metrics for semi-automated ground-truthing

Digital libraries

- Dual format?
- Links between image and text?
- More/different metadata? (glosses, translations)
- Interoperability?
- GUI (Google is often easier!)
- Selective access?
- On-line or automated reference librarians?
- Relations with monster publishers?
- Interface/merge with the semantic web?
- Perpetual exponential growth?

OUTLINE

- The evolution of documents
- Advances in document image capture
- Document image analysis
- Challenges
- To read further

(several thousand papers have been published in the last decades.)

Some tools

- PHOTOSHOP
- MATLAB
- PAINT
- LEADTOOLS
- KHOROS (Khoral)– VisiQuest (Accusoft)
- MATROX
- PEGASUS IMAGING
- PBMPLUS (unix)
- Quick MTF
- R2V
- xv (open source unix, dated)

<http://www.cs.cmu.edu/~cil/v-source.html>

Regular conferences

- **ICDAR** (Int'l Conf on Document Analysis and Recognition) biennial
- **DAS** (Document Analysis Workshop) biennial
- **DRR** (SPIE/IST Document Recognition and Retrieval) annual
- **DOCENG** (ACM Conf on Document Engineering) annual
- **DIAL** (Document Image Analysis for Libraries) two so far
- **SDAIR** (UNLV Doc Analysis and Information Retrieval) 1992-1996

Many papers are presented at pattern recognition, machine learning, and image processing conferences.

The only dedicated journal is **IJDAR**, but many papers are published in IEEE-PAMI, IEEE-IP, PRL, IJPRAI, etc.

My ~recent DIA reviews have many references

- S. Rice, G. Nagy, and T.A. Nartkey, *Optical Character Recognition: An Illustrated Guide to the Frontier*, Kluwer Academic Publishers, Boston/Dordrecht/London, 200 pages, 1999.
- G. Nagy, "Twenty Years of Document Image Analysis in PAMI," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, #1, 20th Anniversary Issue, pp. 38-62, January 2000.
- G. Nagy and D. Lopresti, "Issues in ground-truthing graphic documents," *Lecture Notes in Computer Science*, pp. 46-66, Springer, 2002 (selected papers from the Fourth International Workshop on Graphics Recognition).
- G. Nagy, S. Veeramachaneni, "Adaptive and interactive approaches to document analysis," in *Machine Learning in Document Analysis and Recognition* (S. Marinai, H. Fujisawa, editors), Springer, *Studies in Computational Intelligence*, Vol. 90, ISBN 978-3-540-76279-9, pp. 221-257, 2008.
- G. Nagy and D. Lopresti, "The role of document image analysis in trustworthy elections, in *Document Analysis and Retrieval*" (B.B. Chaudhuri, S.K. Parui, editors), World Scientific, in press, May 2008.
- G. Nagy, "Digitizing, coding, annotating, disseminating, and preserving documents," *Procs. IWRIDL-2006 workshop on Digital Libraries*, Kolkata, India, 2006, ACM 1-59593-608-4, 2008.

<http://www.ecse.rpi.edu/homepages/nagy/>

Summary

- Most new documents are computer produced:
DIA may have already peaked.
- Digitization of hardcopy now essentially lossless.
- There are no universal solutions:
make use of all available information.
- Few systems learn from experience.
- The remaining DIA problems require large contexts:
figures, equations, tables, degraded documents.
- Until computers learn to organize information
autonomously, such problems will require interaction.
- **Interaction must not be wasted.**

Thank you

Questions?

