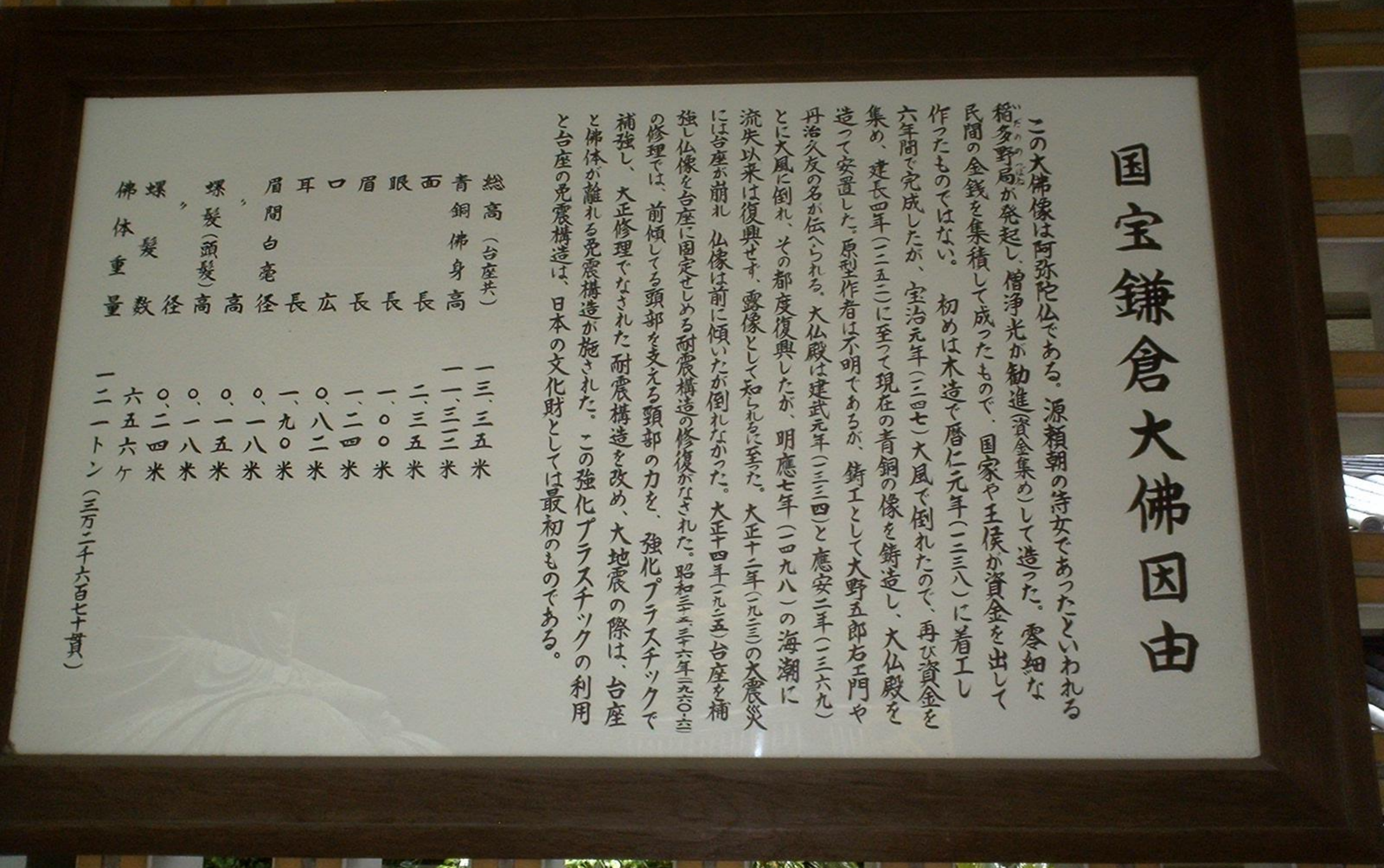
Scanned Documents

LBSC 796/INFM 718R

Douglas W. Oard

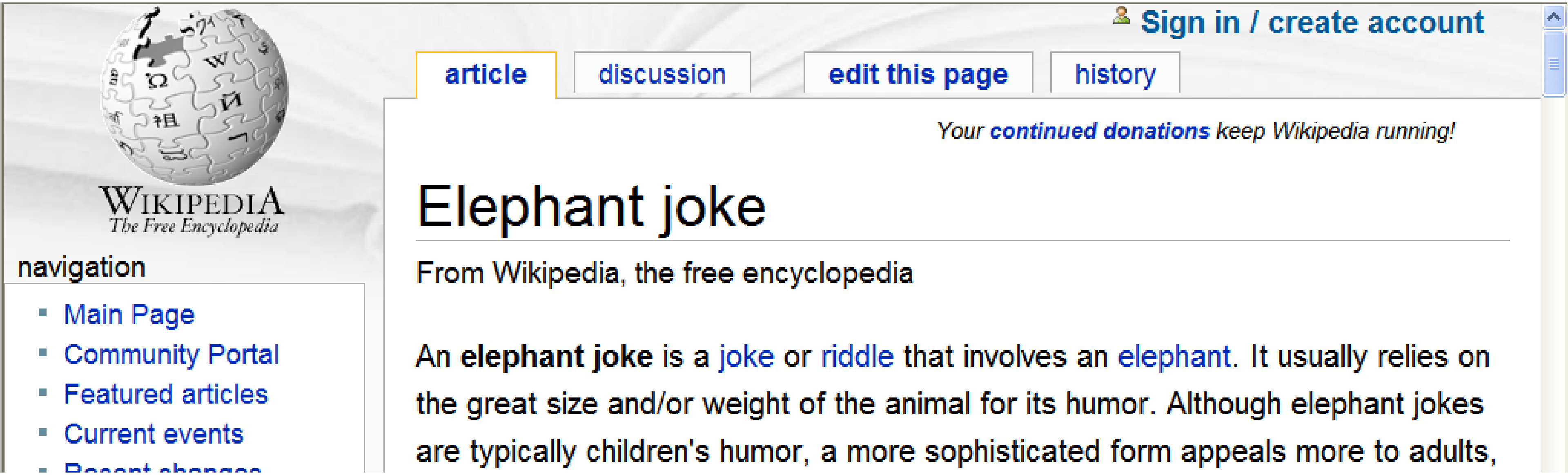
Week 8, March 30, 2011



国宝鎌倉大佛因由

この大佛像は阿弥陀仏である。源頼朝の侍女であったといわれる稲多野局が発起し、僧浄光が勧進（資金集め）して造った。零細な民間の金銭を集積して成ったもので、国家や王侯が資金を出して作ったものではない。　初めは木造で暦仁元年（一二三八）に着工し六年間で完成したが、宝治元年（一二四七）大風で倒れたので、再び資金を集め、建長四年（一二五二）に至って現在の青銅の像を鋳造し、大仏殿を造って安置した。原型作者は不明であるが、鋳工として大野五郎右エ門や丹治久友の名が伝へられる。大仏殿は建武元年（一三三四）と應安二年（一三六九）とに大風に倒れ、その都度復興したが、明應七年（一四九八）の海潮に流失以来は復興せず、露像として知られるに至った。大正十二年（一九二三）の大震災には台座が崩れ　仏像は前に傾いたが倒れなかった。大正十四年（一九二五）台座を補強し仏像を台座に固定せしめる耐震構造の修復がなされた。昭和三十五、三十六年（一九六〇・六一）の修理では、前傾してる頭部を支える頸部の力を、強化プラスチックで補強し、大正修理でなされた耐震構造を改め、大地震の際は、台座と佛体が離れる免震構造が施された。この強化プラスチックの利用と台座の免震構造は、日本の文化財としては最初のものである。

総高（台座共）	一三、三五米
青銅佛身高	一一、三一二米
面長	二、三五米
眼長	一、〇〇米
眉長	一、二〇四米
口広	〇、八二米
耳長	一、九〇米
眉間白毫径	〇、一八米
〃　高	〇、一五米
螺髪（頭髪）高	〇、一八米
〃　径	〇、二四米
螺髪数	六五六ケ
佛体重量	一二一トン（三万二千六百七十貫）

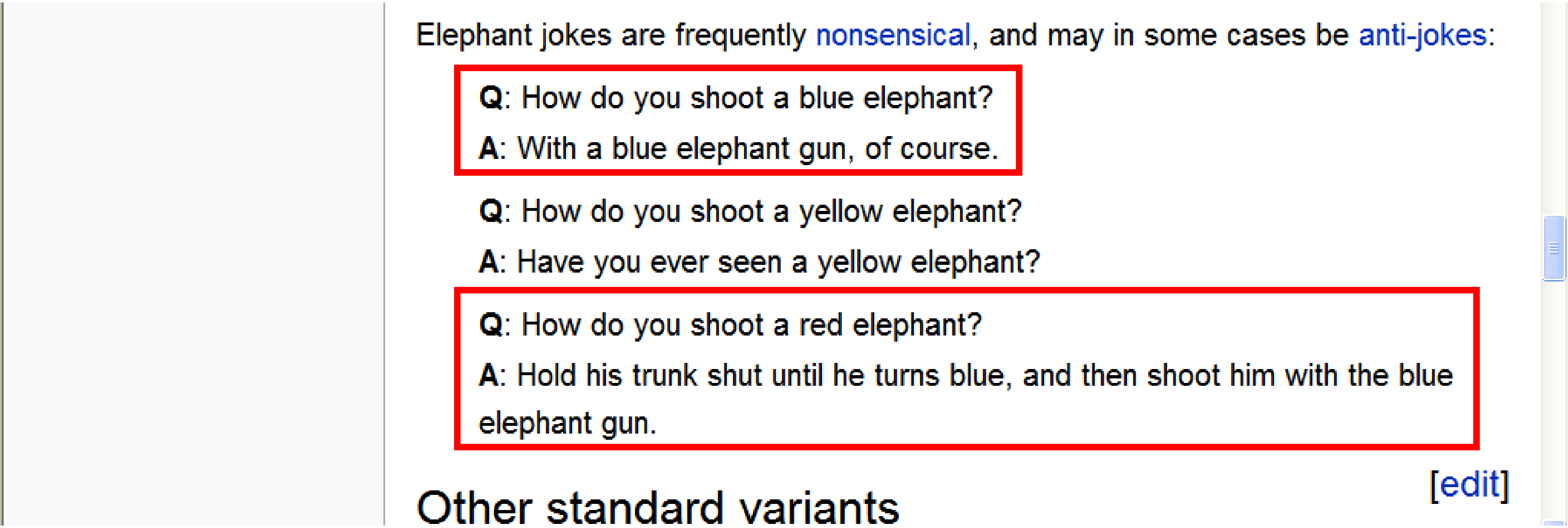


*Your **continued donations** keep Wikipedia running!*

Elephant joke

From Wikipedia, the free encyclopedia

An **elephant joke** is a joke or riddle that involves an elephant. It usually relies on the great size and/or weight of the animal for its humor. Although elephant jokes are typically children's humor, a more sophisticated form appeals more to adults,



Elephant jokes are frequently nonsensical, and may in some cases be anti-jokes:

Q: How do you shoot a blue elephant?

A: With a blue elephant gun, of course.

Q: How do you shoot a yellow elephant?

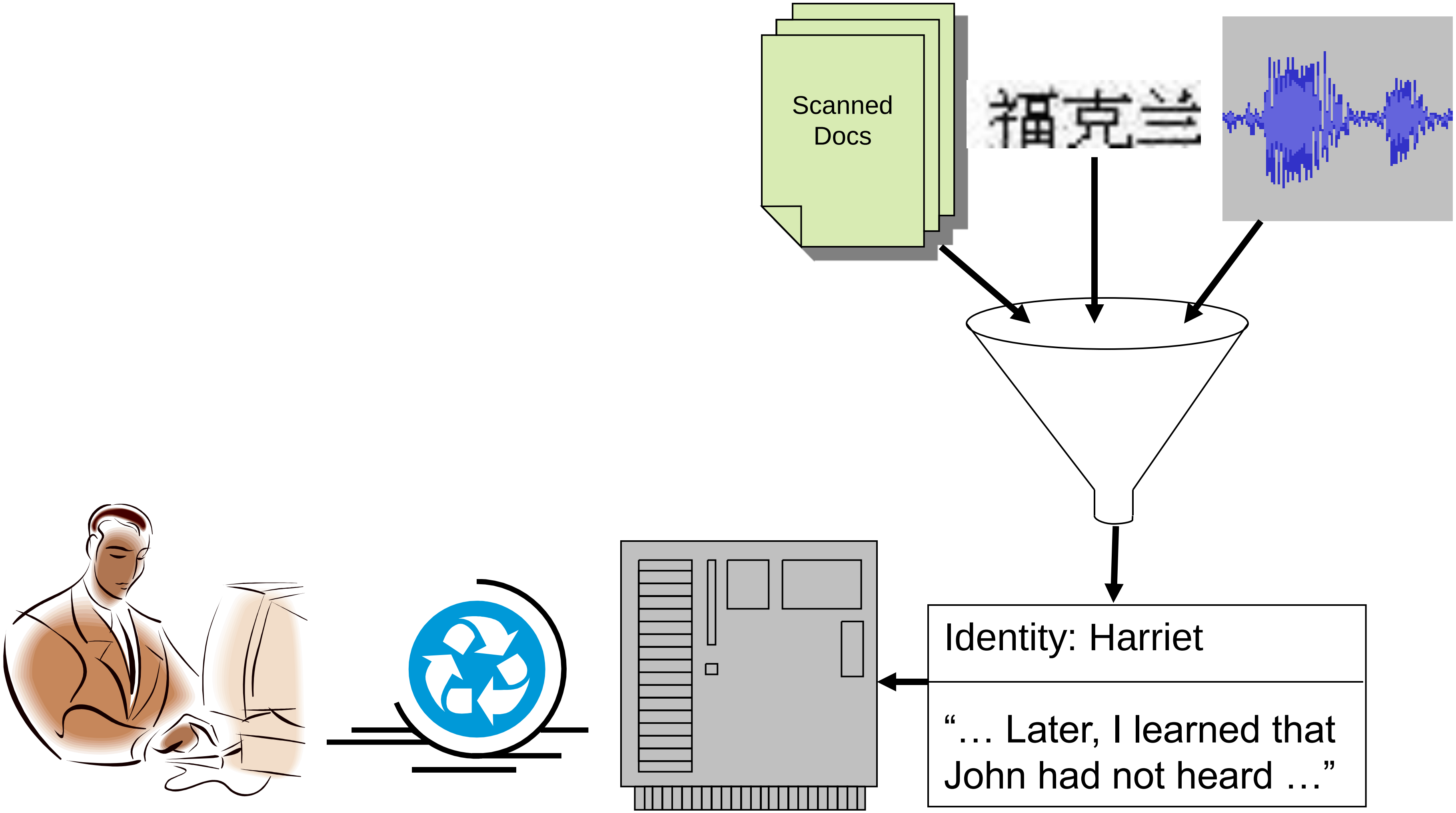
A: Have you ever seen a yellow elephant?

Q: How do you shoot a red elephant?

A: Hold his trunk shut until he turns blue, and then shoot him with the blue elephant gun.

Other standard variants

Expanding the Search Space



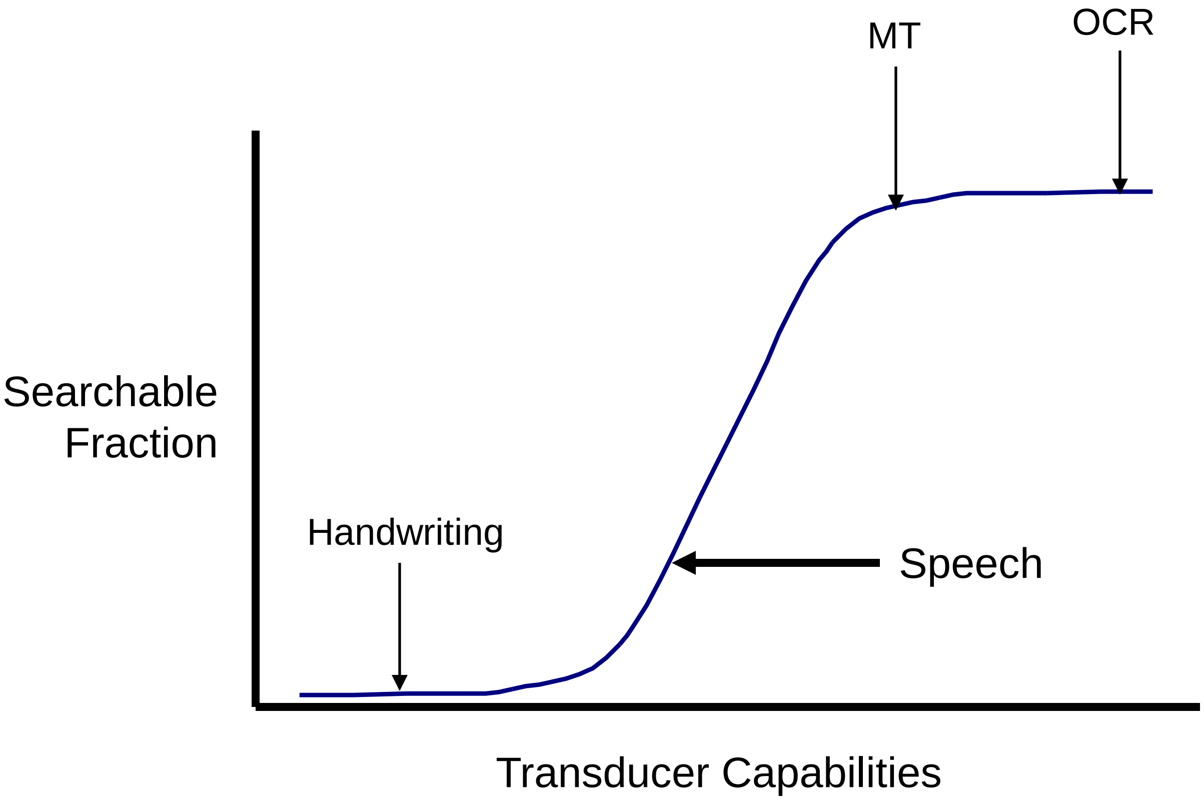
Scanned Docs

福克兰

Identity: Harriet

“… Later, I learned that John had not heard …”

High Payoff Investments



$$\frac{\textit{accurately recognized words}}{\textit{words produced}}$$

Some Applications

- Case management for litigation
- Duplicate detection for declassification productivity and anti-tiling
- Knowledge management from everything I have ever xeroxed or faxed

Indexing and Retrieving Images of Documents

LBSC 796/INFM 718R

David Doermann, UMIACS

Agenda

- Questions
- Definitions - Document, Image, Retrieval
- Document Image Analysis
  - Page decomposition
  - Optical character recognition
- Traditional Indexing with Conversion
  - Confusion matrix
  - Shape codes
- Doing things Without Conversion
  - Duplicate Detection, Classification, Summarization, Abstracting
  - Keyword spotting, etc

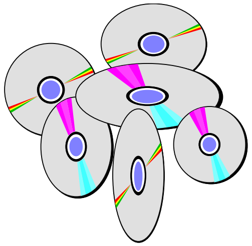
Goals of this Class

- Expand your definition of what is a “DOCUMENT”
- To get an appreciation of the issues in document image analysis and their effects on indexing
- To look at different ways of solving the same problems with different media
- Your job: compare/contrast with other media

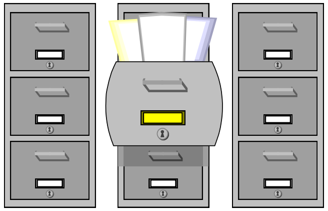
Quiz

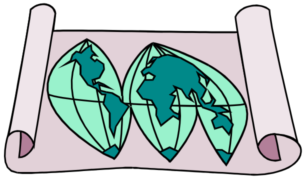
- What is a document?

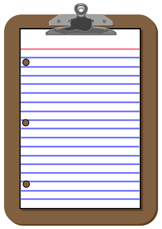
Document

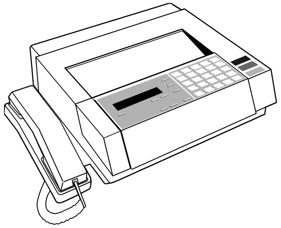


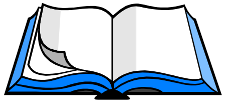
- Basic Medium for Recording Information
- Transient
  - Space
  - Time
- Multiple Forms
  - Hardcopy (paper, stone, ..) / Electronic (CDROM, Internet, …)
  - Written/Auditory/Visual (symbolic, scenic)
- Access Requirements
  - Search
  - Browse
  - “Read”

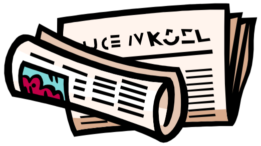




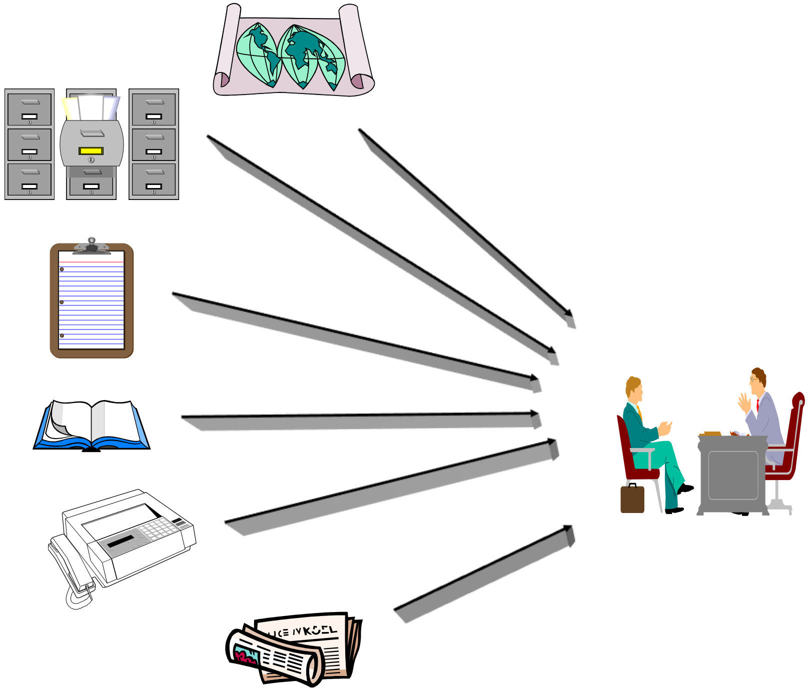








Sources of Document Images



- The Web
  - Some PDF files come from scanned documents
  - Arabic news stories are often GIF images
- Digital copiers
  - Produce “corporate memory” as a byproduct
- Digitization projects
  - Provide improved access to hardcopy documents

Some Definitions

- Modality
  - A means of expression
- Linguistic modalities
  - Electronic text, printed, handwritten, spoken, signed
- Nonlinguistic modalities
  - Music, drawings, paintings, photographs, video
- Media
  - The means by which the expression reaches you
    - Internet, videotape, paper, canvas, …

Quiz

- What is a document?

- What is an image?

Images

- Pixel representation of intensity map
- No explicit “content”, only relations
- Image analysis
  - Attempts to mimic human visual behavior
  - Draw conclusions, hypothesize and verify

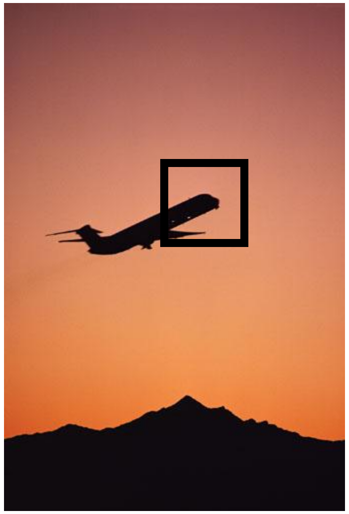
Image databases

Use primitive image analysis to represent content

Transform semantic queries into “image features”

color, shape, texture …

spatial relations



10	27	33	29
27	34	33	54
54	47	89	60
25	35	43	9

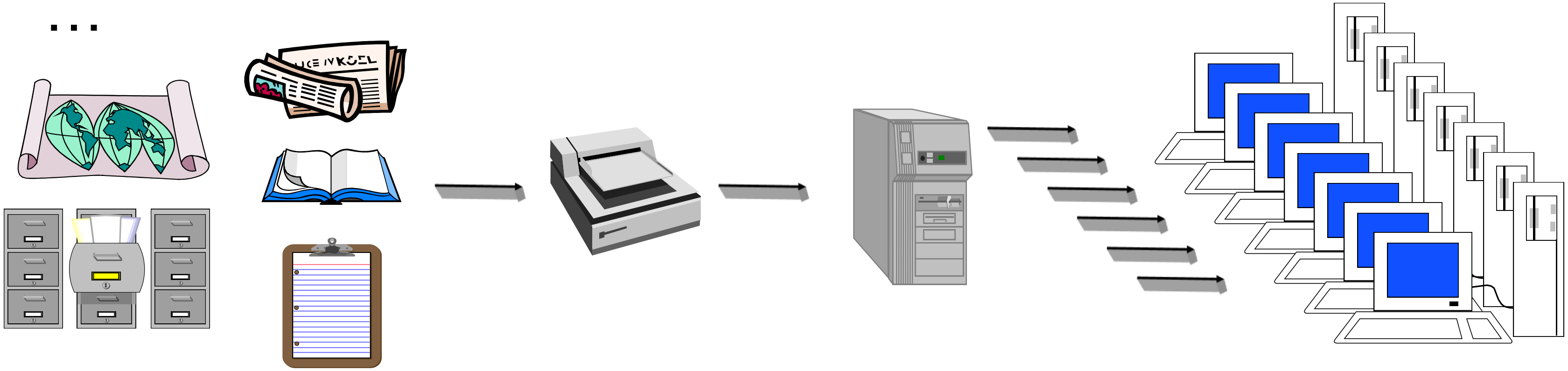
Document Images

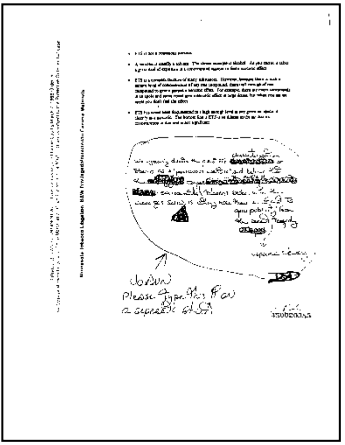
- A collection of dots called “pixels”
  - Arranged in a grid and called a “bitmap”
- Pixels often binary-valued (black, white)
  - But greyscale or color is sometimes needed
- 300 dots per inch (dpi) gives the best results
  - But images are quite large (1 MB per page)
  - Faxes are normally 72 dpi
- Usually stored in TIFF or PDF format

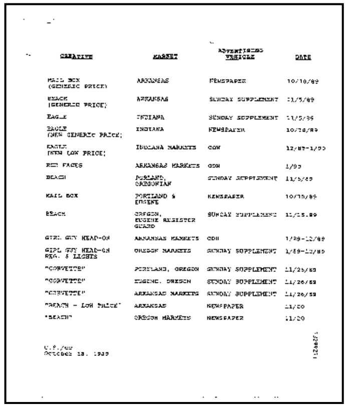
Yet we want to be able to process them like text files!

Document Image "Database"

- Collection of scanned images
- Need to be available for indexing and retrieval, abstracting, routing, editing, dissemination, interpretation …

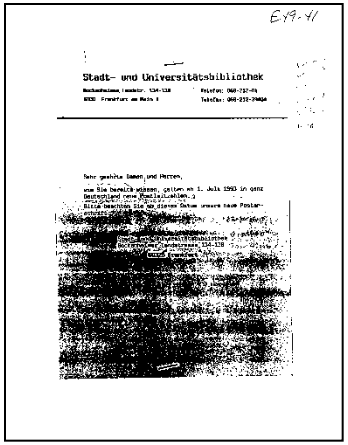






CREATIVE	MARKET	ADVERTISING VEHICLE	DATE
MAIL BOX (GENERIC PRICE)	ARKANSAS	NEWSPAPER	10/10/89
BEACH (GENERIC PRICE)	ARKANSAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (NEW GENERIC PRICE)	INDIANA	NEWSPAPER	10/28/89
EAGLE (NEW LOW PRICE)	INDIANA MARKETS	OON	12/89-1/90
RED FACES	ARKANSAS MARKETS	OON	1/90
BEACH	PORTLAND, OREGONIAN	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & EUGENE	NEWSPAPER	10/30/89
BEACH	OREGON, EUGENE REGISTER GUARD	SUNDAY SUPPLEMENT	11/15.89
GIRL GUY HEAD-ON	ARKANSAS MARKETS	OON	1/89-12/89
GIRL GUY HEAD-ON REG. & LIGHTS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	EUGENE, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	ARKANSAS MARKETS	SUNDAY SUPPLEMENT	11/26/89
"BEACH – LOW PRICE"	ARKANSAS	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

October 18, 1989



E49-41

Stadt- und Universitätsbibliothek

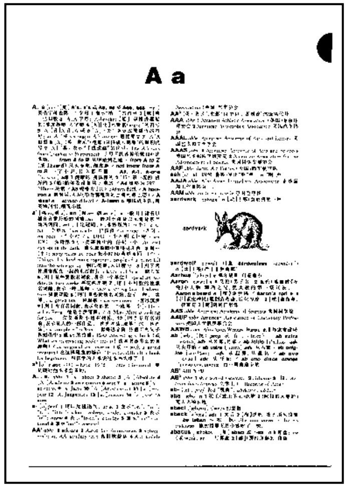
Telefon: 069-212-01
Telefax: 069-212-39404

Sehr geehrte Damen und Herren,

wie Sie bereits wissen, gelten ab 1. Juli 1993 in ganz Deutschland neue Postleitzahlen.

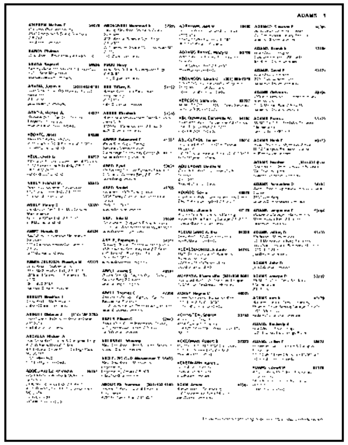
Bitte beachten Sie ab diesem Datum unsere neue Postanschrift:

Stadt- und Universitätsbibliothek
Bockenheimer Landstrasse 134-138
Frankfurt



A a

aardvark



ADAMS 1

The Washington Post

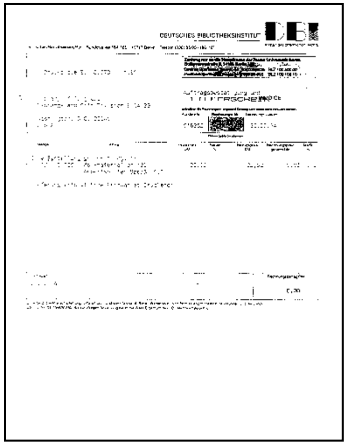
After Zarqawi, No Clear Path In Weary Iraq

The Young Apprentice

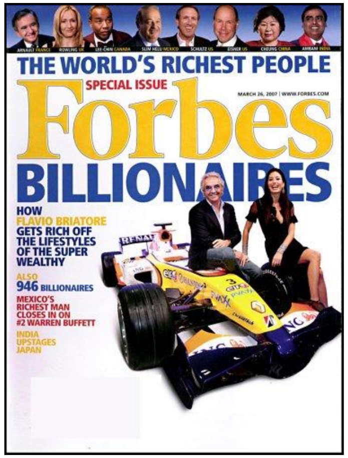
How U.S. Forces Found Iraq's Most-Wanted Man

Kaine Delays Execution Of Inmate for 6 Months

FDA Approves Vaccine That Should Prevent Most Cervical Cancers



DEUTSCHES BIBLIOTHEKSINSTITUT



THE WORLD'S RICHEST PEOPLE

SPECIAL ISSUE

Forbes

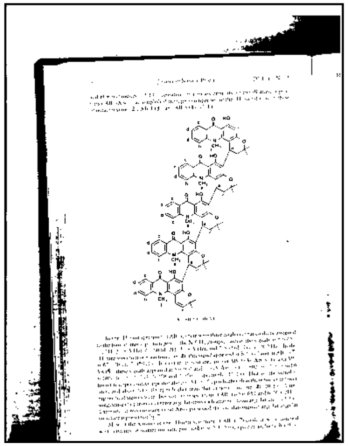
BILLIONAIRES

HOW FLAVIO BRIATORE GETS RICH OFF THE LIFESTYLES OF THE SUPER WEALTHY

ALSO 946 BILLIONAIRES

MEXICO'S RICHEST MAN CLOSES IN ON #2 WARREN BUFFETT

INDIA UPSTAGES JAPAN



FINANCIAL TIMES

Collateral damage

Iraq braced as Saddam rejects Bush ultimatum

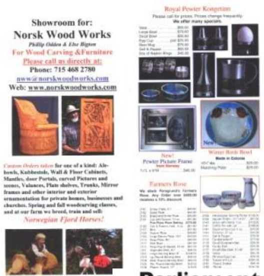
Fed holds rates but signals uncertain future

US probes Ahold collusion claims

Language And Media Processing

LAMP

Other “Documents”


Showroom for:
Norsk Wood Works
Phillip Odden & Else Bigton
For Wood Carving &Furniture
Please call us directly at:
Phone: 715 468 2780
nww@norskwoodworks.com
Web: www.norskwoodworks.com

Custom Orders taken for one of a kind: Ale-bowls, Kubbestole, Wall & Floor Cabinets, Mantles, door Portals, carved Pictures and scenes, Valances, Plate shelves, Trunks, Mirror frames and other interior and exterior ornamentation for private homes, businesses and churches. Spring and fall woodcarving classes, and at our farm we breed, train and sell:

Norwegian Fjord Horses!

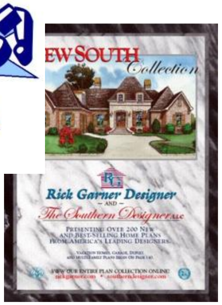
Royal Pewter Kongetinn
Please call for prices. Prices change frequently.
We offer many specials.

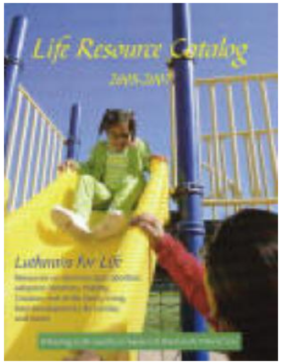
New!
Pewter Picture Frame

Winter Birds Bowl
Made in Colorado

Farmers Rose

DELUXE PACKAGE
Business Letter Professional
2005
Best letter-making software for marketing managers, home businesses, consultants, accountants, secretaries, customer service, HR managers, and CEOs!


EW SOUTH
Collection
Rick Garner Designer
— AND —
The Southern Designer
PRESENTING OVER 200 NEW AND BEST-SELLING HOME PLANS FROM AMERICA'S LEADING DESIGNERS



Order online now at www.poker-wear.com
Prices shown here are subject to change without notice.

Poker Tables from Poker-Wear

You can't play poker without a poker table and Poker-Wear.com has a great selection of folding poker tables you can set up for the big poker game and store away until the next big poker game. Or, you can leave these great looking poker tables set up all the time.

Poker-Wear.com is offering FREE SHIPPING on poker table orders anywhere in the 48 states. This is a limited time offer. Order your Poker-Wear poker table today!

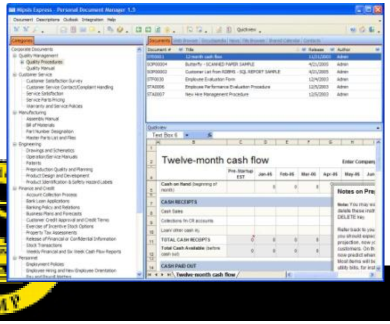
Poker Table

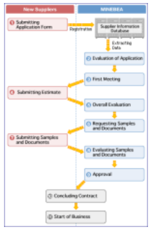
Poker Table

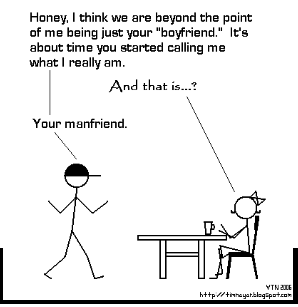
Poker Table

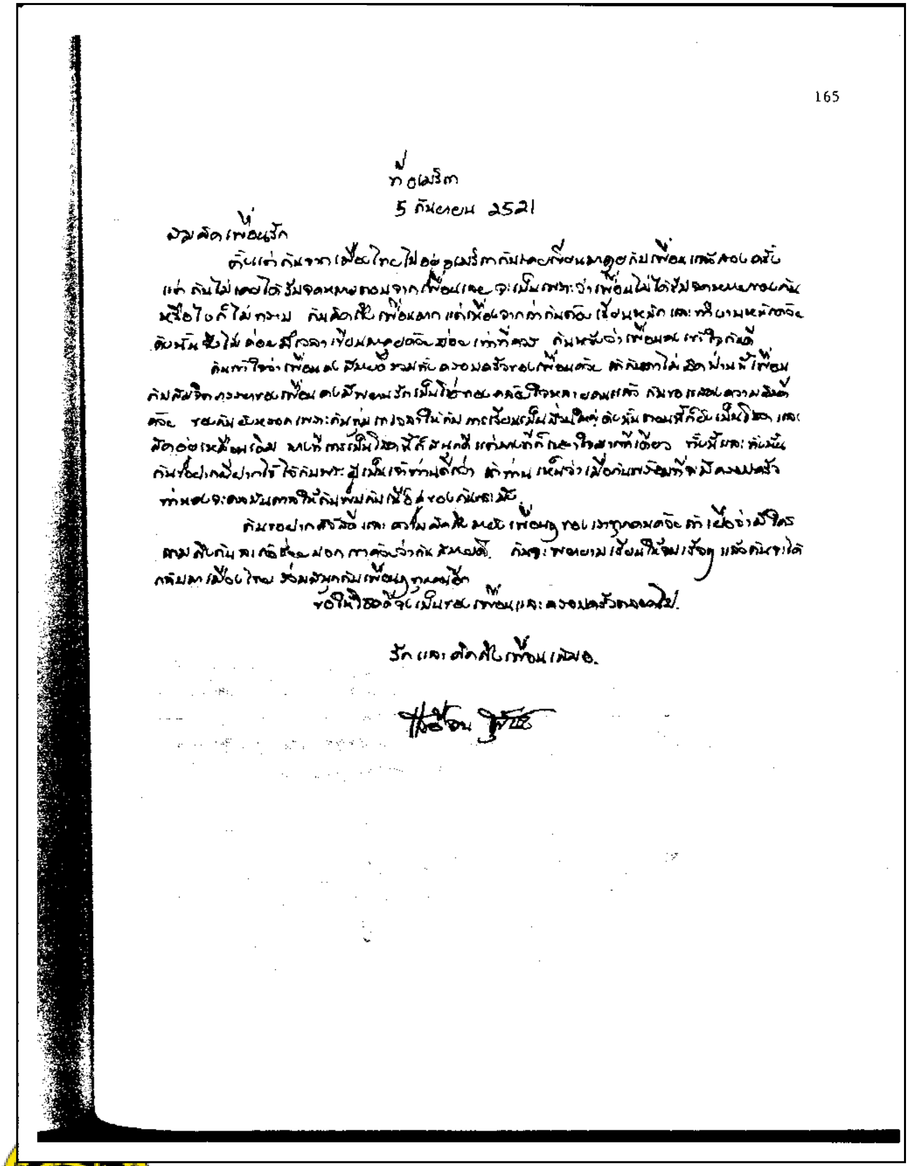
Poker Table

Poker-Wear.com 3121 Potts Crossing Smyrna, TN 37167 USA	Phone 615.666.9112	Hours Monday-Friday 9 AM to 5 PM Central


Twelve-month cash flow


New Suppliers
MINERVA
Submitting Application Form
Registration
Supplier Information Database
Extracting Data
Evaluation of Application
First Meeting
Submitting Estimate
Overall Evaluation
Requesting Samples and Documents
Submitting Samples and Documents
Evaluating Samples and Documents
Approval
Concluding Contract
Start of Business


Honey, I think we are beyond the point of me being just your "boyfriend." It's about time you started calling me what I really am.
And that is...?
Your manfriend.
VTN 2005
http://timmeyer.blogspot.com



ที่อเมริกา
5 กันยายน 2521

สมคิดเพื่อนรัก

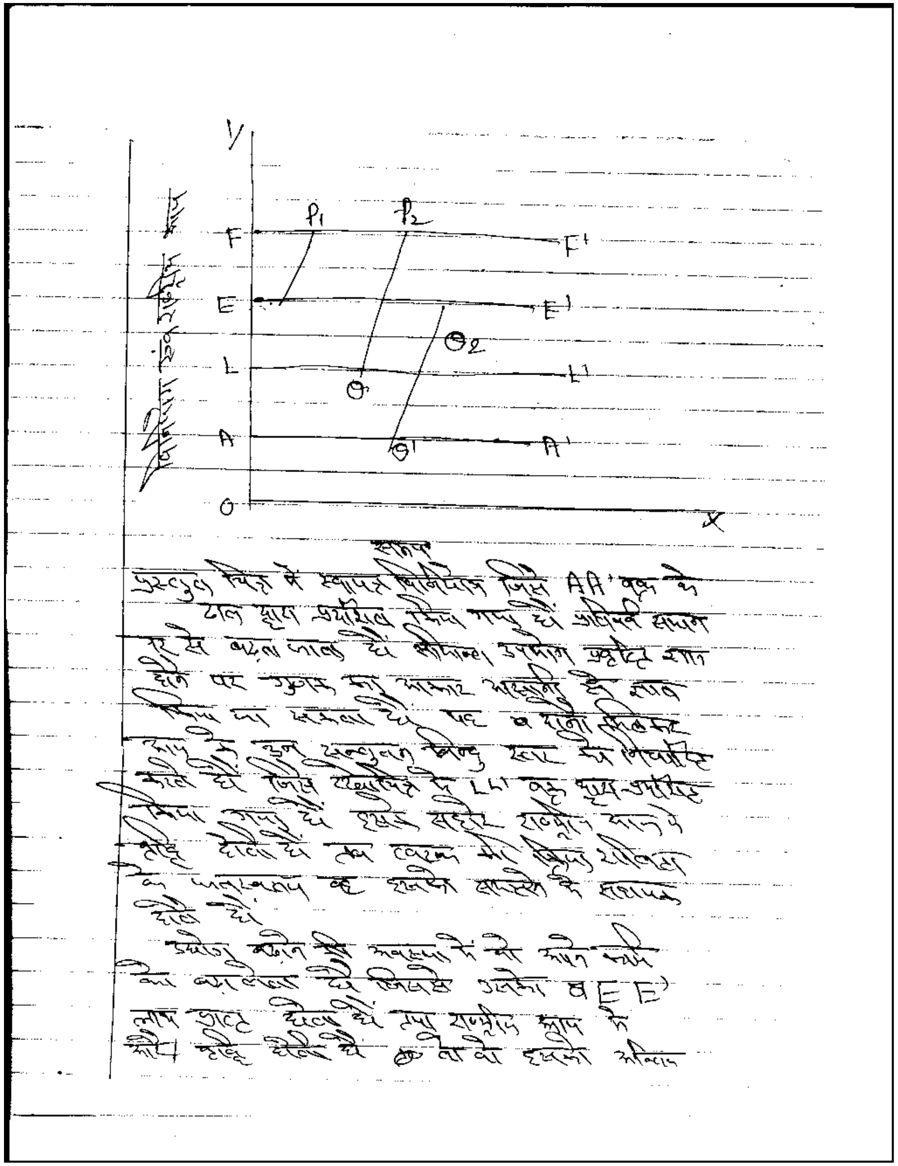
ตั้งแต่ฉันจากเมืองไทยไปอยู่อเมริกาฉันเคยเขียนมาถึงกับเพื่อนหลายครั้ง แต่ฉันไม่เคยได้รับจดหมายตอบจากเพื่อนเลย จะเป็นเพราะว่าเพื่อนไม่ได้รับจดหมายของฉัน หรือไงก็ไม่ทราบ ฉันคิดถึงเพื่อนมาก แต่เมื่อจากกันก็ต้อง [illegible] และ [illegible] ฉันหวังว่าจะได้มีเวลาเขียนจดหมายถึงฉันบ้างเท่าที่ควร ฉันหวังว่าเพื่อนคงเข้าใจฉันดี

ฉันหวังว่าเพื่อนคงสบายดี รวมทั้งครอบครัวของเพื่อนด้วย [illegible] เพื่อนคงสบายดีนะ [illegible] ฉันขอแสดงความยินดีด้วย [illegible] เพราะฉันรู้ [illegible] เป็นส่วนใหญ่ [illegible] และ [illegible] ฉันนี้ [illegible] ก็ [illegible] ท่าน [illegible] ครอบครัว [illegible]

ฉันขอบอกว่า [illegible] และ [illegible] เพื่อน [illegible] ฉันจะพยายามเขียนให้เพื่อน [illegible] แล้วฉันจะได้กลับมาเมืองไทย [illegible] เพื่อน [illegible]

ขอให้โชคดีจงเป็นของเพื่อนและครอบครัวตลอดไป.

รักและคิดถึงเพื่อนเสมอ.



Y
F
E
L
A
O
X
P1
P2
F'
E'
L'
A'
θ
θ1
θ2

[illegible]

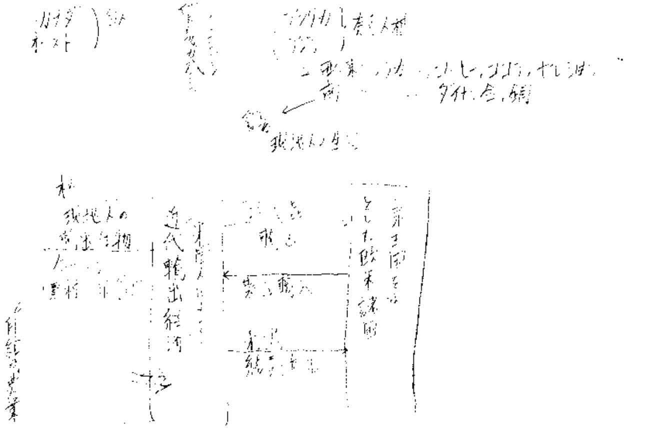
प्रस्तुत चित्र में स्थायी [illegible] जिसे AA' वक्र से [illegible] द्वारा प्रदर्शित किया गया है प्रतिफल समान रहते हैं [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] LL' वक्र [illegible] किया गया है [illegible] [illegible] [illegible] [illegible] होता है तब [illegible] में [illegible] के [illegible] होते हैं

[illegible] [illegible] अवस्था में [illegible] [illegible] [illegible] जिसे [illegible] EE' [illegible] [illegible] [illegible] है [illegible] [illegible] होते हैं तथा [illegible] [illegible]

II. 途上国の社会経済的特質

1. 基本的構造 ～ 植民地型二重経済

[illegible]は歪んだ経済構造を作り出した。これは植民地型二重経済と呼ばれるもので、他の白人居住植民地（オーストラリア、NZ、カナダ、[illegible]）が自力の発展によって急速に伸びたのに比し、これら植民地は停滞を余儀なくされた。



植民地型二重経済 [illegible] 特化し [illegible] 特徴を [illegible] 多くは自給農業に依存し [illegible] 低い。
[illegible] 多くは [illegible]

1 6. The Three Kingdoms and the Six Dynasty

p85	férment	n.	불안. 동요. 정치적 동요
	usúrp	v.	강탈하다
	incéssant	a.	끊임없는
	shrink	v.	움츠러들다, 줄어들다. 기가 죽다 (p.p. shrunken)
	rashly	ad.	무모하게. 경솔하게, 성급하게
	mássacre	v.	학살하다
	flee	v.	달아나다. 도피하다 (p. fled)
	tribe	n.	부족. 종족
	abándon	v.	버리다. 포기하다
	nomádic	a.	유목민의. 유목 생활을 하는
	* Sínicize	v.	중국화하다. 중국식으로 만들다
	cávalry	n.	기병대, 기마부대
	refugée	n.	망명자, 피난자. 난민
	perpétual	a.	영구한. 부단한
	tórmoil	n.	혼란. 동요. 불안
p87	undermind	v.	약화시키다. 서서히 퇴화시키다
	* mónastery	n.	수도원. 사원 사찰(절)
	vast	a.	광대한. 거대한. 엄청난
	* propórtion	n.	크기. 비례. 비율
	realm /relm/	n.	왕국. 영역
	bureáucracy	n.	관료제. 관료정치, 관료주의
	* exért	v.	(힘.권력 등을) 행사하다, (영향을) 미치다

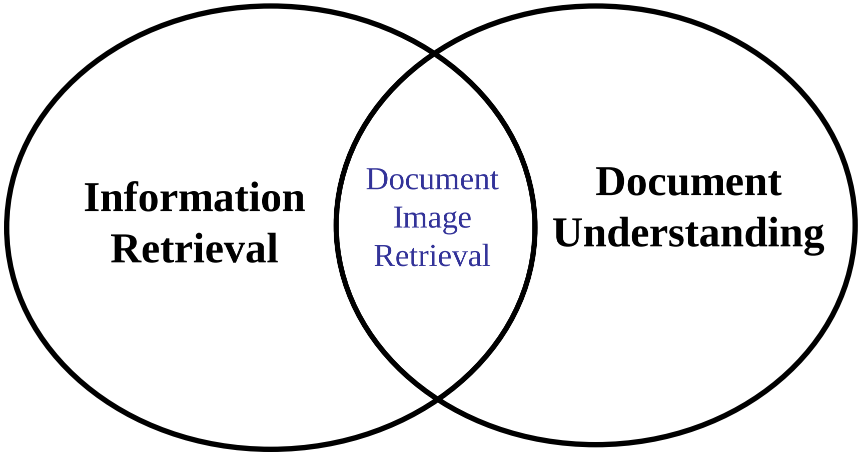
〈 Taoism 〉 도교

Táoism	n.	도교
*callígraphy	n.	서예. 서도, 달필. 필적
conglomerátion	n.	응집. 집합

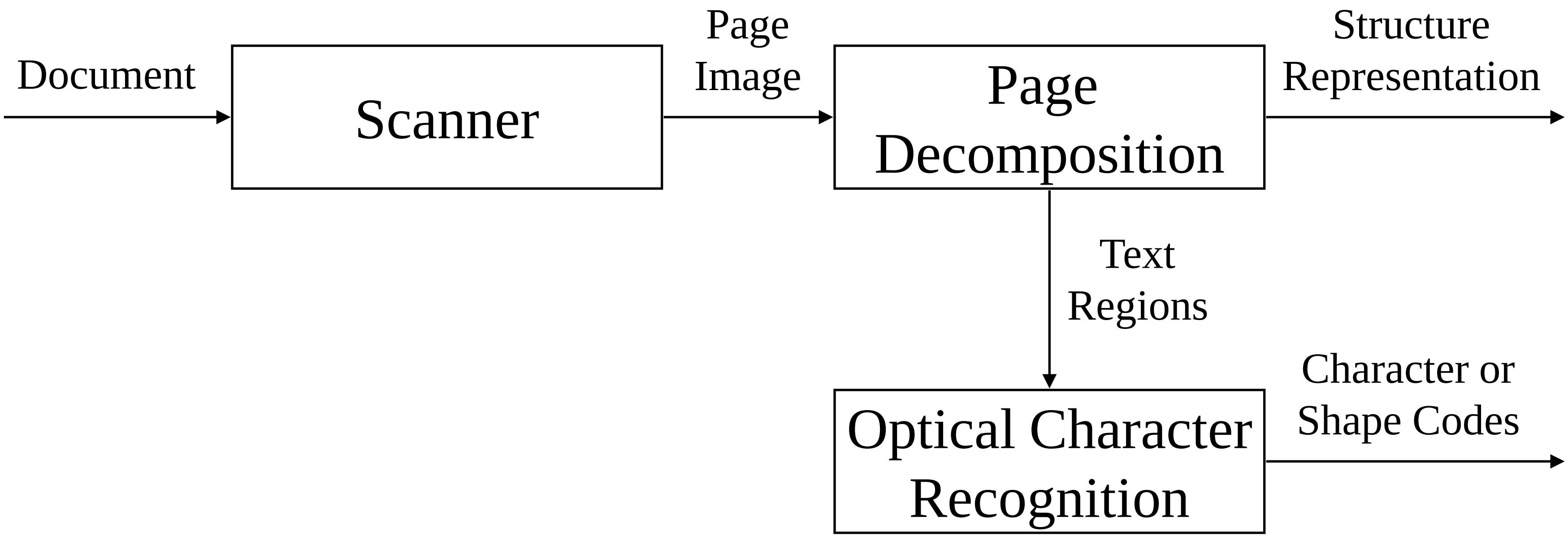
Quiz

- What is a document?
- What is an image?

- How can we *index and retrieve* document images?



Indexing Page Images



Document

Scanner

Page Image

Page Decomposition

Structure Representation

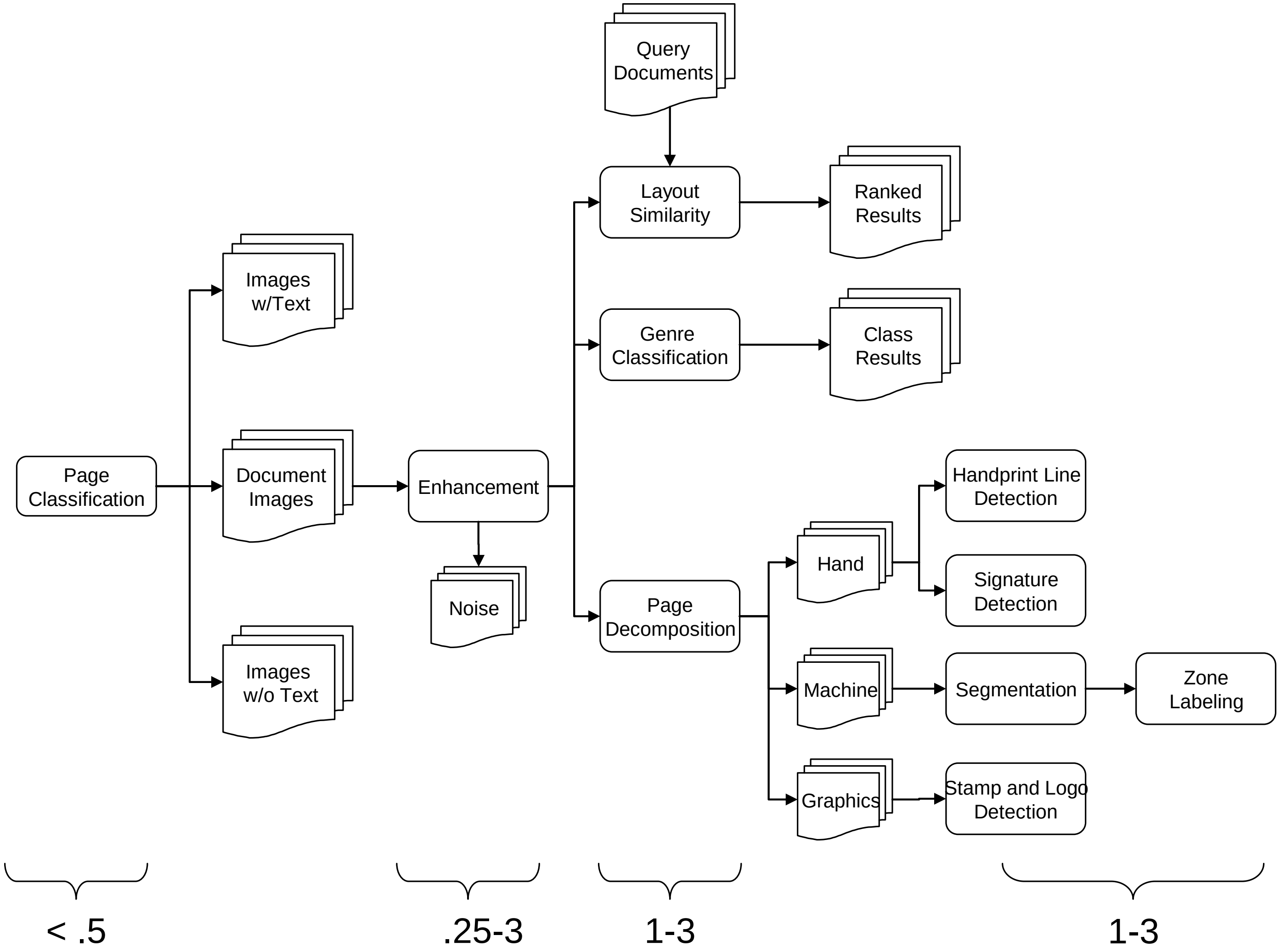
Text Regions

Optical Character Recognition

Character or Shape Codes

Document Image Analysis

- General Flow:
  - Obtain Image - Digitize
  - Preprocessing
  - Feature Extraction
  - Classification
- General Tasks
  - Logical and Physical Page Structure Analysis
  - Zone Classification
  - Language ID
  - Zone Specific Processing
    - Recognition
    - Vectorization



Query Documents

Page Classification

Images w/Text

Document Images

Images w/o Text

Enhancement

Noise

Layout Similarity

Ranked Results

Genre Classification

Class Results

Page Decomposition

Hand

Machine

Graphics

Handprint Line Detection

Signature Detection

Segmentation

Zone Labeling

Stamp and Logo Detection

< .5

.25-3

1-3

1-3

Target Processing Speed in Seconds

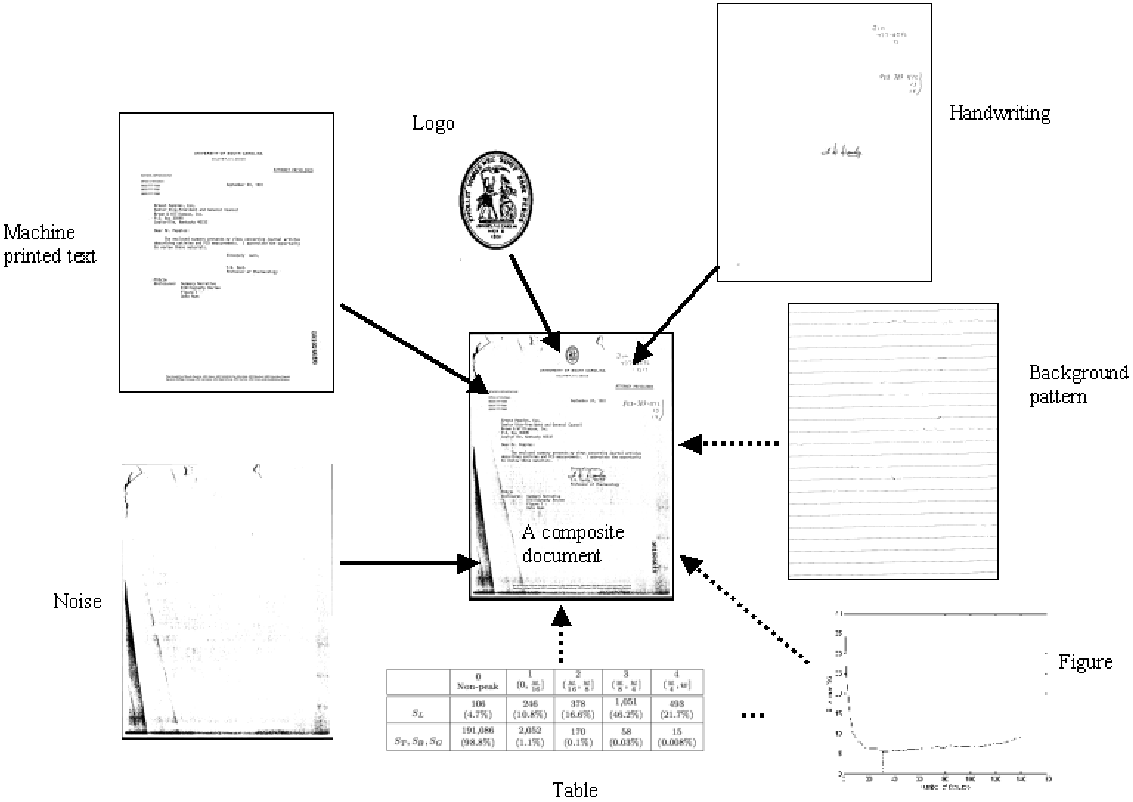
Quiz

- What is a document?
- What is an image?
- How can we *index and retrieve* document images?

- Why is document analysis difficult?

Page Layer Segmentation

- Document image generation model
  - A document consists many layers, such as handwriting, machine printed text, background patterns, tables, figures, noise, etc.

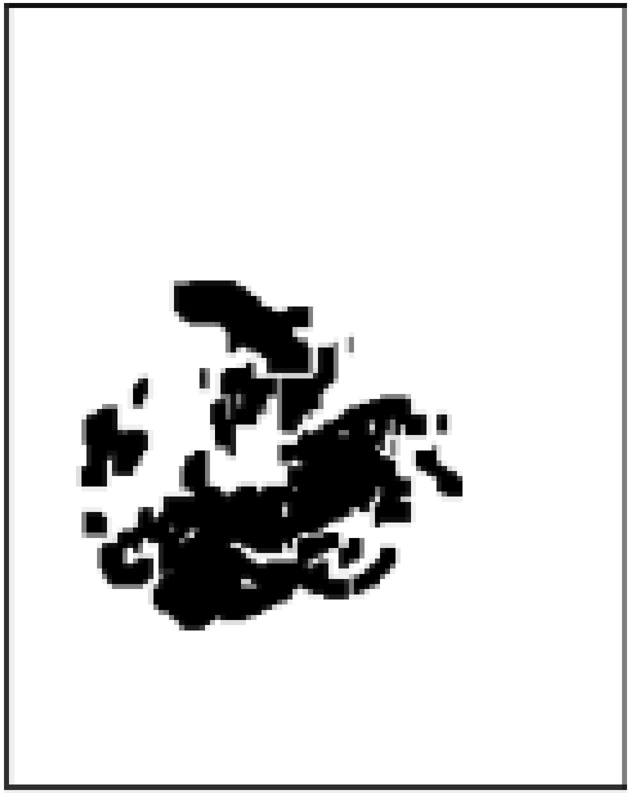


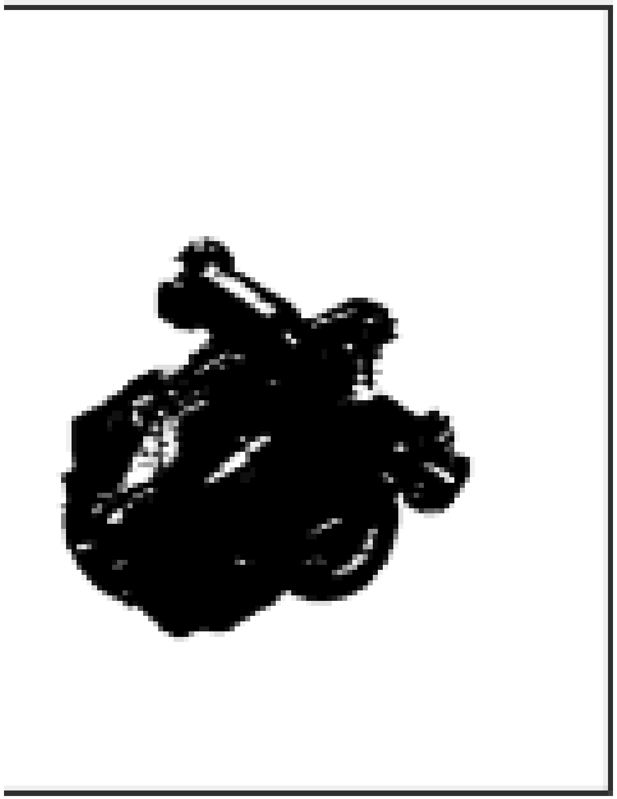
	0 Non-peak	1 $[0, \frac{w}{16}]$	2 $(\frac{w}{16}, \frac{w}{8}]$	3 $(\frac{w}{8}, \frac{w}{4}]$	4 $(\frac{w}{4}, w]$
S_L	106 (4.7%)	246 (10.8%)	378 (16.6%)	1,051 (46.2%)	493 (21.7%)
S_T, S_B, S_O	191,086 (98.8%)	2,052 (1.1%)	170 (0.1%)	58 (0.03%)	15 (0.008%)

Page Analysis

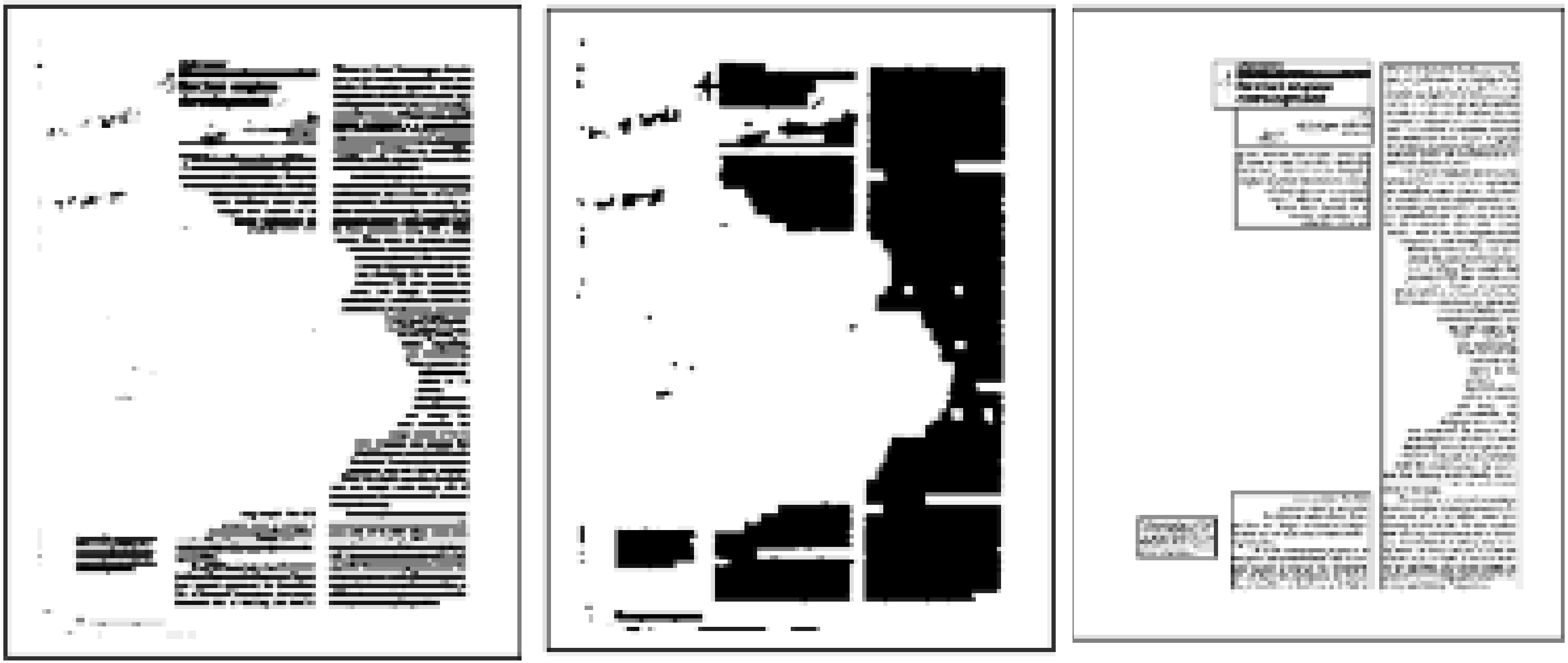
- Skew correction
  - Based on finding the primary orientation of lines
- Image and text region detection
  - Based on texture and dominant orientation
- Structural classification
  - Infer logical structure from physical layout
- Text region classification
  - Title, author, letterhead, signature block, etc.

Image Detection





Text Region Detection

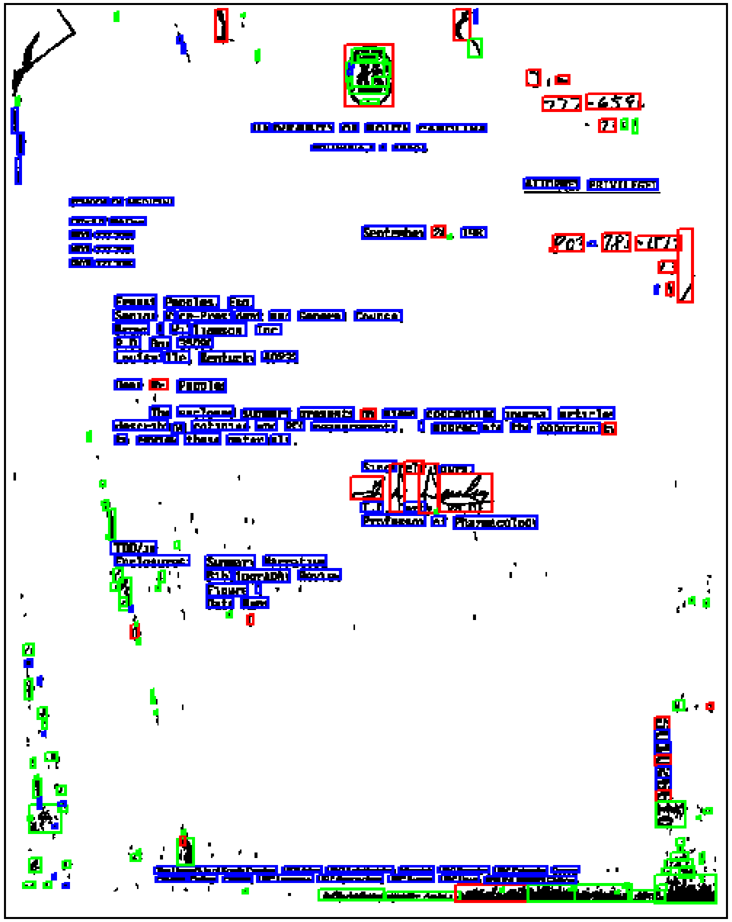


More Complex Example

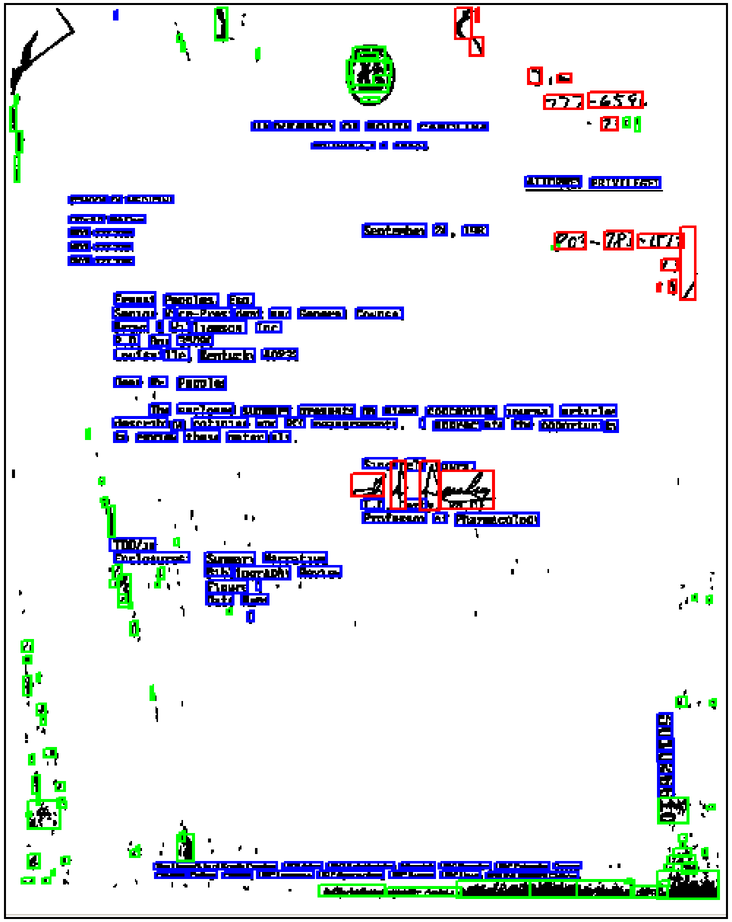
Printed text

Handwriting

Noise

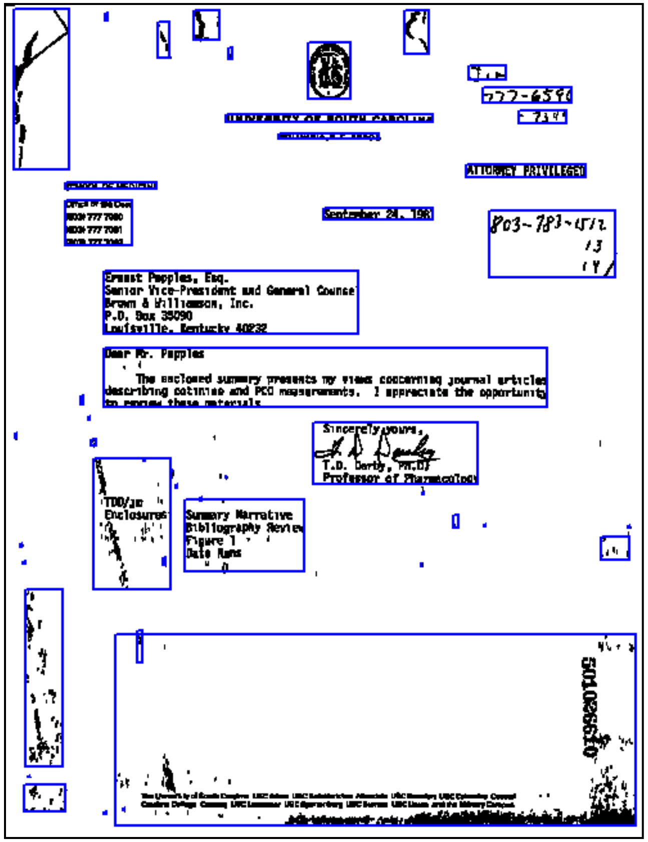


Before MRF-based postprocessing



After MRF-based postprocessing

Application to Page Segmentation



777-6590

7349

UNIVERSITY OF SOUTH CAROLINA

ATTORNEY PRIVILEGED

(803) 777 7080
(803) 777 7081
(803) 777 7082

September 24, 198

803-783-4572
13
14

Ernest Pepples, Esq.
Senior Vice-President and General Counsel
Brown & Williamson, Inc.
P.O. Box 35090
Louisville, Kentucky 40232

Dear Mr. Pepples

The enclosed summary presents my views concerning journal articles describing cotinine and PCO measurements. I appreciate the opportunity to review these materials.

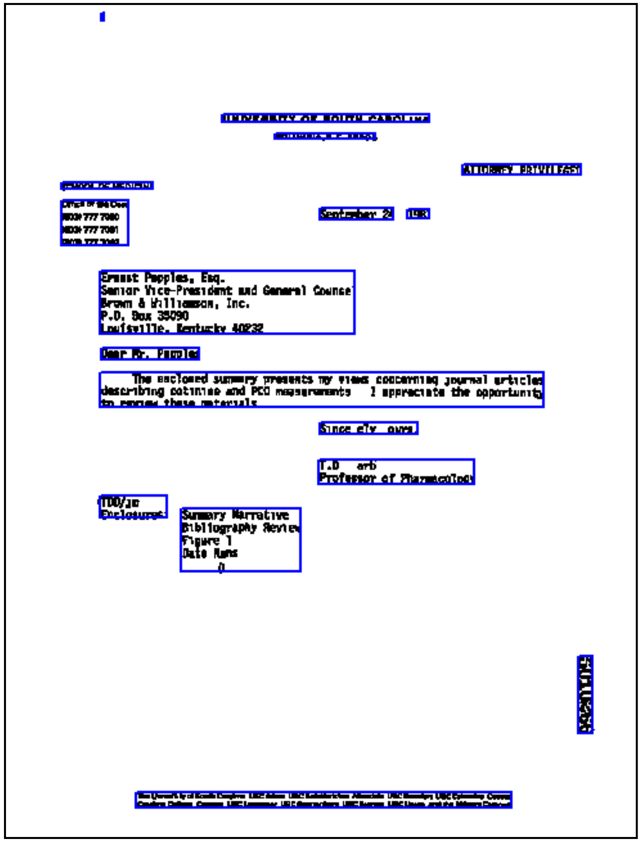
Sincerely yours,

T.D. Darby, Ph.D.
Professor of Pharmacology

TDD/jm
Enclosures: Summary Narrative
Bibliography Review
Figure 1
Data Runs

5010[illegible]610

Before enhancement



UNIVERSITY OF SOUTH CAROLINA

ATTORNEY PRIVILEGED

(803) 777 7080
(803) 777 7081
(803) 777 7082

September 24 198

Ernest Pepples, Esq.
Senior Vice-President and General Counsel
Brown & Williamson, Inc.
P.O. Box 35090
Louisville, Kentucky 40232

Dear Mr. Pepples

The enclosed summary presents my views concerning journal articles describing cotinine and PCO measurements. I appreciate the opportunity to review these materials.

Since ely ours

T.D arb
Professor of Pharmacology

TDD/jm
Enclosures: Summary Narrative
Bibliography Review
Figure 1
Data Runs

After enhancement

Language Identification

- Language-independent skew detection
  - Accommodate horizontal and vertical writing
- Script class recognition
  - Asian scripts have blocky characters
  - Connected scripts can’t be segmented easily
- Language identification
  - Shape statistics work well for western languages
  - Competing classifiers work for Asian languages

What about handwriting?

Optical Character Recognition

- Pattern-matching approach
  - Standard approach in commercial systems
  - Segment individual characters
  - Recognize using a neural network classifier
- Hidden Markov model approach
  - Experimental approach developed at BBN
  - Segment into sub-character slices
  - Limited lookahead to find best character choice
  - Useful for connected scripts (e.g., Arabic)

Quiz

- What is a document?
- What is an image?
- How can we *index and retrieve* document images?
- Why is document analysis difficult?

- Is the (Doc Image IR) problem solved? Why or Why not?

OCR Accuracy Problems

- Character segmentation errors
  - In English, segmentation often changes “m” to “rn”
- Character confusion
  - Characters with similar shapes often confounded
- OCR on copies is much worse than on originals
  - Pixel bloom, character splitting, binding bend
- Uncommon fonts can cause problems
  - If not used to train a neural network

Improving OCR Accuracy

- Image preprocessing
  - Mathematical morphology for bloom and splitting
  - Particularly important for degraded images
- “Voting” between several OCR engines helps
  - Individual systems depend on specific training data
- Linguistic analysis can correct some errors
  - Use confusion statistics, word lists, syntax, …
  - But more harmful errors might be introduced

OCR Speed

- Neural networks take about 10 seconds a page
  - Hidden Markov models are slower
- Voting can improve accuracy
  - But at a substantial speed penalty
- Easy to speed things up with several machines
  - For example, by batch processing - using desktop computers at night

Problem: Logical Page Analysis (Reading Order)

- Can be hard to guess in some cases
  - Newspaper columns, figure captions, appendices, …
- Sometimes there are explicit guides
  - “Continued on page 4” (but page 4 may be big!)
- Structural cues can help
  - Column 1 might continue to column 2
- Content analysis is also useful
  - Word co-occurrence statistics, syntax analysis

Processing Converted Text

Typical Document Image Indexing

- Convert hardcopy to an “electronic” document
  - OCR
  - Page Layout Analysis
  - Graphics Recognition
- Use structure to add metadata
- Manually supplement with keywords

Use traditional text indexing and retrieval techniques?

Information Retrieval on OCR

- Requires robust ways of indexing
- Statistical methods with large documents work best
- Key Evaluations
  - Success for high quality OCR (Croft et al 1994, Taghva 1994)
  - Limited success for poor quality OCR (1996 TREC, UNLV)

N-Grams

- Powerful, Inexpensive statistical method for characterizing populations
- Approach
  - Split up document into n-character pairs fails
  - Use traditional indexing representations to perform analysis
  - “DOCUMENT” -> DOC, OCU, CUM, UME, MEN, ENT
- Advantages
  - Statistically robust to small numbers of errors
  - Rapid indexing and retrieval
  - Works from 70%-85% character accuracy where traditional IR fails

Matching with OCR Errors

- Above 80% character accuracy, use words
  - With linguistic correction
- Between 75% and 80%, use n-grams
  - With n somewhat shorter than usual
  - And perhaps with character confusion statistics
- Below 75%, use word-length shape codes

Handwriting Recognition

- With stroke information, can be automated
  - Basis for input pads
- Simple things can be read without strokes
  - Postal addresses, filled-in forms
- Free text requires human interpretation
  - But repeated recognition is then possible

Outline

- Processing Converted Text
- Manipulating Images of Text
  - Title Extraction
  - Named Entity Extraction
  - Keyword Spotting
  - Abstracting and Summarization
- Indexing based on Structure
- Graphics and Drawings
- Related Work and Applications

Processing Images of Text

- Characteristics
  - Does not require expensive OCR/Conversion
  - Applicable to filtering applications
  - May be more robust to noise

- Possible Disadvantages
  - Application domain may be very limited
  - Processing time may be an issue if indexing is otherwise required

Proper Noun Detection
(DeSilva and Hull, 1994)

- Problem: Filter proper nouns in images of text
  - People, Places, Things
- Advantages of the Image Domain:
  - Saves converting all of the text
  - Allows application of word recognition approaches
  - Limits post-processing to a subset of words
  - Able to use features which are not available in the text
- Approach:
  - Identify Word Features
    - Capitalization, location, length, and syntactic categories
  - Classify using rule-set
  - Achieve 75-85% accuracy without conversion

Keyword Spotting

Techniques:

- Work Shape/HMM - (Chen et al, 1995)
- Word Image Matching - (Trenkle and Vogt, 1993; Hull et al)
- Character Stroke Features - (Decurtins and Chen, 1995)
- Shape Coding - (Tanaka and Torii; Spitz 1995; Kia, 1996)

Applications:

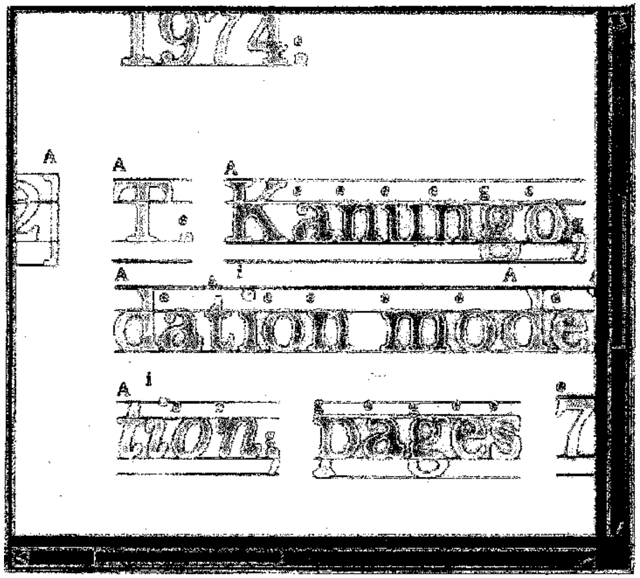
- Filing System (Spitz - SPAM, 1996)
- Numerous IR
- Processing handwritten documents

Formal Evaluation :

- Scribble vs. OCR (DeCurtins, SDIUT 1997)

Shape Coding

- Approach
  - Use of Generic Character Descriptors
  - Make Use of Power of Language to resolve ambiguity
  - Map Character based on Shape features including ascenders, descenders, punctuation and character with holes



Shape Codes

- Group all characters that have similar shapes
  - {A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, 2, 3, 4, 5, 6, 7, 8, 9, 0}
  - {a, c, e, n, o, r, s, u, v, x, z}
  - {b, d, h, k, }
  - {f, t}
  - {g, p, q, y}
  - {i, j, l, 1}
  - {m, w}

Why Use Shape Codes?

- Can recognize shapes faster than characters
  - Seconds per page, and very accurate

- Preserves recall, but with lower precision
  - Useful as a first pass in any system

- Easily extracted from JPEG-2 images
  - Because JPEG-2 uses object-based compression

Evaluation

- The usual approach: Model-based evaluation
  - Apply confusion statistics to an existing collection
- A bit better: Print-scan evaluation
  - Scanning is slow, but availability is no problem
- Best: Scan-only evaluation
  - Few existing IR collections have printed materials

Summary

- Many applications benefit from image based indexing
  - Less discriminatory features
  - Features may therefore be easier to compute
  - More robust to noise
  - Often computationally more efficient
- Many classical IR techniques have application for DIR
- Structure as well as content are important for indexing
- Preservation of structure is essential for in-depth understanding

Closing thoughts….

- What else is useful?
  - Document Metadata? – Logos? Signatures?

- Where is research heading?
  - Cameras to capture Documents?

- What massive collections are out there?
  - Tobacco Litigation Documents
    - 49 million page images
  - Google Books
  - Other Digital Libraries

Additional Reading

- A. Balasubramanian, et al. Retrieval from Document Image Collections, *Document Analysis Systems VII*, pages 1-12, 2006.

- D. Doermann. The Indexing and Retrieval of Document Images: A Survey. *Computer Vision and Image Understanding*, 70(3), pages 287-298, 1998.

Some Applications

- Legacy Tobacco Documents Library
  - http://legacy.library.ucsf.edu/

- Google Books
  - http://books.google.com/

- George Washington’s Papers
  - http://ciir.cs.umass.edu/irdemo/hw-demo/