



<http://www.flickr.com/photos/56685562@N00/565216/>

Document Image Retrieval

David Kauchak

cs160

Fall 2009

adapted from:

David Doermann

<http://terpconnect.umd.edu/~oard/teaching/796/spring04/slides/11/796s0411.ppt>



Assign 4 writeups

- Overall, I was very happy
- See how big a difference the modifications make!
- Some general comments
 - explain data set and characteristics
 - explain your evaluation measure(s)
 - think about the points you're trying to make, then use the data to make that point
 - comment on anything abnormal or surprising in the data
 - dig deeper if you need to
 - if you have multiple evaluation measures, use them to explain/ understand different behavior
 - try and explain why you got the results you obtained



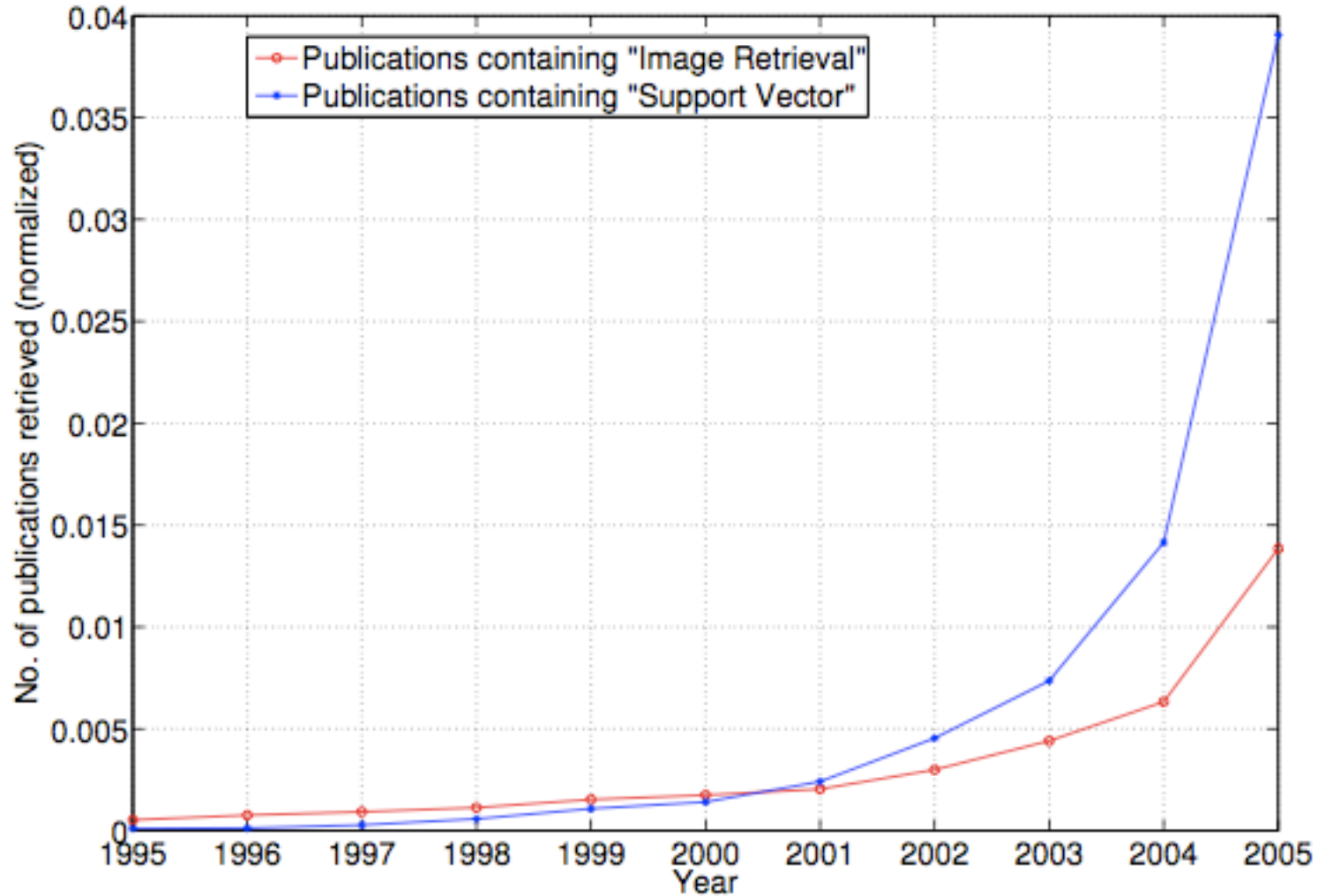
Information retrieval systems

- Spend 15 minutes playing with three different image retrieval systems
 - http://en.wikipedia.org/wiki/Image_retrieval has a number
 - What works well?
 - What doesn't work well?
 - Anything interesting you noticed?
- You won't hand anything in, but we'll start class on Monday with a discussion of the systems



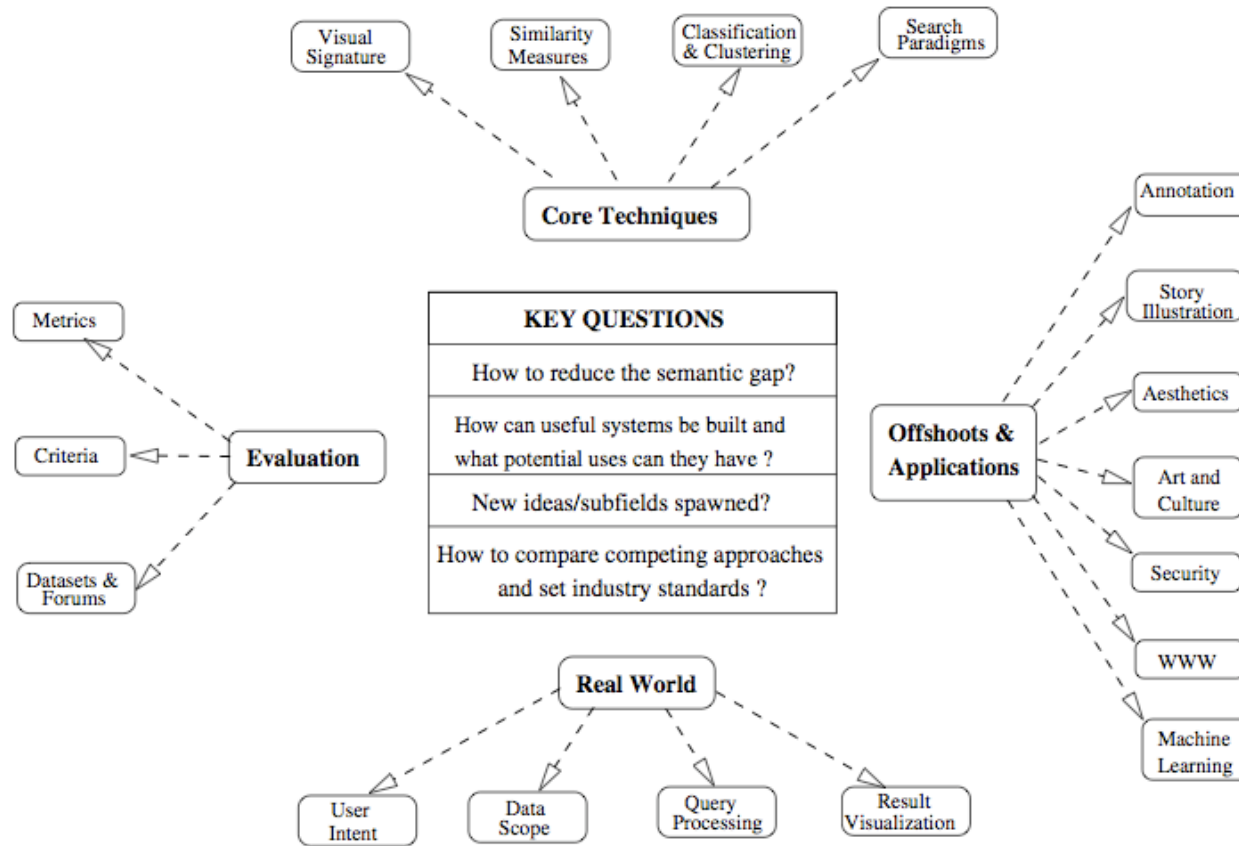
Image Retrieval

Plot of (normalized) trends in publication over the last 10 years as indexed by Google Scholar



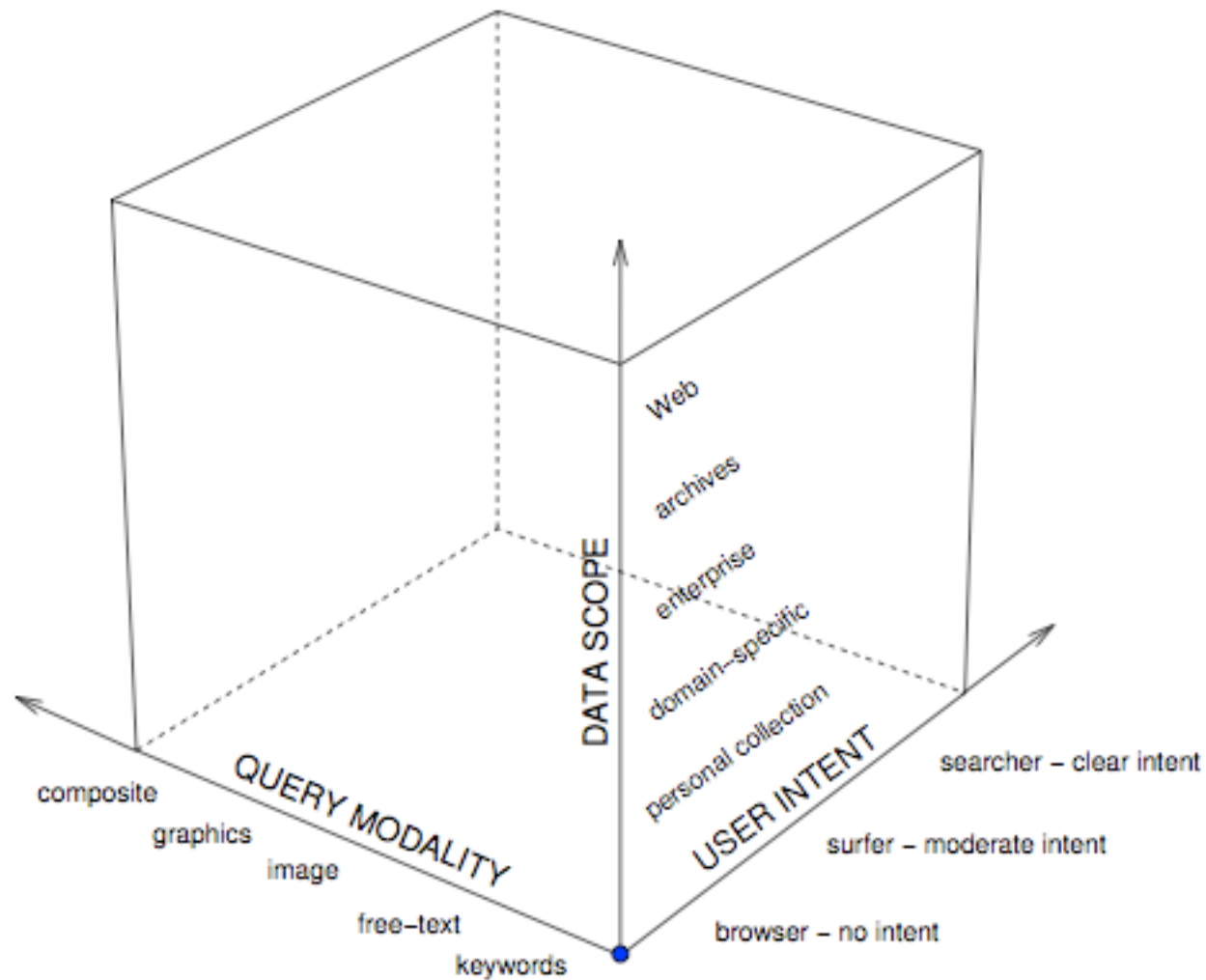
http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Image Retrieval Problems



http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Different Systems



http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta_TR.pdf

Information retrieval: data

amount of data

data characteristics

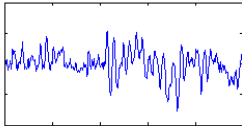
Text retrieval



trillions of web pages
within an order of
magnitude in “private” data

- user generated
- some semi-structured
- link structure

Audio retrieval



order of a few billion?
last fm has 150M songs

- mostly professionally generated
- co-occurrence statistics

Image retrieval



somewhere in between

- user generated
- becoming more prevalent
- some tagging
- incorporated into web pages (context)



Information retrieval: challenges

challenges

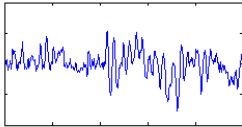
other dimensions?

Text retrieval



- scale
- ambiguity of language
- link structure
- spam

Audio retrieval



- query language
- user interface
- features/pre-processing

Image retrieval



- query language
- user interface
- features/pre-processing
- ambiguity of pictures



What's in a document?

- I give you a file I downloaded
- You know it has text in it
- What are the challenges in determining what characters are in the document?
 - File format:

1. What file types are returned in a Google search?

There are 13 main file types searched by Google in addition to standard web formatted Microsoft Office formats:


- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)
- Lotus WordPro (lwp)
- MacWrite (mw)
- Microsoft Excel (xls)
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks, wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Shockwave Flash (swf)
- Text (ans, txt)



http://www.google.com/help/faq_filetypes.html

What is a document?

Assigned to the 3842 Regiment of Infantry U.S. Army.

STATE OF  TOWN OF

Missouri Saint-Louis

I, William Cathey, born in Independence in the State of Missouri, aged twenty-two years, and by occupation a Cook Do HEREBY ACKNOWLEDGE to have voluntarily enlisted this fifteenth day of November 1865, as a **Soldier** in the Army of the United States of America, for the period of **THREE YEARS**, unless sooner discharged by proper authority: Do also agree to accept such bounty, pay, rations, and clothing, as are, or may be, established by law. And I, William Cathey, do solemnly swear, that I will bear true faith and allegiance to the **United States of America**, and that I will serve them honestly and faithfully against all their enemies or opposers whomsoever; and that I will observe and obey the orders of the President of the United States, and the orders of the officers appointed over me, according to the Rules and Articles of War.

Sworn and subscribed to, at St. Louis, Mo. this 15th day of November 1865. William Cathey

BEFORE Henry Johnson Major 18th Reg't Inf'y.

W. M. Powers
Act. Capt. Surg. U.S.A.
EXAMINING SURGEON.

I CERTIFY, ON HONOR, That I have carefully examined the above named recruit, agreeably to the General Regulations of the Army, and that in my opinion he is free from all bodily defects and mental infirmity, which would, in any way, disqualify him from performing the duties of a soldier.

I CERTIFY, ON HONOR, That I have minutely inspected the Recruit, William Cathey previously to his enlistment, and that he was entirely sober when enlisted; that, to the best of my judgment and belief, he is of lawful age; and that, in accepting him as duly qualified to perform the duties of an able-bodied soldier, I have strictly observed the Regulations which govern the recruiting service. This soldier has black eyes, black hair, black complexion, is 5 feet 9 inches high.

Henry Johnson
Major 18th Reg't Inf'y.
RECRUITING OFFICER.

14. G. O. No. 721

Superintendent.

The Washington Post

Friday, July 9, 2004

Price: 50¢

DISTRICT OF COLUMBIA

U.S. MAIL PERMIT NO. 1000 WASHINGTON, D.C. 20001

BEING A BLACK MAN



The 11-year-old boy, Marcus, is shown in a classroom at the St. Albans School in Washington, D.C. He is being held by his father, Marcus, and his mother, Hilma.

Marcus is shown in a classroom at the St. Albans School in Washington, D.C. He is being held by his father, Marcus, and his mother, Hilma.

After Zargawi, No Clear Path In Weary Iraq

Difficult Questions Surround Legacy of Insurgent Leader

By Peter Krause

It is not clear what the end of the war will mean for Iraq. A legacy of insurgency remains, and the path forward is unclear. The legacy of an insurgent leader, Zargawi, is a difficult one to navigate. The path forward is unclear, and the legacy of an insurgent leader, Zargawi, is a difficult one to navigate.

The Young Apprentice

Marcus's Parents Argue Over How to Protect — and Prepare — Him

Story by Susan E. Smart (Photos by Michael Grecco / The Washington Post)



Marcus is shown in a classroom at the St. Albans School in Washington, D.C. He is being held by his father, Marcus, and his mother, Hilma.

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

A U.S. soldier is shown in a classroom at the St. Albans School in Washington, D.C. He is being held by his father, Marcus, and his mother, Hilma.

A U.S. soldier is shown in a classroom at the St. Albans School in Washington, D.C. He is being held by his father, Marcus, and his mother, Hilma.

Kaine Delays Execution of Inmate for 6 Months

Inquiry Into Killer's Mental State Overlooked

By Thomas H. Dyer

Gov. Tim Wainwright has delayed the execution of a man convicted of murdering a woman for six months. The delay is due to a mental health inquiry that was overlooked.

INSIDE

Paul's Personal Mission

Paul's personal mission is to help the poor and the sick. He is a man of faith and a man of action.

Gene's Drive to the Max

Gene's drive to the max is to help the poor and the sick. He is a man of faith and a man of action.

14. G. O. No. 721

U.S. MAIL PERMIT NO. 1000 WASHINGTON, D.C. 20001

Document Images

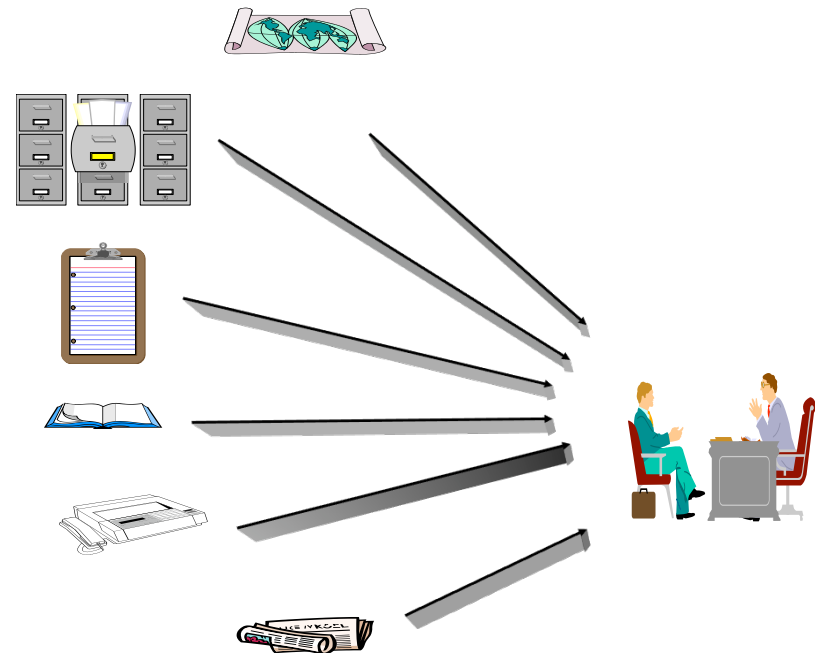
- A document image is a document that is represented as an image, rather than some predefined format
- Like normal images, contain pixels
 - often binary-valued (black, white)
 - But greyscale or color sometimes
- 300 dots per inch (dpi) gives the best results
 - But images are quite large (1 MB per page)
 - Faxes are normally 72 dpi
- Usually stored in TIFF or PDF format

Want to be able to process them like text files



Sources of document images

- Web
 - <http://dli.iiit.ac.in/>
 - Arabic news stories are often GIF images
 - Google Books, Project Gutenberg (though these are a bit different)
- Library archives
- Other
 - Tobacco Litigation Documents
 - 49 million page images



ST-1 is an extremely sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.

ST-1 is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.

ST-1 is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.

Handwritten notes:
 - **ST-1** is an extremely sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.
 - **ST-1** is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.
 - **ST-1** is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.

Other notes:
 - **ST-1** is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.
 - **ST-1** is a highly sensitive, highly specific, and highly accurate assay for the detection of **ST-1** in a wide range of samples.

ORDERER	MAGAZINE	ADVERTISING PERIOD	DATE
MAIL BOX (BESTIC PRICE)	ARABIAS	NEWSPAPER	10/20/89
BEACH (BESTIC PRICE)	ARABIAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE (BESTIC PRICE)	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (BESTIC PRICE)	INDIANA	NEWSPAPER	10/20/89
MAIL BOX	INDIANA MARKETS	OH	12/89-1/90
RED PAGES	ARKANSAS MARKETS	OH	1/90
BEACH	PORTLAND	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & DOUGIE	NEWSPAPER	10/20/89
BEACH	INDONESIA	SUNDAY SUPPLEMENT	11/25/89
GULF CITY HEAD-OS	INDONESIA MARKETS	OH	1/89-12/89
GULF CITY HEAD-OS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/24/89
"CORVETTE"	ARABASIA MARKETS	SUNDAY SUPPLEMENT	11/24/89
"BEACH - LOW PRICE"	ARABASIA	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

O.P./NO OCTOBER 18, 1989

Stadt- und Universitätsbibliothek
 Hochschulbibliothek, L34-128
 6900 Frankfurt am Main 1
 Telefon: 089-232-01
 Telefax: 089-232-2064

Satz: geschlo. Oasen und Herren.
 am 12. Juli 1983 in der garten ab 8. Juli 1983 in garten
 Deutschland neue (BUNDESBIBLIOTHEK)
 Die Bibliothek der Stadt- und Universitätsbibliothek Frankfurt am Main

A a

Association (Association)
 1. A group of people who are organized together to achieve a common purpose or goal.
 2. A group of people who are organized together to represent a common interest or profession.

Association (Association)
 1. A group of people who are organized together to achieve a common purpose or goal.
 2. A group of people who are organized together to represent a common interest or profession.

Association (Association)
 1. A group of people who are organized together to achieve a common purpose or goal.
 2. A group of people who are organized together to represent a common interest or profession.

Association (Association)
 1. A group of people who are organized together to achieve a common purpose or goal.
 2. A group of people who are organized together to represent a common interest or profession.

ORDERER	MAGAZINE	ADVERTISING PERIOD	DATE
MAIL BOX (BESTIC PRICE)	ARABIAS	NEWSPAPER	10/20/89
BEACH (BESTIC PRICE)	ARABIAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE (BESTIC PRICE)	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (BESTIC PRICE)	INDIANA	NEWSPAPER	10/20/89
MAIL BOX	INDIANA MARKETS	OH	12/89-1/90
RED PAGES	ARKANSAS MARKETS	OH	1/90
BEACH	PORTLAND	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & DOUGIE	NEWSPAPER	10/20/89
BEACH	INDONESIA	SUNDAY SUPPLEMENT	11/25/89
GULF CITY HEAD-OS	INDONESIA MARKETS	OH	1/89-12/89
GULF CITY HEAD-OS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/24/89
"CORVETTE"	ARABASIA MARKETS	SUNDAY SUPPLEMENT	11/24/89
"BEACH - LOW PRICE"	ARABASIA	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

The Washington Post

After Zargawi, No Clear Path In Weary Iraq
 Difficult Questions Remain Legacy of Inaugural Leader

The Young Apprentice
 Mother's Parents Argue Over How to Protect - and Prepare - Him

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

Keine Delays Execution Of Inmate for 6 Months

FDA Approves Vaccine That Should Prevent Most Cervical Cancers

DEUTSCHES BIBLIOTHEKSGESAMTVEREIN (DBV)

Druck: J. K. S. 0,070 1,10

AUFTRAGSABGABE UND LIEFERUNG

Zeitschriftendienstleistungen
 Lieferung und Druck: 0,070 1,10
 Werbung: 0,070 1,10

Bestellnummer: 0,070 1,10

THE WORLD'S RICHEST PEOPLE SPECIAL ISSUE

Forbes BILLIONAIRES

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

946 BILLIONAIRES

MEXICO'S RICHEST MAN CLOSES IN ON U.S. WARREN BUFFETT

INDIA UPDATES JAPAN

Journal of Neural Protection

FLAVIO BRITTORE

INDIA UPDATES JAPAN

MEXICO'S RICHEST MAN CLOSES IN ON U.S. WARREN BUFFETT

946 BILLIONAIRES

FINANCIAL TIMES

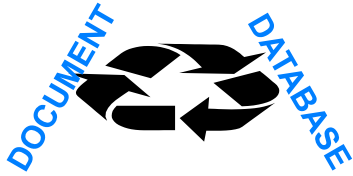
Collateral damage

Iraq braced as Saddam rejects Bush ultimatum

Fed holds rates but signals uncertain future

US probes Abold collusion claims

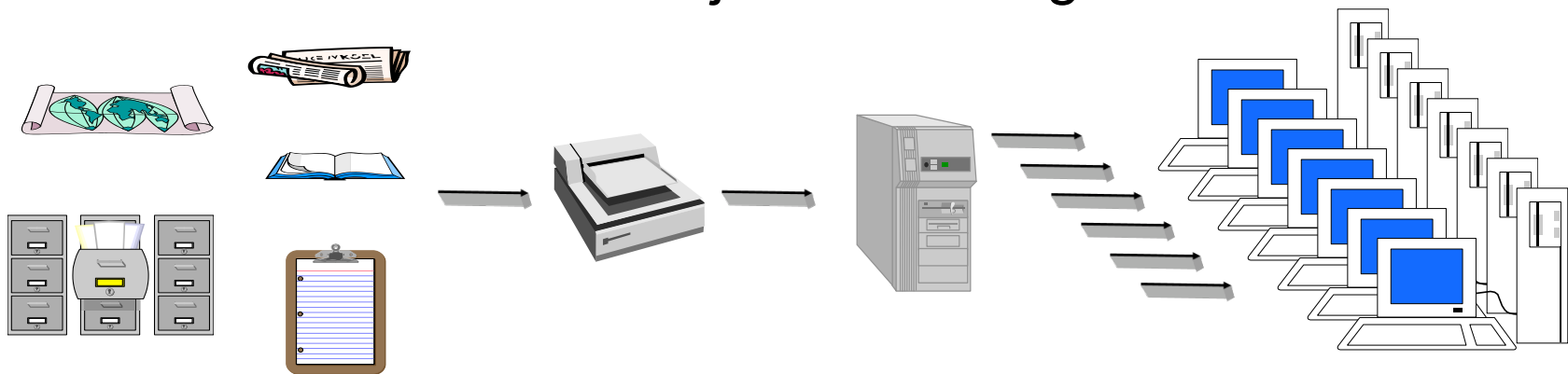
Chopard



Document Image Database

IMAGE

- Collection of scanned images
- Need to be available for indexing and retrieval, abstracting, routing, editing, dissemination, interpretation
- NOTE: more needs than just searching!



What are the challenges?

What are the sub-problems?



Document images

- So far, we've only been interested in documents as strings of text
- Document images introduce contain additional information
 - embedded images
 - formatting
 - handwritten annotations
 - figures/diagrams/tables
 - classes of documents
 - memo
 - newspaper article
 - book page



Challenges

- They're an image 😊
- Quality
 - scan orientation
 - noise
 - contrast
- Hand-written text
- Hand-written diagrams



Sub-problems

- Classification - what type of document image is this?
- Page segmentation
 - structure
 - identify images
 - identify text
 - identify handwritten text
 - diagram identification
- Meta-data identification
 - title, author
 - language
- OCR
- Reading ordering
- Indexing



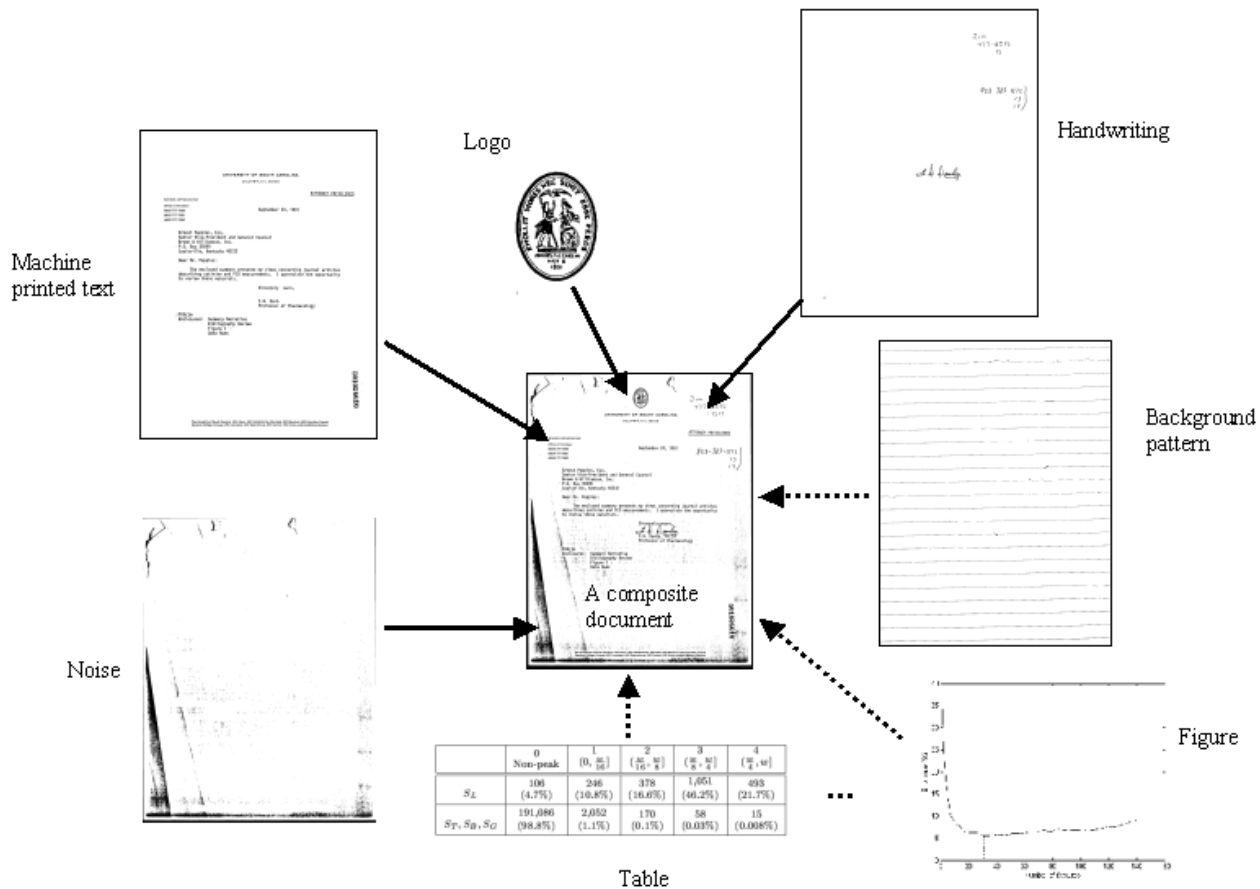
Problems we'll discuss today...

- Preprocessing issues
 - Page Layer Segmentation
 - OCR
 - Reading order
- IR issues

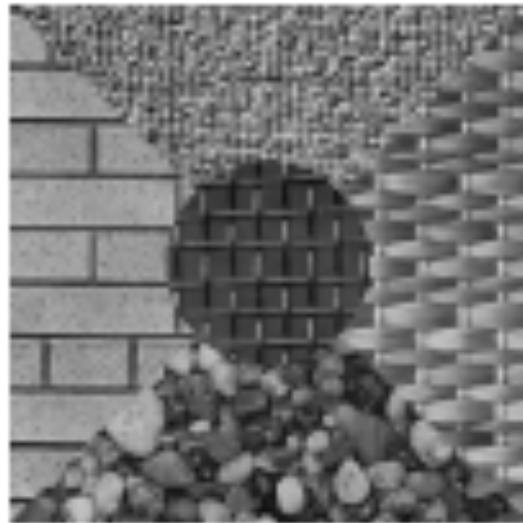


Problem: Page Layer Segmentation

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



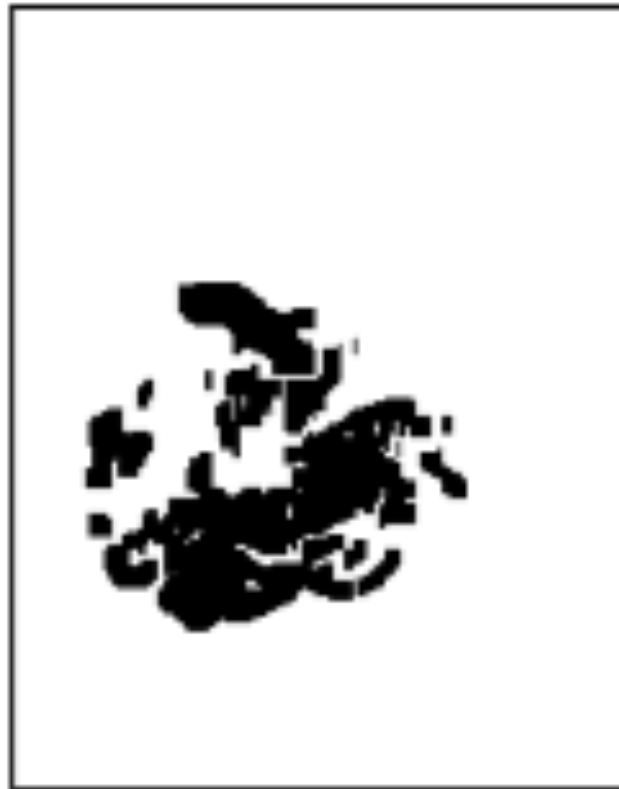
Step 1 - segmentation



Segmentation



1



2

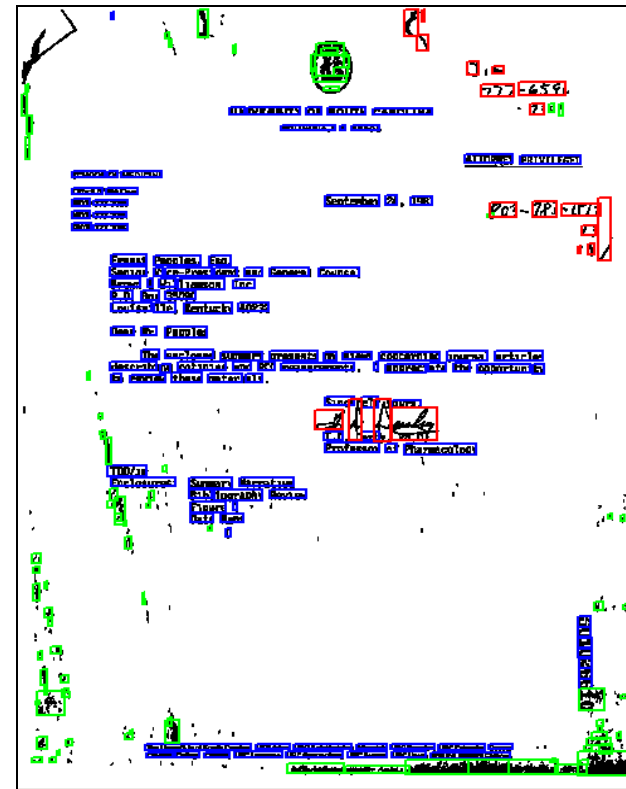
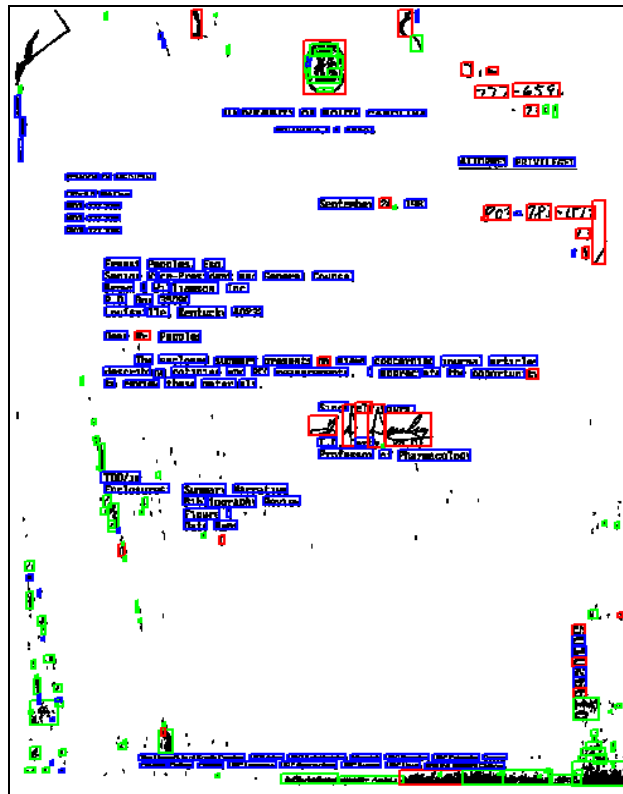


Segmentation



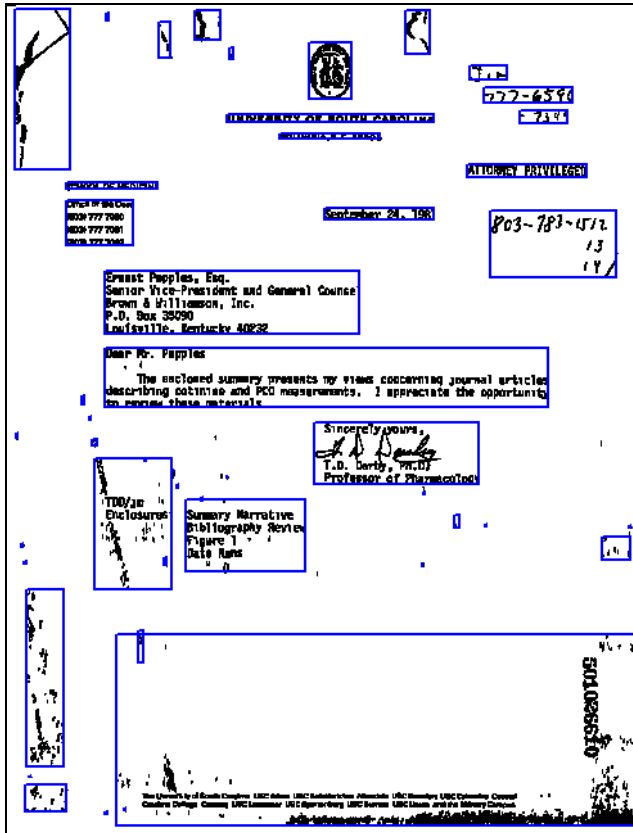
Step 2 – classify the segments

Printed text
Handwriting
Noise

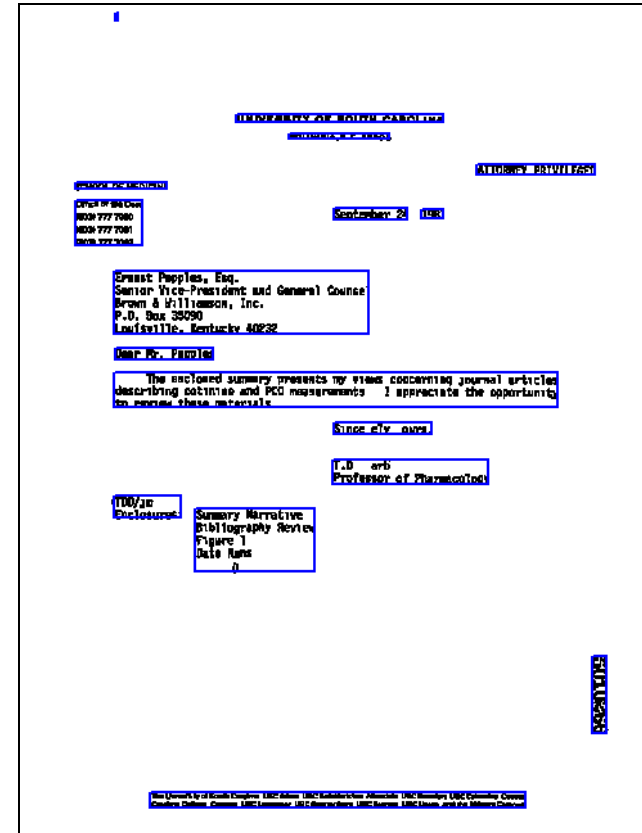


We can use features of the “segment” as well as positional information about the other segments

Segmentation Classification



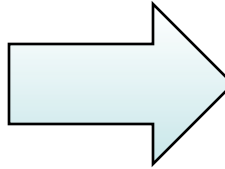
Before enhancement



After enhancement

Problem: OCR

- One of the more successful applications of computer vision

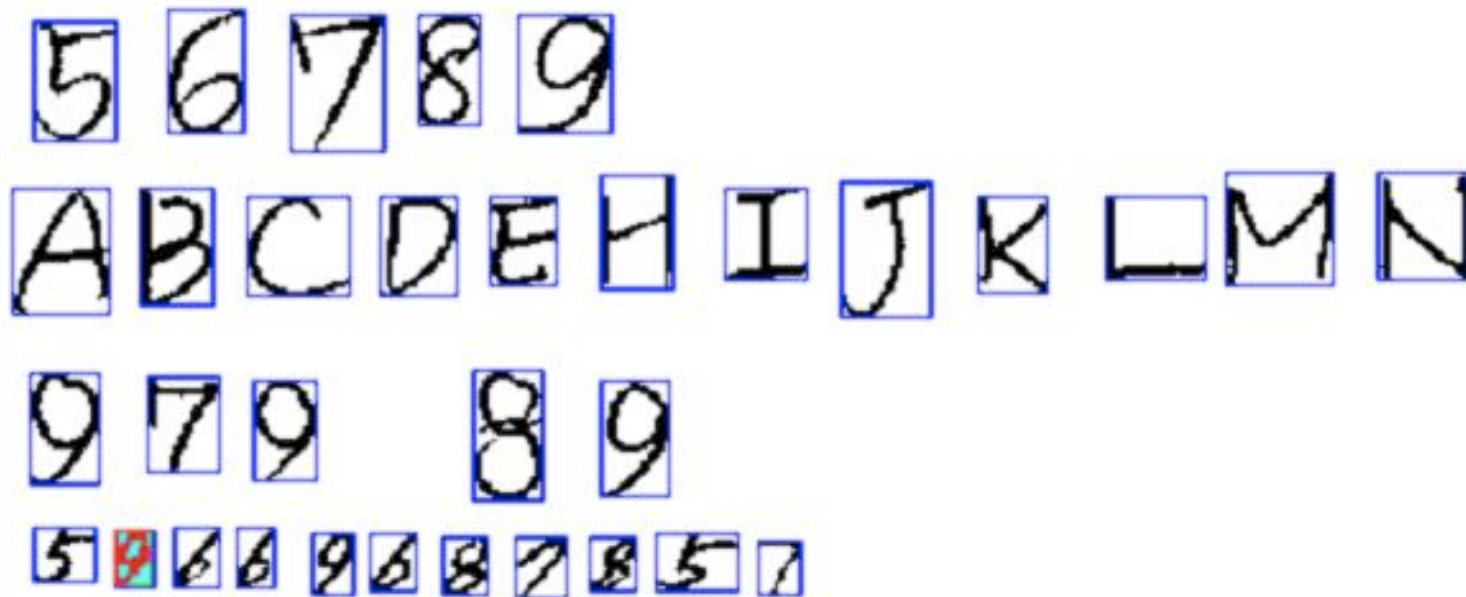


How does this happen?



OCR: One solution

- Pattern-matching approach
 - Standard approach in commercial systems
 - Segment individual characters
 - Recognize using a neural network classifier



OCR

एक बार एक सज्जन आश्रम आये थे, उन्होंने बातचीत के सिलसिले में श्री श्री ठाकुर से कहा कि मुझ पर कृपा कीजियेगा इस पर श्री श्री ठाकुर ने कहा : 'कृपा' में पहला अक्षर 'कृ' है तब 'पा' यानी करने से पाओगे ।"
पुनः उस सज्जन ने कहा, सुनते हैं कि दुर्गापाठ करने

مرحبا بكم على موقع مجلس ايسلنجتون – هو افضل موقع للحصول على معلومات عن الزيارة والعيش والعمل في ايسلنجتون. يمكنكم ايضا معرفة اين تقع اقرب صالة سينما وكيفية دفع ضريبة المجلس. اقرؤا عن الخدمات التي يقدمها المجلس للراشدين والأطفال و اقرؤا ايضا عن الديمقراطية في ايسلنجتون وكيف يمكنكم اعطاء آرائكم بخصوص قرارات المجلس.

Ideas?

Optical Character Recognition

- Hidden Markov model approach
 - Experimental approach developed at BBN
 - Segment into sub-character slices
 - Limited lookahead to find best character choice

Determining character segmentation is difficult!



- Uniform slices
- View as a sequential prediction problem



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

Challenge with OCR is there is a often a trade-off between speed and accuracy

- 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: Reading Order

What is the sequence of words from this document?

The Washington Post
DISTRICT FINAL
Friday, June 9, 2008
Washington Post
Circulation: 444,000
Subscription: \$10.00 per month
Retail: \$0.50 per copy
Masthead: The Washington Post
Motto: Democracy Dies in Darkness

BEING A BLACK MAN

The world's fastest-growing city, Atlanta, is a starkly different place where freedom has long been a reality. The city is now a mix of black and white, and the city's future is uncertain. The city's future is uncertain. The city's future is uncertain.

After Zarqawi, No Clear Path In Weary Iraq

Difficult Questions Surround Legacy of Insurgent Leader

By David E. Sanger
Iraq's most notorious leader, Abu Musab al-Zarqawi, died in a U.S. military airstrike Sunday, but his death has not ended the conflict. The U.S. military and its coalition partners are still engaged in a struggle to bring stability to the war-torn country. The U.S. military and its coalition partners are still engaged in a struggle to bring stability to the war-torn country.

The Young Apprentice

Marcus's Parents Argonize Over How to Protect — and Prepare — Him

By Ronan Farrow
Marcus's parents are torn over how to protect their son from the dangers of the world. Marcus's parents are torn over how to protect their son from the dangers of the world.

At the age of 10, Marcus is a bright, curious child. His parents are torn over how to protect him from the dangers of the world. Marcus's parents are torn over how to protect him from the dangers of the world.

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

By David E. Sanger
U.S. forces in Iraq have found a cache of weapons and explosives, including a cache of weapons and explosives. U.S. forces in Iraq have found a cache of weapons and explosives, including a cache of weapons and explosives.

Kaine Delays Execution Of Inmate for 6 Months

Inquiry Into Killer's Mental State Ordered

By Catherine Reynolds
Gov. Mitt Romney has ordered a review of the execution of a man convicted of a heinous crime. Gov. Mitt Romney has ordered a review of the execution of a man convicted of a heinous crime.

FDA Approves Vaccine That Should Prevent Most Cervical Cancers

By David E. Sanger
The FDA has approved a new vaccine that could prevent most cervical cancers. The FDA has approved a new vaccine that could prevent most cervical cancers.

INSIDE

- Senate Panel Report**
The Senate panel report on the Iraq war is expected to be released soon. The Senate panel report on the Iraq war is expected to be released soon.
- Obama's Health Plan**
President Obama's health care plan is facing opposition in Congress. President Obama's health care plan is facing opposition in Congress.
- Obama's Health Plan**
President Obama's health care plan is facing opposition in Congress. President Obama's health care plan is facing opposition in Congress.

© 2008 The Washington Post Company

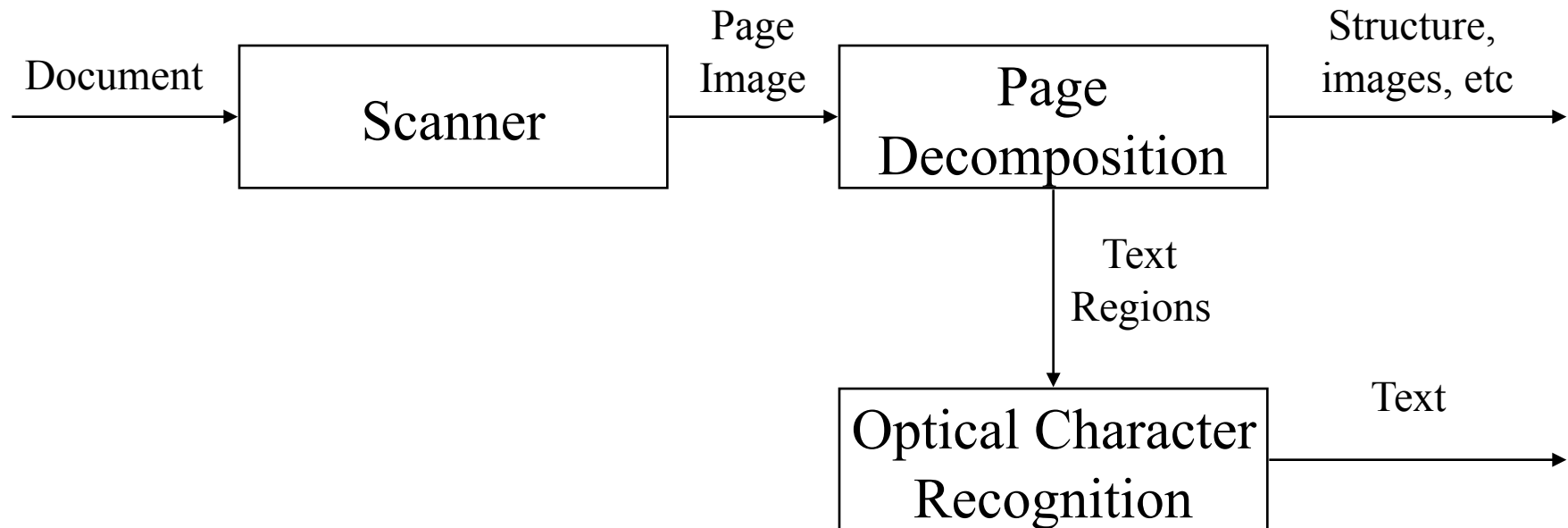
Ideas?

Logical Page Analysis

- 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



Traditional Approach



Remember our goal

- Create an IR system over image documents
- Challenge: OCR is not perfect
 - Success for high quality OCR (Croft et al 1994, Taghva 1994)
 - Limited success for poor quality OCR (1996 TREC, UNLV)

Ideas?

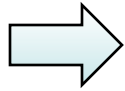


Proposed Solutions

- Improve OCR 😊
- Again, speed is always a concern
- Similar to spelling correction
 - Automatic Correction
 - Characters N-grams
 - Statistically robust to small numbers of errors
 - Rapid indexing and retrieval
 - Works from 70%-85% character accuracy where traditional IR fails

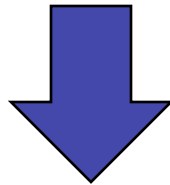


Matching with OCR errors



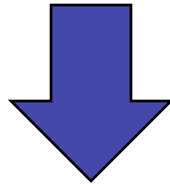
5

with confidence *X%*



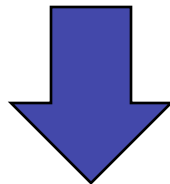
> 80%

Keep base system answer



75% - 80%

Character n-grams



<75%

More intensive image techniques
(e.g. shape codes)



Conversion to Text?

- Full Conversion often required
- Conversion is difficult!
 - Noisy data
 - Complex Layouts
 - Non-text components

Points to Ponder

- Do we really need to convert?
- Can we expect to fully describe documents without assumptions?



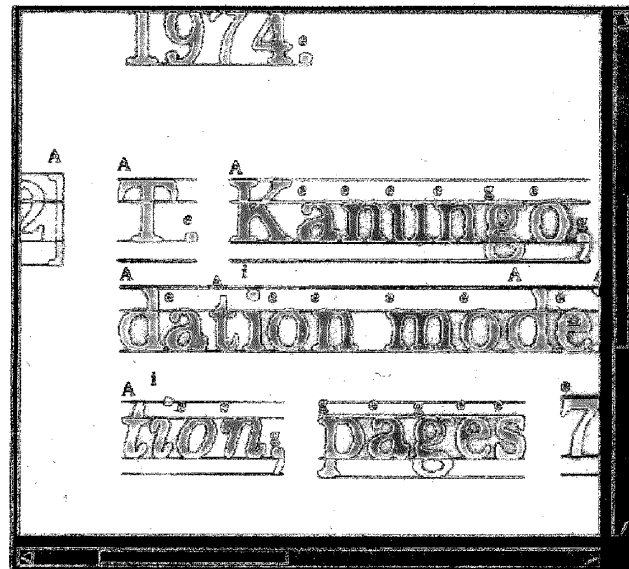
Idea: do processing on images

- 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
 - Indexing?



Shape Coding

- Approach
 - Use of Generic Character Descriptors
 - Map Character based on Shape features including ascenders, descenders, punctuation and character with holes



Shape Codes

- Group all characters that have similar shapes
 - {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, l}
 - {m, w}
- Shape codes whether a subset of an image belongs to a given character set
- Sub-process later based on linguistic and/or OCR



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?
 - 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.



Fun Stuff

- <http://www.sr.se/P1/src/sing/#>

