# Document Image Analysis

State of the Art
and
Technology Roadmap

March, 2008

Eric Saund
Area Manager, Perceptual Document Analysis
Intelligent Systems Laboratory
Palo Alto Research Center

# Every Document Tells a Story...

Goal: Connect the Dots

Hugo Yepez hosted some Japanese v

Sudan procured oil drilling equipment...

mer...

---

Enterprise 1-800 rent-a-car

Flitsch

---

LA HORA Martes, 27 de enero del 2004

BUSCAN UN ACUERDO DE COOPERACIÓN

**Ecuador y Japón intercar**
**experiencias sobre vol**

**El objetivo de los encuentros técnicos es que se firme un acuerdo de cooperación que incluye la donación de una nueva te nología de monitoreo para los volcanes Co topaxi y Tungurahua**

A FIN de mejorar la vigilancia en los volcanes Tungurahua y Cotopaxi técnicos del Instituto Geofísico de la Politécnica Nacional (IGEPN) y una misión de expertos del Japón comparten experiencias sobre las nuevas tecnologías.

Según el director del IGEPN, Hugo Yépez, la visita de los japoneses responde a un pedido del Ecuador para lograr un convenio de cooperación que podría incluir la transferencia de equipos para la vigilancia de volcanes.

"Una misión de japoneses evaluó la propuesta hecha por el país y otra misión se encuentra evaluando las condiciones de campo para la vigilancia de los volcanes", confirmó Yépez, quien espera que luego del trabajo técnico se logre la firma definitiva del convenio.

De lograrse un acuerdo, el Gobierno de Japón donaría sismógrafos de banda ancha y micrófonos con censores acústicos

Los Técnicos del Institut

que responden a muy baja frecuenci

Estos equipos, dentro de lo pre to, serían instalados en los volcanes Cotopaxi y Tungurahua, que mantienen distintos tipos de actividad y que, actualmente son vigilados por una red de sismógrafos de baja frecuencia.

**VIGILANCIA INTEGRAL**

Tras dejar claro que todo depende de la aprobación del Gobierno japonés, Yépez explicó las ventajas que tendría Ecuador con la instalación de los nuevos equipos, los cuales se usan para la vigilancia de

ría el r aparatos a una c ñales producidas

Los equipo seen una banda permitirá vigilar amplio de señale

Esta instal la incorporación en el registro y que emiten los v no como externo

---

Sudan Embassy
In South Africa

ESZA/OG/X112/CF 01/09/2006

Our Ref:..... Your Ref:..... Date.....

Dear Sir,

FROM THE DESK OF HONOURABLE MINISTER OF ENERGY AND MINERAL RESOURCES REPUBLIC OF SUDAN: DR AWAD AL-JAZ

SOURCE FOR AGENT/PURCHASE OF MACHINERY

xcuse my communication through email, on an issue as important as this one. We shall municate as soon as I receive a response from you. I write to inform you on behalf of the Director of Contracts and Finance Allocations country (Republic of Sudan).

e Machinery are available and can be obtained in nt. I have therefore been directed to inquire if you that this mail is purely for Republic of

e only source of income for the re taking advantage of the situation to even execute contracts for individuals,

st us by forwarding to me, your detail umbers for easy communication. The e have agreed to highly compensate you in your participation this transaction; nonetheless,

gh my above stated email address or ties for seeing this transaction through. Your e a positive response from you, I shall give you a complete

Dr Awad al-Jaz
Energy and Mining ministry
Government House, Khartoum
Republic of Sudan

Be notified that The Minister have a personal interest in this transaction; as a result I want my own company to handle this transaction. Due to the current war in the Darfur region, our head office has been moved to Johannesburg, South Africa due to security reasons and enabling environment. Be informed that all correspondence and meetings will be held in Johannesburg till conclusive stages which would be held in Khartoum at the Government Lodge.

SIGNED
South African Contact
Contact: Mr. Monday S. Kumba: Third Secretary
Contact Telephone and Fax numbers to send you upon reply due to security reasons.

---

Ervin Leasing Company
Dept. 77228
P.O. Box 77000
Detroit, MI 48277-0228 (734) 332-5400

INVOICE DATE 04/05/2001

| CONTRACT NUMBER | PAYMENT NUMBER | DESCRIPTION | CONTRACT PAYMENT | SALES/USE TAX | LATE CHARGES | AMOUNT |
|---|---|---|---|---|---|---|
| | | | | | 956082) | $435.72 |
| | | | | | 001 | $435.72 |
| | | | | | | 251.76 |
| | | | | | Subtotal | 251.76 |
| | | | | | | 143.96 |
| | | | | | Subtotal | 143.96 |

void late charges. **
***
***
228 ***

TOTAL DUE 395.72

DIT ON YOUR ACCOUNT
H, MI 48277-0228

| ract # | Amt Due |
|---|---|
| | 251.76 |
| | 143.96 |

al Due $ 395.72

Date: 04/20/2001

2510001 0000 3

# Outline

1. The classical document recognition pipeline
    - image preprocessing
    - page segmentation
    - text recognition
        - OCR
        - Handwriting
    - page layout analysis
    - text analysis

State of the art:
It all depends...

2. Why is this so hard?

3. Trends
    - What is industry up to?
        - the paper/digital divide
    - What are academics up to?
        - digital libraries
        - camera document imaging

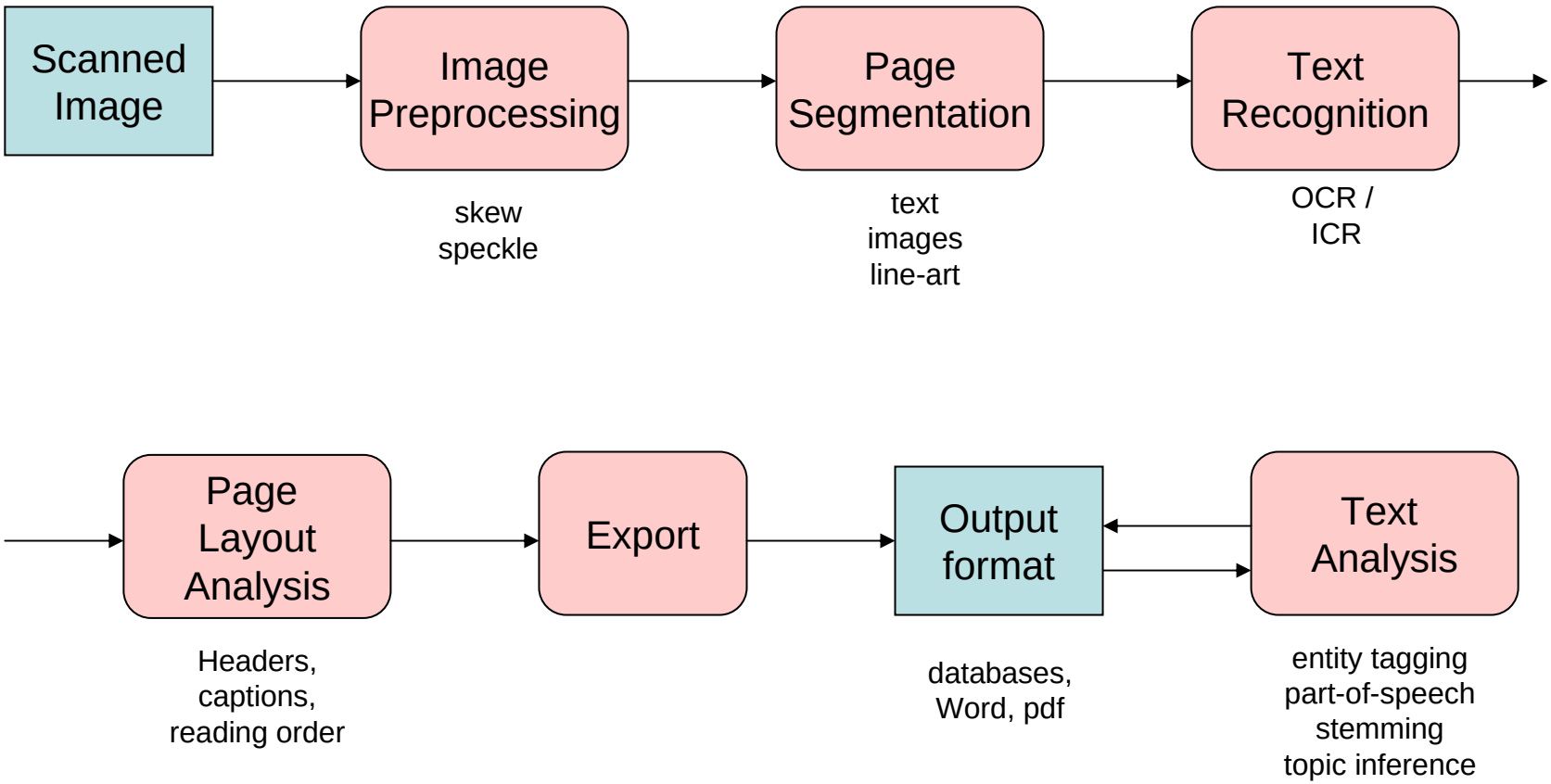
4. Structure of the Field
    - Machine learning
    - Where is research done?

5. Roadmap:  pressing problems and opportunities

# Acknowledgements

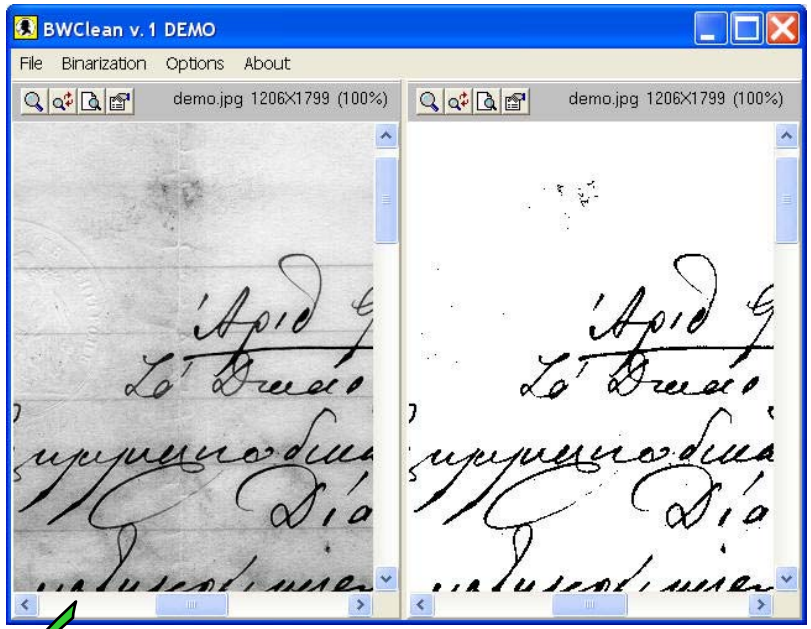
Dave Doermann, U. Maryland
R. Manmatha, U. Mass
Dan Lopresti, Lehigh U.
George Nagy, RPI
Yann LeCunn, NYU Courant Inst.
Marco Bressan, XRCE
Prateek Sarkar, PARC
Tracy King, PARC
Danny Bobrow, PARC
Stu Card, PARC

# The Classical Document Recognition Pipeline



# The Classical Document Recognition Pipeline

## Image Preprocessing




Adaptive thresholding for binarization

ScanFix XPRESS
ScanFix® Xpress v5.1
Bitonal Image Enhancement Toolkit

- Image Input, Image Output, and Image Handling
- Deskew
- Noise Removal
- Dot Shading Removal
- Hole Punch & Blob Removal
- Line Removal
- Comb Removal (commonly used in ICR forms)
- Character Completion & Smoothing
- Auto Inverse Text Correction
- Auto Negate
- Auto Binarize
- Auto Border Crop
- Image Registration
- Blank Page / Blank Rectangle Detection

Dozens of commercial products

Not available:

- separate multiple content layers
- non-uniform skew
- detect multiple documents in one scan image

# The Classical Document Recognition Pipeline

## Page Segmentation

### Purpose:

- identify zones of text, image, line-art, page graphics (separators)
- used widely to improve compression and image quality of scanners/copiers

### Main approaches:

- recursive decomposition by analyzing whitespace runs
- image morphology to form blobs of like image material


TECHNOLOGY

The Gist of Translation

How long will it be before machines make the Web multilingual?

BY JENNIFER L. SCHENKER PARIS

IMAGINE IF CYBERSHOPPERS IN PARIS, Beijing and Rio could go to the Gap's website and find a description of the clothes—and an order form—in their native language. That day isn't far off. Machine-language translation—a process that uses computer software to translate text from one language into another—is maturing precisely at the time when the need for it, particularly on international websites, is escalating. Analysts figure that by 2005 about three-quarters of the 1 billion people *communicating* and doing business online will live outside the U.S., and most of them won't speak English.

Late last month Autodesk, a U.S. software company, began offering machine-generated translation to its European customers at 50% less than it costs to have humans do the job. In its first application of the software, Autodesk is providing French and Spanish versions of 5,000 online customer-support documents—the ones that tell you what to do when you run into technical problems. German and Italian will be added in about a month, with Asian languages to follow, says Mirko Plitt, a process analyst at Autodesk's global center for localizing content in Neuchâtel, Switzerland.

The company supply ing Autodesk's transl ation software is Systran, which is based in Soisy-sous

CAN WE TALK? Autodesk's Mirko Plitt says German, Italian and Asian languages are next

HELP DESK: Employees at Autodesk in Switzerland may find their jobs easier with Systran

Montmorency, France (www.systran-soft.com). It makes software that can translate more than 20 language pairs and provides the technology behind the free online translation services offered by AltaVista's BabelFish, Google, Lycos and AOL, as well as a wireless portal run by Oracle.

The quality of these free translation services tends to be uneven because the subject matter is so diverse and the process relatively imprecise. (It is called gisting, because it yields translations that provide the gist of a document.) What Systran offers Autodesk—and other multinational companies—is improved translation software customized for the content of the website. "Multinational organizations trying to provide quality service to all their customers, business partners and employees will want to have this capability," says Steve McClure, a research vice president at the technology consultancy IDC in Framingham, Massachusetts. Given Systran's head start, IDC says it holds the No.1 position worldwide for machine-language translation, ahead of IBM and Belgium's Mendez.

Machine translation has long been an elusive goal of computer experts. Like its human counterpart, the software has to take into account the grammatical structure of each language; it can't just give a substitution for every word unless it understands all of the words in a given sentence and how one influences the other. Early attempts at decoding natural languages through mathematical techniques using closet-sized computers produced poor results. Then Peter Toma, a linguistics researcher who began his work in 1957 at the California Institute of Tech-

nology, introduced linguistic rules into the process. He founded Systran in 1968. Among its first clients was the U.S. Air Force, which asked the company to develop a Russian-into-English machine-translation system. Later work for NASA caught the attention of the European Commission, which asked Systran to develop systems to translate European languages. In 1986 the company was purchased by Gachot, a French maker of industrial valves and fittings. Systran, which trades on the Nouveau Marché, last year had profits of $876,000 on revenues of $8.5 million.

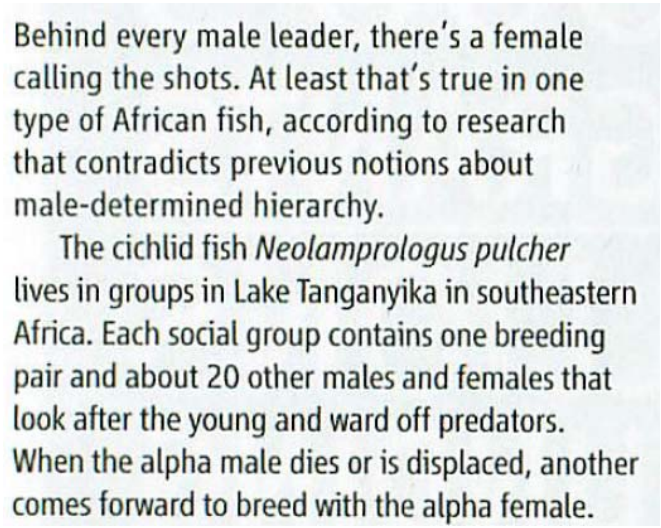
For decades Systran's machine-translation technology was available only to U.S. and European government agencies and other public institutions via mainframe computers. Today its software runs on many different hardware platforms—including the type of Linux and Unix servers used by multinational corporations. At Autodesk the Systran software will not replace the people manning the multilingual help desk in Neuchâtel, says Plitt, but it's likely to make their jobs easier.

Systran has other multinational clients, including Ford and Price Waterhouse Coopers, that use the technology internally. But to create a mass market, Systran will need a lot more customers like Autodesk, which Systran CEO Dimitris Sabatakakis calls a "showcase client," putting the software on their external websites. If analysts are right, translating customer-support information is a significant first step, after which e-commerce applications should follow. Beyond that, the World Wide Web will be not just multinational but multilingual as well.

[ICDAR Page Segmentation Competitions]

# The Classical Document Recognition Pipeline

## Optical Character Recognition



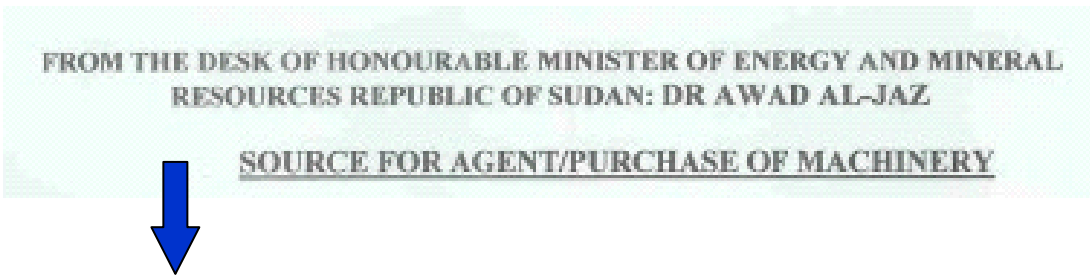
Behind every male leader, there's a female calling the shots. At least that's true in one type of African fish, according to research that contradicts previous notions about male-determined hierarchy.

The cichlid fish *Neolamprologus pulcher* lives in groups in Lake Tanganyika in southeastern Africa. Each social group contains one breeding pair and about 20 other males and females that look after the young and ward off predators. When the alpha male dies or is displaced, another comes forward to breed with the alpha female.

Behind every male leader, there's a female calling the shots. At least that's true in one type of African fish, according to research that contradicts previous notions about male-determined hierarchy.

The cichlid fish Neolampmlogus pulcher lives in groups in Lake Tanganyika in southeastern Africa. Each social group contains one breeding pair and about 20 other males and females that look after the young and ward off predators. When the alpha male dies or is displaced, another comes forward to breed with the alpha female.

*(ABBYY FineReader)*



FROM THE DESK OF HONOURABLE MINISTER OF ENERGY AND MINERAL RESOURCES REPUBLIC OF SUDAN: DR AWAD AL-JAZ

SOURCE FOR AGENT/PURCHASE OF MACHINERY

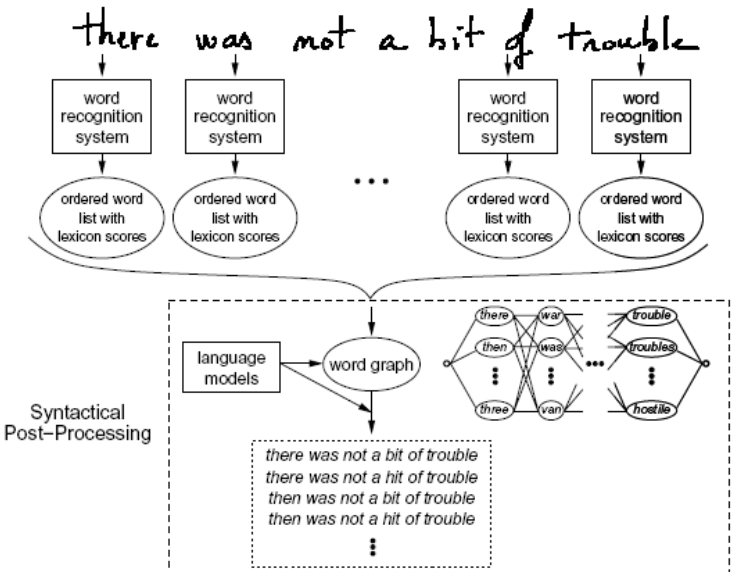
from Tin: DCSK OF HOftOftJltABLB minIsti.k09 ENKR£¥ and MINERAL
KFSC>1IK<T.S HICniBl.K Of SI DAN: OR A WAI) AL-JAZ
SOURCE FOR AOENT/PIKCHASK OF MACHI.N KRY

### State of OCR:

- commercial market mature
- results depend heavily on image quality
- “up to” 99.9% accuracy: ~ 2 errors/page
- non-Roman alphabets still lagging

# The Classical Document Recognition Pipeline

## Handwriting Recognition



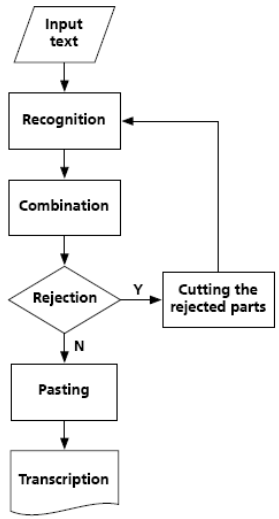

Brown corpus

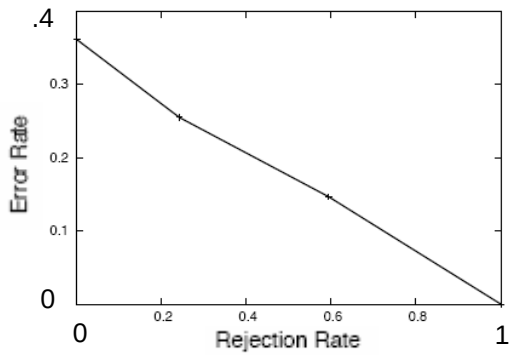
| Model | Word rec. rate |
|---|---|
| Baseline | 83.53 % |
| Bigram | 91.72 % |
| Trigram | 92.30. % |

*Quiniou & Anquetil, IWFHW 2006*

## State of handwriting recognition:

- Online writer-trained HW rec. is very good
- Offline general HW rec. unusable for transcription, but may support some search and word-spotting applications
- Isolated numerals very good: > 99%
- Finding and segmenting handwriting still an issue
- Much work to be done on non-English scripts




Word recognition error vs. reject

*Bertolami, Halter, Bunke, IWFHW 2006*

# The Classical Document Recognition Pipeline

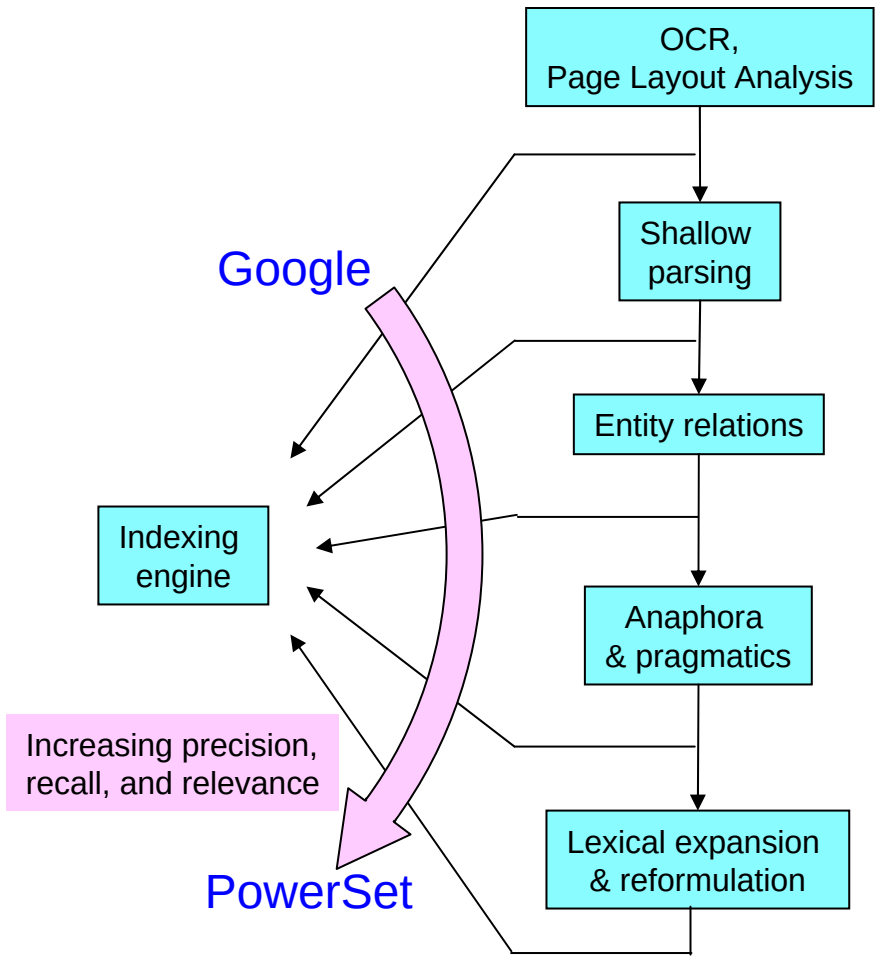
## Page Layout Analysis

- Reading order
- Functional Role Labeling (Zone typing)
  e.g. title, header, caption, page number.
- Table parsing

- Current methods mostly rule-based, heuristic, error-prone.
- Current formulations are in terms of
  - •document logical structure vs.
  - •document layout structure.
- Lack general theories and formulations for combining weak evidence.
- Lack ontologies of layout elements.
- Lack taxonomies of document genres and layout patterns.

# The Classical Document Recognition Pipeline

## Text Analysis



# Outline

1. The classical document recognition pipeline
    - image preprocessing
    - page segmentation
    - text recognition
        - OCR
        - Handwriting
    - page layout analysis
    - text analysis

State of the art:
It all depends...

2. Why is this so hard?

3. Trends
    - What is industry up to?
        - the paper/digital divide
    - What are academics up to?
        - digital libraries
        - camera document imaging

4. Structure of the Field
    - Machine learning
    - Where is research done?

5. Roadmap:  pressing problems and opportunities

# What Makes Document Recognition Difficult?

- variability
  - 100000s of journal, magazine, newspaper formats
- unpredictability, exceptions
  - handwritten annotations, post-it notes, strange fonts, languages
  - figures, graphics can include anything
- context
  - meaning of glyphs, graphics, whitespace depend on---and define---genre and style

- You can engineer a solution to any narrowly constrained problem.
- Real applications seldom obey narrow problem constraints.

# Outline

1. The classical document recognition pipeline
    - image preprocessing
    - page segmentation
    - text recognition
        - OCR
        - Handwriting
    - page layout analysis
    - text analysis

2. Why is this so hard?

3. Trends
    - What is industry up to?
        - the paper/digital divide
    - What are academics up to?
        - digital libraries
        - camera document imaging

4. Structure of the Field
    - Machine learning
    - Where is research done?

5. Roadmap:  pressing problems and opportunities

# Trends: Commercial Document Recognition

*Driving Force: Bridge the paper/digital divide*

# Trends: Commercial Document Recognition

## Market Segments

*Scale*

**Office** ⟷ **Production**

- Office: convert to word, PDF
- Production: indexing, data capture

*Document Type*

**Structured** ⟷ **Semi-structured** ⟷ **Unstructured**

- Structured: forms
- Semi-structured: invoices, bills receipts
- Unstructured: letters, slideware, email

*Value Chain*

### Components

image ~~enhancement~~,
document classification,
forms ~~layout~~, ~~OMR~~ (mark recog.)
OCR, ICR (handwriting),
key entry tools,
vectorization

ABBYY
A2iA
IRIS
Nuance

### Solutions

data capture from forms
invoice processing
health care records processing
bank check processing
engineering drawing management
litigation support

EMC / Captiva
Kofax
ReadSoft
OCE

### Services

scanning services
forms indexing
mailroom
forms data extraction
document repositories
document archiving
legal document coding
medical records scanning & indexing
mortgage document processing
drawing vectorization

Xerox Global Services
Source Corp

## Accessible $ ≠ connect the dots

- large volume of predictable documents
- offshoring
- value to customer

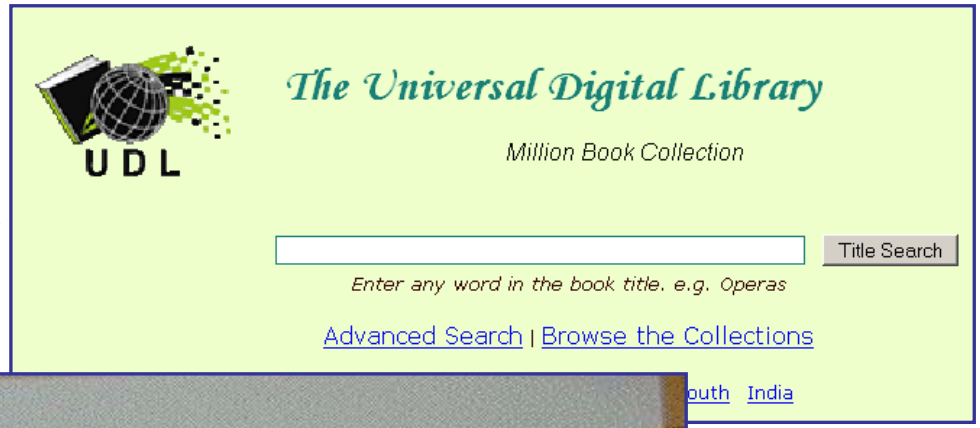
- large volume of variable documents
- human assist
- need flexible access for fishing

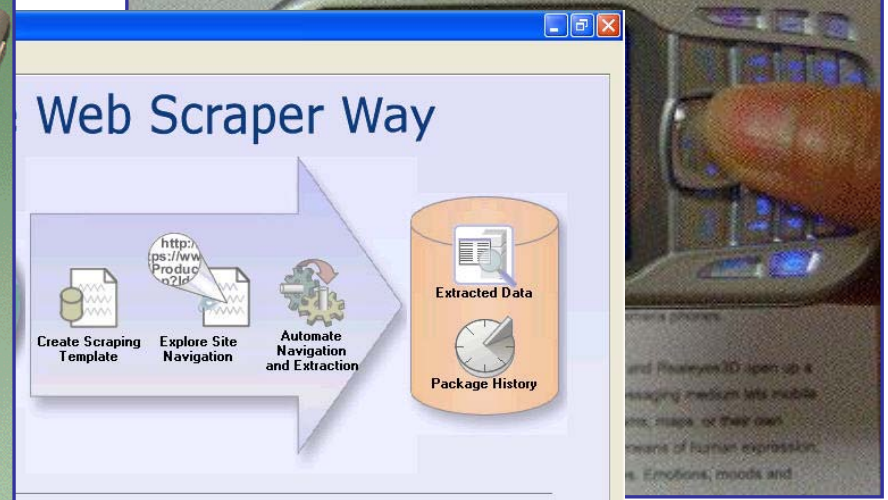
# Trends: Academic Research

Drivers:

- Digital libraries

  Google books, Million book project, historical document mining, geneology
- Cameras
- Web data capture
- Pen computing



# Trends: Academic Research

Drivers:

- Digital libraries

  Google books, Million book project,
  historical document mining, geneology
- Cameras
- Web data capture
- Pen computing

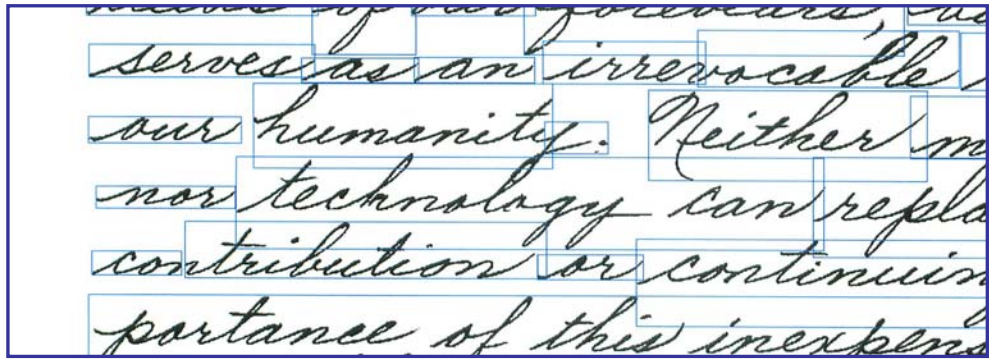
Topics:

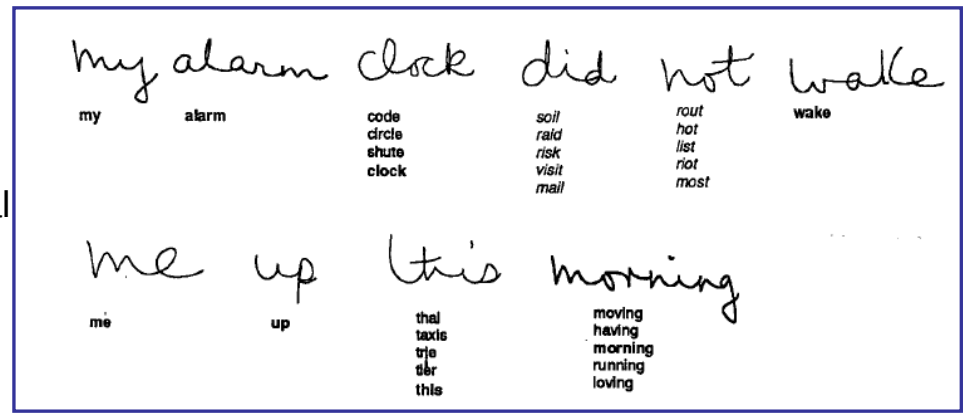
- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps

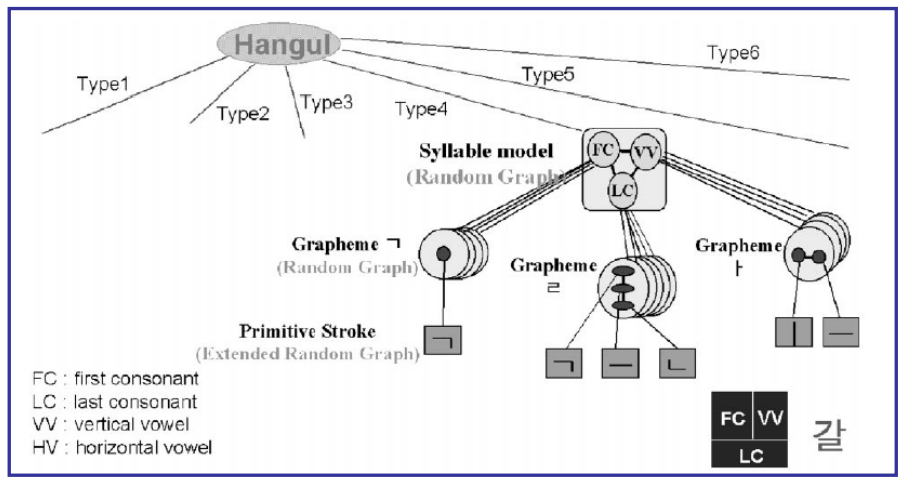
# Trends: Academic Research

Topics:

- **handwriting**
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps





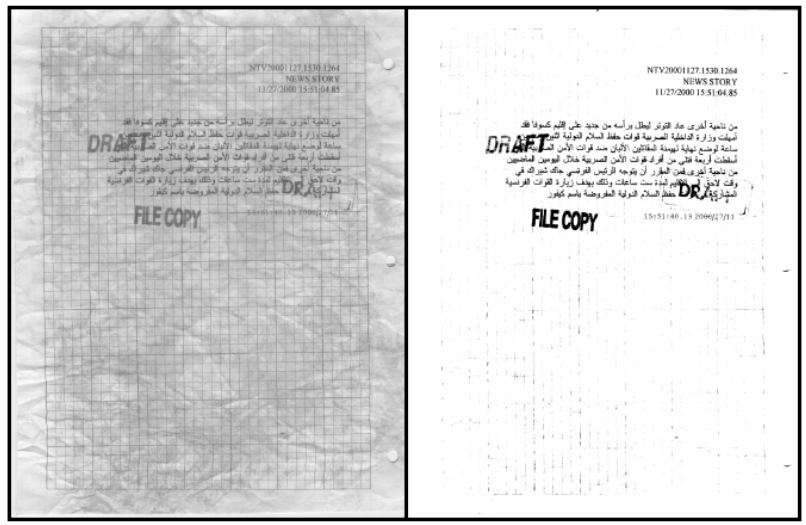




- The “face recognition” of document image analysis.

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- **OCR on degraded documents**
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps




**Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents**

Paul Herceg, Ben Huyck, Chris Johnson, Linda Van Guilder, Amlan Kundu*
The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102-7508

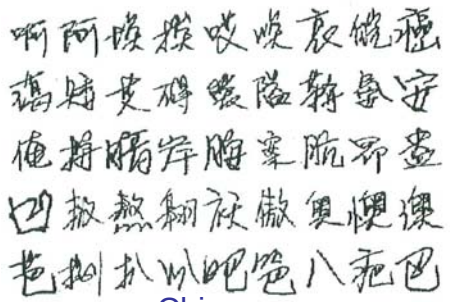
- Lots of very helpful image processing can be done.
- The problem is knowing what to apply to any given image.

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- **languages, non-roman alphabets**
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps

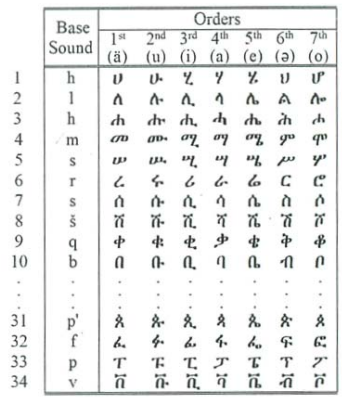
Devnagari



Chinese

Arabic

Bangla



Ethiopic

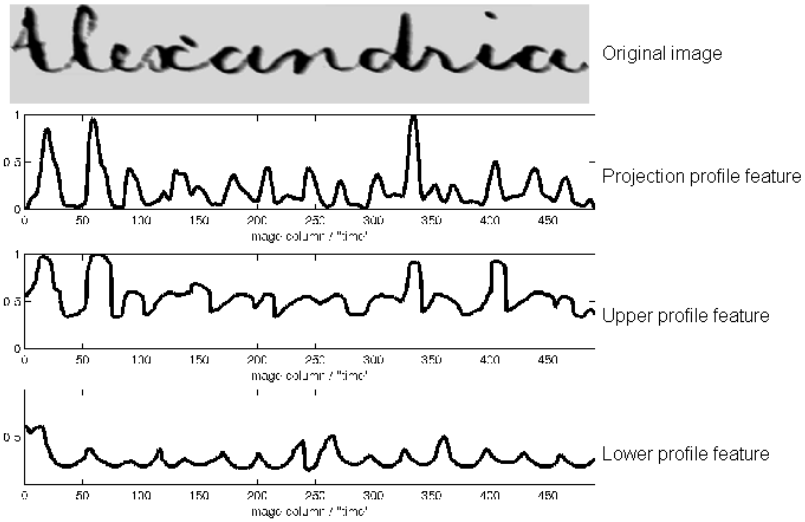
Tamil

- Lots of activity overseas is training a world full of document recognition practitioners.

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- **word spotting**
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps





- Task-driven workaround for difficulty of handwriting recognition & OCR on difficult documents.

*R. Manmatha, U Mass*

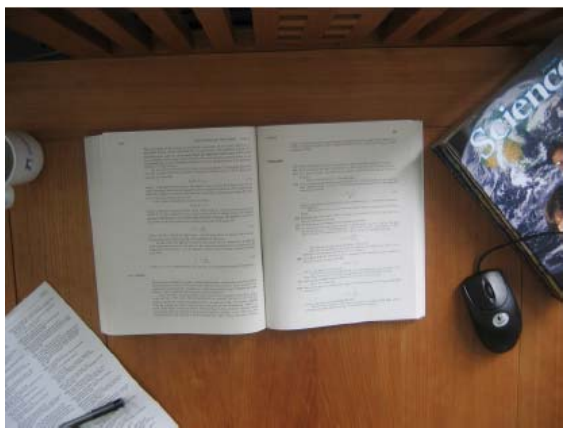
# Trends: Academic Research

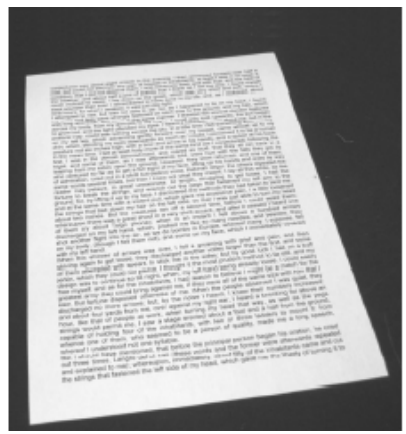
Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- **image preprocessing for lighting, deformation, degraded sources**
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps


*Landon, Lin, Seales, U. Kentucky*

- Ubiquitous computing: potential to radically transform work.





*Lampert, Braun, Ulges, Keysers, Breuel, IUPR, Germany*

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- **scene text - finding and reading**
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps



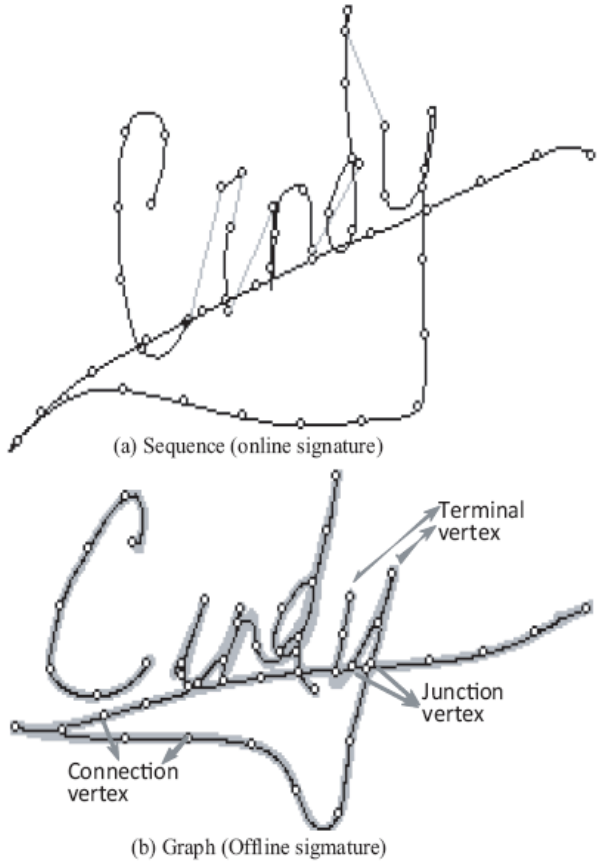
- Expect language translastion in Japanese cameras soon...

*Silapachote, Weinman, Hanson, Weissy, Mattar, U Mass*

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- **writer ID / signature verification**
- document image matching and retrieval
- layout modeling
- graphics recognition
- maps




*Qiao, Liu, Tang, Chinese U. of Hong Kong*

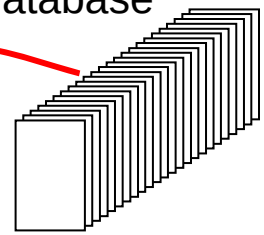
- Useful for filtering obvious cases, but even experts argue.

# Trends: Academic Research

Topics:

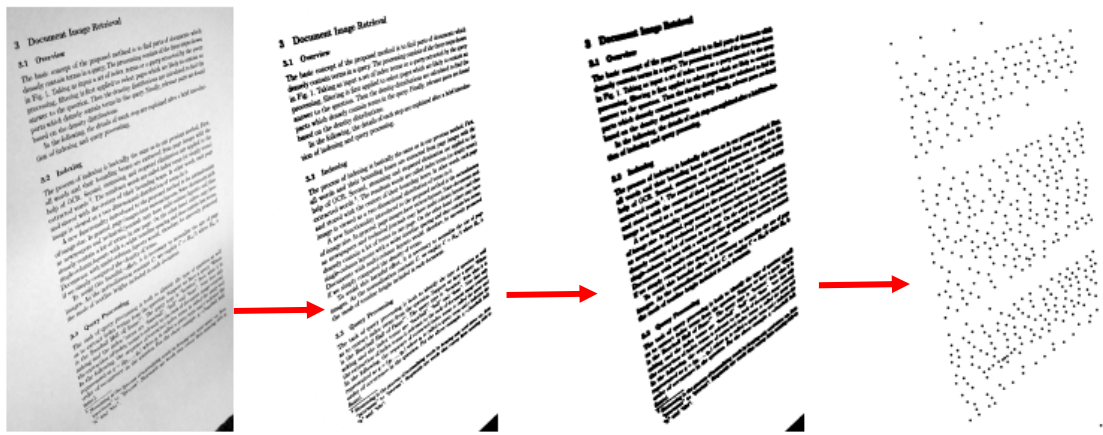
- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- **document image matching and retrieval**
- layout modeling
- graphics recognition
- maps

Document Image Database



fingerprint index

- Breakthrough is in the lab now, looking for customers.



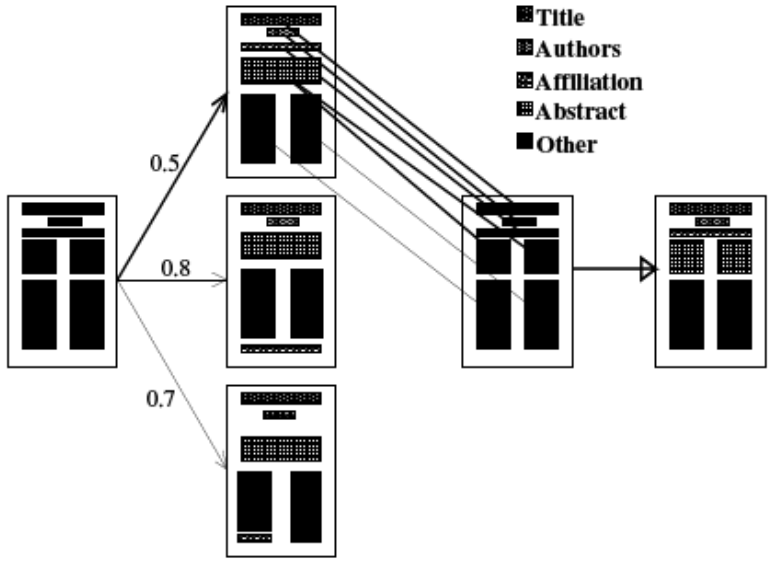
(a) Input image. (b) Binarized image. (c) Connected components. (d) Feature points.

*Nakai, Kise, Iwamura, Osaka Prefecture University*

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- **layout modeling**
- graphics recognition
- maps



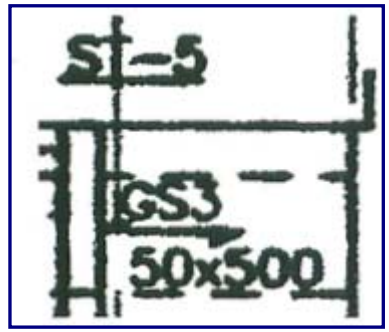

*Beusekom, Keysers, Shafait, Breuel, DFKI, Germany*

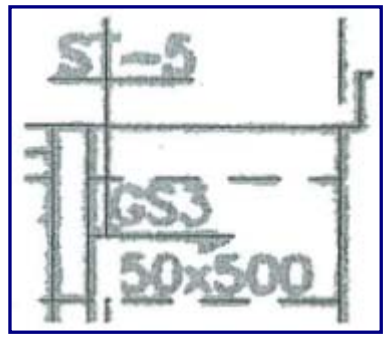
- No one has a clue here.

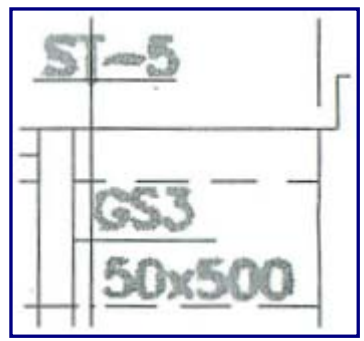
# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retrieval
- layout modeling
- **graphics recognition**
- maps



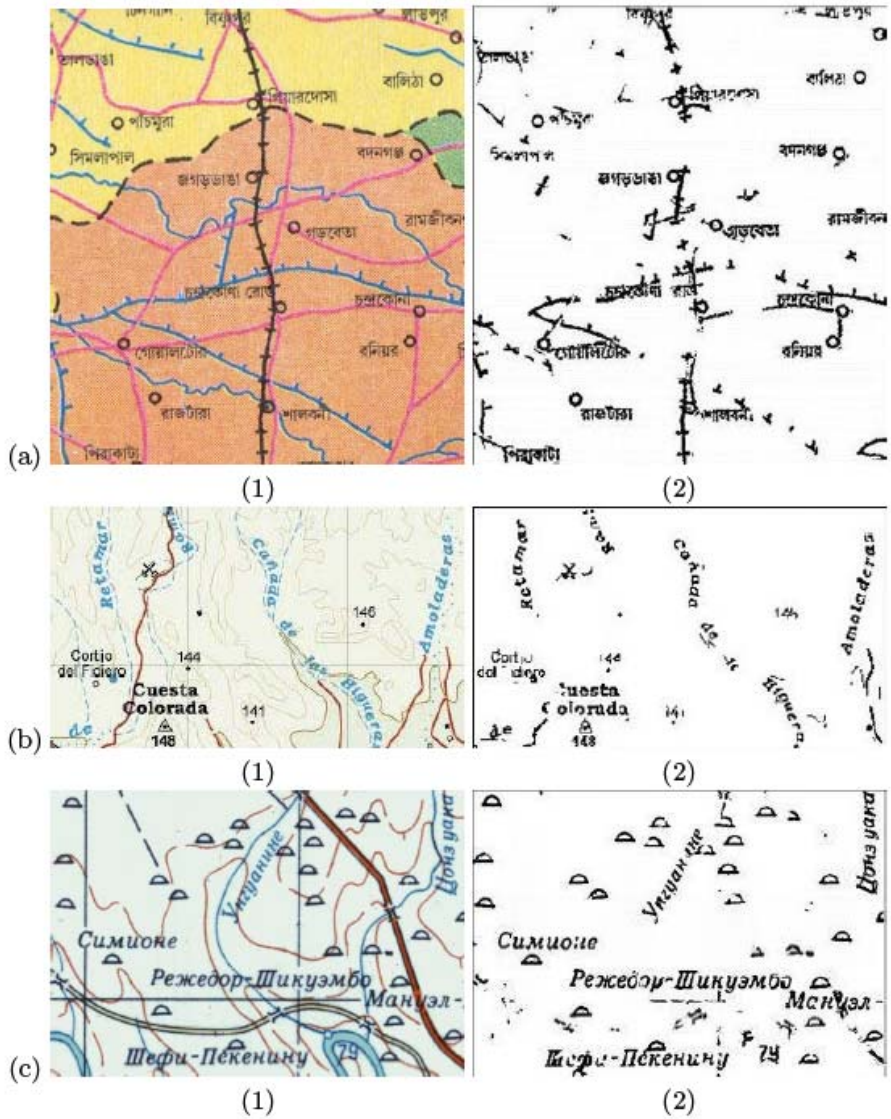

*Su, Song, Cai, Nanjing U.*

- A whole field unto itself, no longer stuck at vectorization.

# Trends: Academic Research

Topics:

- handwriting
  - parsing and segmentation
  - lexical constraints
  - learning and inference methods
- OCR on degraded documents
- languages, non-roman alphabets
- word spotting
- image preprocessing for lighting, deformation, degraded sources
- scene text - finding and reading
- writer ID / signature verification
- document image matching and retriev
- layout modeling
- graphics recognition
- **maps**




*Roy, Vazquez, Llados, Baldrich, Pal, Autonoma de Barcelona*

- Incredibly challenging, human assist seems the only way forward for now.

# Outline

1. The classical document recognition pipeline
    - image preprocessing
    - page segmentation
    - text recognition
        - OCR
        - Handwriting
    - page layout analysis
    - text analysis

2. Why is this so hard?

3. Trends
    - What is industry up to?
        - the paper/digital divide
    - What are academics up to?
        - digital libraries
        - camera document imaging

4. Structure of the Field
    - Machine learning
    - Where is research done?

5. Roadmap:  pressing problems and opportunities

# Machine Learning in Document Recognition

- Primary use: adjusting parameters of fixed models based on statistics of data
  - character classification
  - document type classification
  - HMMs for string recognition (e.g. handwritten words)

  } Kernel / Support Vector Machine
  Neural Net / Multilayer Perceptron
  Decision Tree

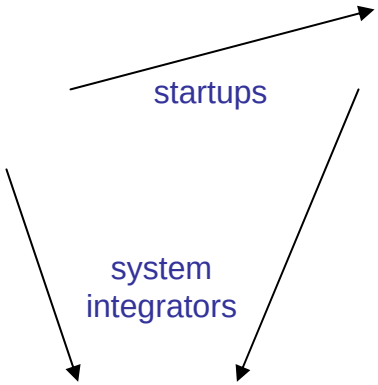
*Data* is critical
- large quantities
- preferably real, not simulated
- groundtruthing
- competitive advantage in industry
- data obfuscation

- Exploratory:
  - learning of features as well as classes (e.g. deep belief networks)
  - unsupervised e.g. clustering to learn novel glyph sets
  - build structural models of layout patterns
  - MRF, CRF models for page segmentation
  - discriminative learning for structured output

# Where is Progress in Document Recognition?

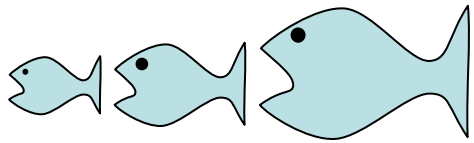
## Universities

- harvest advances in computer vision, machine learning, natural language, digitial libraries
- imaginitive but need connection to real-world problems




## Industry

- short-term outlook
- secretive
- R&D by acquistion

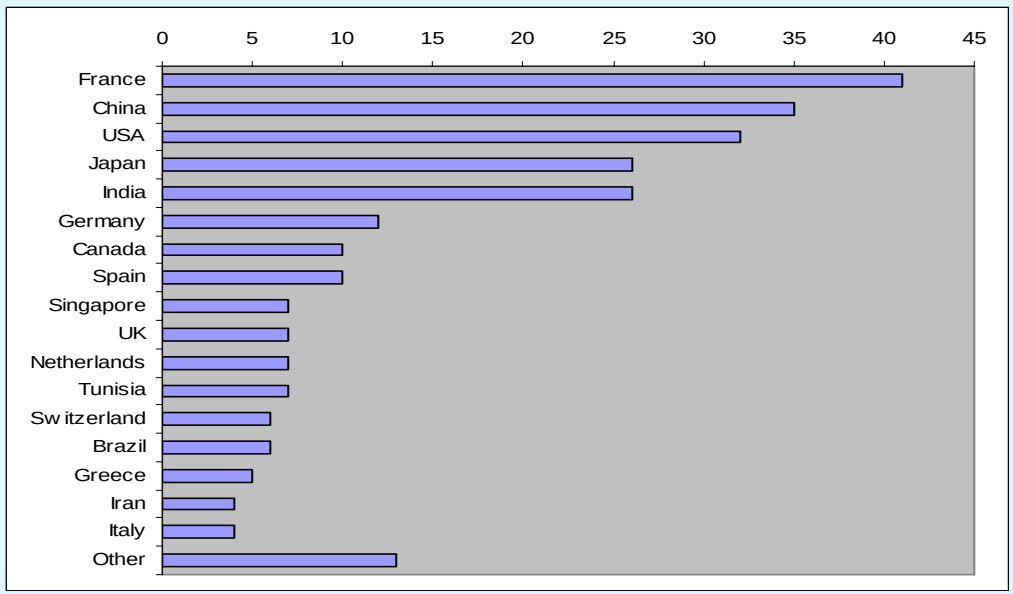


Neurascript Kofax DICOM

SWT Captiva EMC

## Government

ICDAR 2007 papers by country




✓ Microsoft Research

# Outline

1. The classical document recognition pipeline
    - image preprocessing
    - page segmentation
    - text recognition
        - OCR
        - Handwriting
    - page layout analysis
    - text analysis

2. Why is this so hard?

3. Trends
    - What is industry up to?
        - the paper/digital divide
    - What are academics up to?
        - digital libraries
        - camera document imaging

4. Structure of the Field
    - Machine learning
    - Where is research done?

5. Roadmap:
    - bottom line state of the art
    - pressing problems and opportunities

# Bottom Line State of the Art

| | Available Commercially | Current research / Could be engineered | Future Research |
|---|---|---|---|
| *Office OCR* | • convert clean office memo for pdf, editing | • extract main fields and topics of memos in company correspondence | • scan and sensibly organize the contents of an office (mixture of memos, published material, receipts, notes, invoices, etc.) |
| *Books* | • scan and OCR books for search | • extract structured representations for Russian language proceedings of one publication | • parse all encyclopedias since 1900 |
| *Handwriting* | • index online handwritten notes from a Tablet PC | • index and search names and places in neatly handwritten papers | • transcribe the notes off a typical conference room whiteboard |
| *Cameras and scene text* | | • spot and translate street signs in camera images | • translate a posted bus schedule |
| *Database search* | • find potential document duplicates through text string matching (depends on good OCR) | • find document matches in a database from a low-res camera image | • automatically construct a version history of documents scanned mulitple times over their authoring cycle |

# Roadmap: Pressing Problems and Opportunities

*Topics*

- structural models for document layout
  - combined segmentation and labeling
  - representations for variation, patterns, degrees of freedom
  - learning
  - inference
- principles for document genre
  - multidimensional types and taxonomies
  - multivalent definitions of similarity
  - learn from corpora
- document collections
  - relations among documents and items within a “case”
  - probabilistic logic – combine rules with statistics

*Process*

- build data sets reflecting problems of interests
- build community
  - workshops
  - competitions
  - evaluation methodologies
  - “a few good researchers”
- leverage cross-discipline interaction
  - computer vision, machine learning, natural language, Artificial Intelligence
- mine open source

saund@parc.com

International Conference on Document Analysis and Recognition (ICDAR)
International Journal of Document Analysis and Recognition (IJDAR)
Document Image Analysis for Libraries (DIAL)
International Conference on Pattern Recognition (ICPR)
IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
International Workshop on the Frontiers of Handwriting Recognition (IFWHR)
Association for Image and Information Management (AIIM)

# What Makes Document Recognition Difficult?

- variability
  - 100000s of journal, magazine, newspaper formats
- unpredictability, exceptions
  - handwritten annotations, post-it notes, strange fonts, languages
  - figures, graphics can include anything
- context
  - meaning of glyphs, graphics, whitespace depend on---and define---genre and style
- problem definition

simple transcription ← ? → build a resource for exploring knowledge

- You can engineer a solution to any narrowly constrained problem.
- Real applications seldom obey narrow problem constraints.