CLEAT:

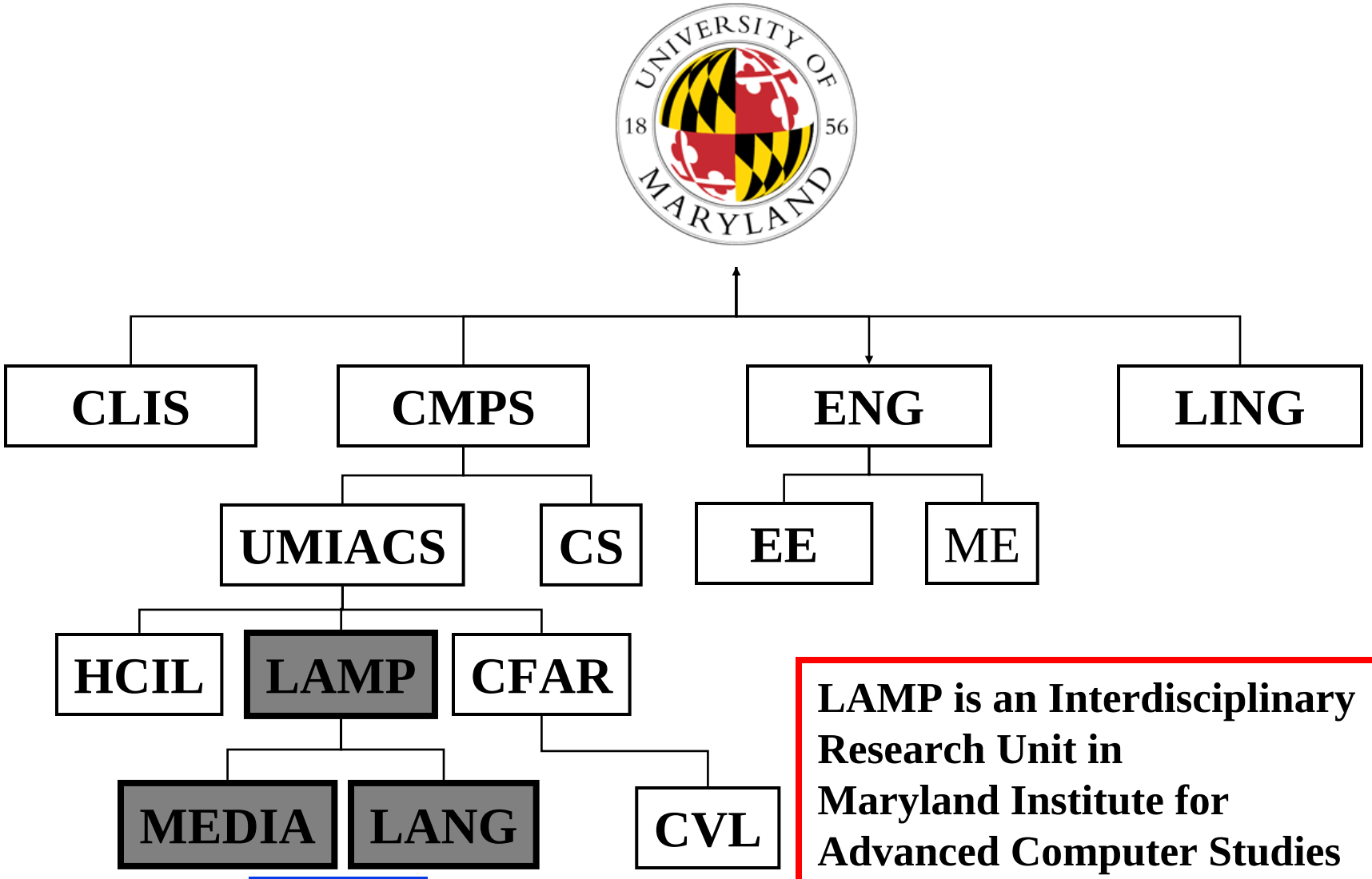
A CLassification, Enhancement and Analysis Toolkit for Heterogeneous Document Image Collections

LAMP History

- Began in 1996 with a focus on documents
- Produced 9 PhD (2 more expected in 2007)
- Over 200 scientific publications
- Almost 50 Students (Undergrad-Graduate)
- Numerous Technology Transfer Opportunities

Mission

To conduct research and education in analysis and processing of multimedia information sources including documents, images and video, to develop natural language tools for real world applications, and to foster collaboration in these areas between researchers at the university and representatives of government agencies and industry



LAMP is an Interdisciplinary Research Unit in Maryland Institute for Advanced Computer Studies (UMIACS)

Media Personnel

RESEARCH FACULTY

DOCUMENTS		***VIDEO***
Stefan Jaeger	***David Doermann***	***Daniel DeMenthon*****

Graduate Students

Guangyu Zhu	**Burcu Karagol-Ayan**	**David Mihalcik***
Yi Li	**Sameer Kibey***	**Yang Yu**
Rajat Ahuja	**May Huang**	**Xiaodong Yu**
Nagia Ghanem*	**Zhe Lin**	**Mudit Agrawal**
Xu Liu		

Undergraduate Researchers

Mike Roth*

Faculty Research Associate

Tandeep Sidhu

Affiliates

- Research/Sponsor Affiliations
  - Ricoh Japan/CRC (Scanning Equipment)
  - Xerox (OCR software), Panasonic, KLA-Tencor, Hitachi
  - University of Oulu (Multimedia Documents in Telecommunications)
  - Army Research Laboratory, Other Intelligence Community Participants

National Security Agency

ARL

HITACHI
Inspire the Next

RICOH

THE DOCUMENT COMPANY
XEROX

Panasonic
ideas for life

Research Focal Areas

- Document image analysis
  - Providing fundamental tools for the enhancement summarization, navigation, indexing and retrieval in document image databases
- Content based video analysis
  - Providing access to video content through extraction, structure representation, classification, visualization and indexing
- In General
  - Ability to access large heterogeneous collections of material
  - Adaptable systems – OCR, MT
  - Low density to resource poor languages
  - Enhancing low quality input – document images, OCR

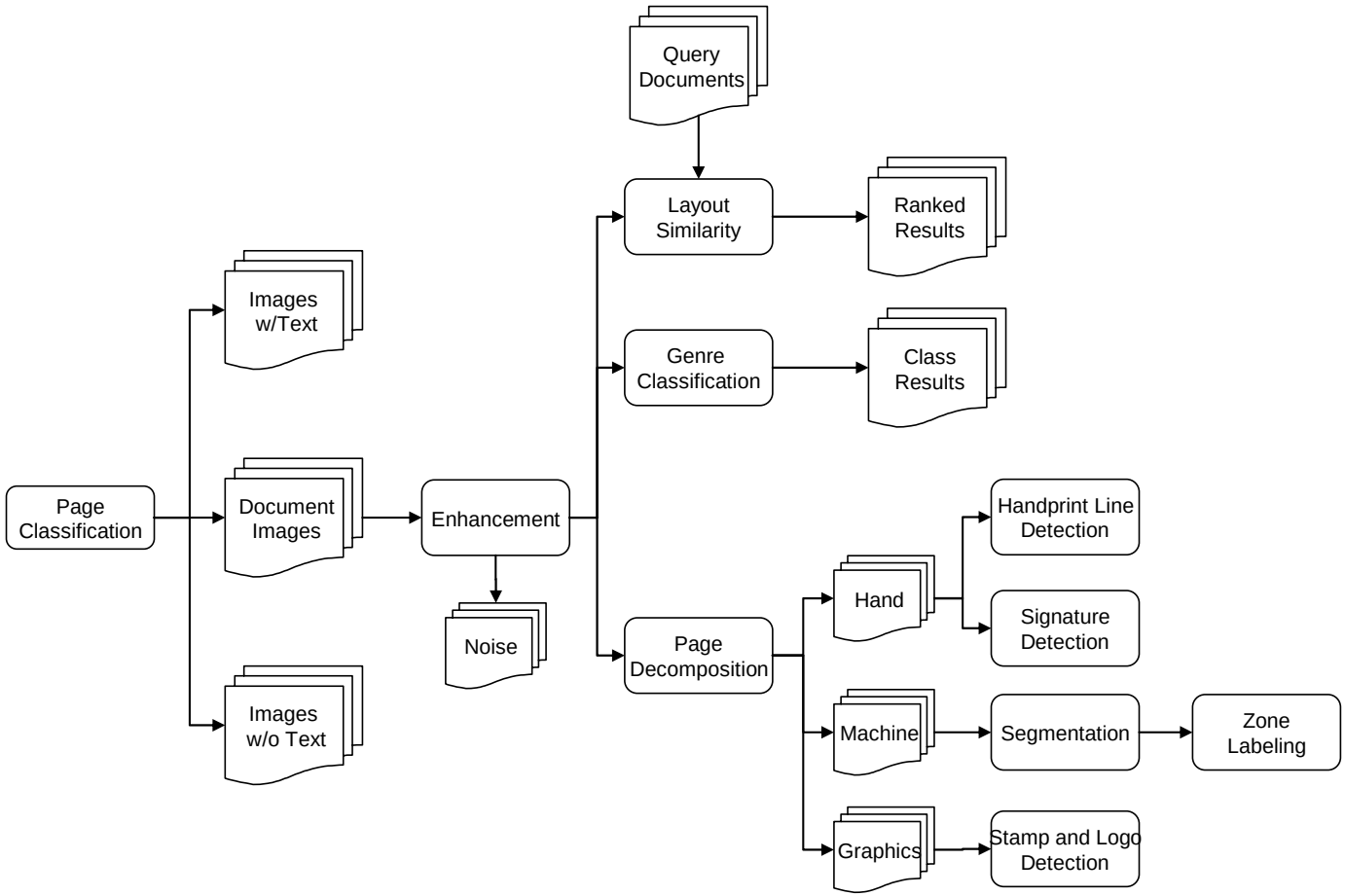
Outreach

- Bi-Annual SDIUT Conference
  - Soon to be included in Google Books Project
- Host of workshops and short courses
- Editorial Office of IJDAR
- Data Collection and Evaluations
- LAMP Seminar Series
- Chairing Program Committee for ICDAR 2007
- Organizing Arabic OCR competition at ICDAR’07

Agenda

- **Project Overview**
  - **Introduction**
  - **Goals and Objectives**
- **Data Collection and Ground Truth**
- **GEDI and Evaluation Framework**
- **Evaluation and Research Components**
  - **Unconstrained Signature and Logo Detection and Matching for Off-line Document Image Retrieval**
  - **Document/Non-Document Discrimination**
- **Technical Presentations**
  - **Adaptive OCR**
  - **Document Classification by Layout**
  - **Learning from Web for Shape-based Object Recognition**

Project Overview



Target Processing Speed in Seconds

Task Objectives

Task 1: Data Collection
Task 2: Ground Truthing
Task 3: Evaluation Framework
Task 4: Evaluation and Visualization Tool
Task 5: Page Classification Module
Task 6: Enhancement Module
Task 7: Layout Analysis Module
Task 8: Content Labeling module
Task 9: Evaluation
Task 10: Training

Performance Goals

Task	Performance Goal
Page Classification	80% precision across all three classes
Enhancement	10-30% increase in accuracy of downstream processes – segmentation, detection
Layer Separation	90% coverage at the pixel level
Segmentation (Print and Hand)	85% using implementation of existing methods
Logo and Stamp Detection	75% precision at 85% recall
Signature Detection	75% precision at 85% recall

- **Phase 1 - March 31, 2007**
  - Deliver completed CLEAT data collection.
  - Provide ground truth for subset of data including signatures, stamps, logos, handwritten, and machine printed text.
  - Provide document describing evaluation framework.

Data Collection and Evaluation

Type	Number
Class 1: Traditional Document Images	9000
Class 2: Camera captured, Text in Scene, and Color documents	500
Class 3: Non-document Images	500

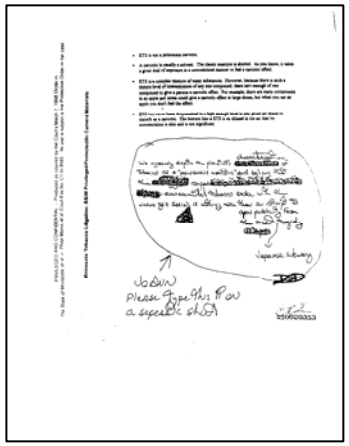
Genre	Number
Forms, Drawing, Tables	1000
Business Documents, Memos, Letters	2500
Journal and Conference Papers, Articles	2500
Newsletters, Flyers	1000
Structured Documents – phone books, dictionaries	1000
Handwritten	1000
Foreign Language – handwritten and machine printed	1000
Highly Degraded	500
Mixed Annotation	2000

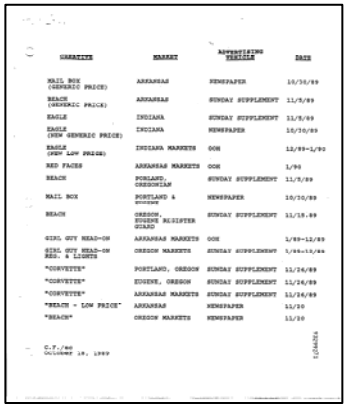
Document Image Acquisition

- Sampling of Existing Databases
  - 20-25%
- Google Image Search
  - 60%
- Scanning hardcopy Document Images
  - 15-20%

Document Genres

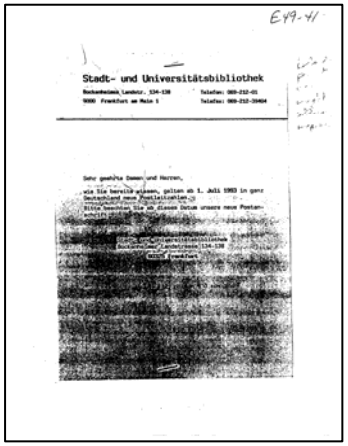
Genre	
Forms, Drawing, Tables et at.	
Forms	650
Drawing	80
Tables	100
Chemistry formulae	25
Math equations	165
Figures	40
Total	1060
Business documents and Memo letters	
Business documents	50
Business documents degraded	2700
Business documents with annotations	160
Memo letters	900
Total	3810
Journal and Conference Papers, Articles	
English	2785
German	360
Japanese	480
Total	3625
Newsletters and Flyers	
Google images	2400
Structured Documents	
Phonebook	229
Dictionaries (Chinese English, English Chinese)	1150
Yellowpage	80
Total	1459
Structured Documents	
Phonebook	229
Dictionaries (Chinese English, English Chinese)	1150
Yellowpage	80
Total	1459
Handwritten	
Chinese	146
Cyrillic	410
Japanese	47
Korean	80
Thai	319
Hindi	281
	1283

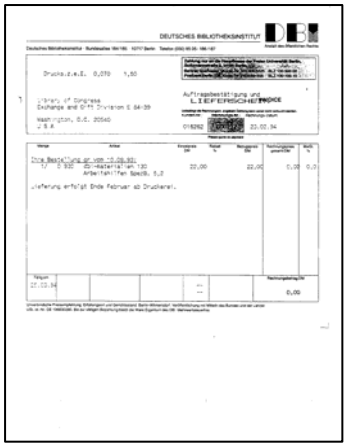


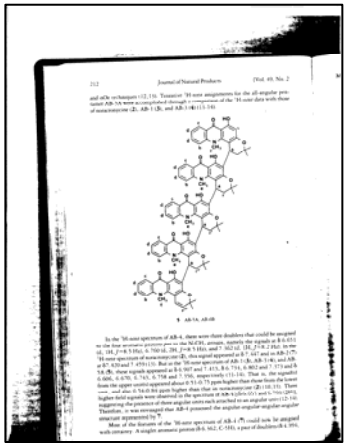


CIGARETTE	MARKET	ADVERTISING VEHICLE	DATE
MAIL BOX (GENERIC PRICE)	ARKANSAS	NEWSPAPER	10/30/89
BEACH (GENERIC PRICE)	ARKANSAS	SUNDAY SUPPLEMENT	11/5/89
EAGLE	INDIANA	SUNDAY SUPPLEMENT	11/5/89
EAGLE (NEW GENERIC PRICE)	INDIANA	NEWSPAPER	10/30/89
EAGLE (NEW LOW PRICE)	INDIANA MARKETS	OOH	12/89-1/90
RED FACES	ARKANSAS MARKETS	OOH	1/90
BEACH	PORTLAND, OREGONIAN	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & environs	NEWSPAPER	10/30/89
BEACH	OREGON, EUGENE REGISTER GUARD	SUNDAY SUPPLEMENT	11/18.89
GIRL GUY HEAD-ON	ARKANSAS MARKETS	OOH	1/89-12/89
GIRL GUY HEAD-ON RED, & LIGHTS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/90-11/90
"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

C.T./mc
October 18, 1989







Internet Downloads

Genre	
Figure	
Good	240
Medium	755
Low	548
Form	
Good	66
Medium	69
Low	32
Letter-Memo	
Good	55
Medium	88
Low	31
LIST	
Good	6
Medium	34
Low	11
Newspaper	
Good	22
Medium	37
Low	17
Publication Cover	
Good	130
Medium	425
Low	128
Receipt	
Good	10
Medium	50
Low	20
Screenshot	
Good	184
Medium	848
Low	566
Table	
Good	52
Medium	124
Low	42

Maryland Datasets

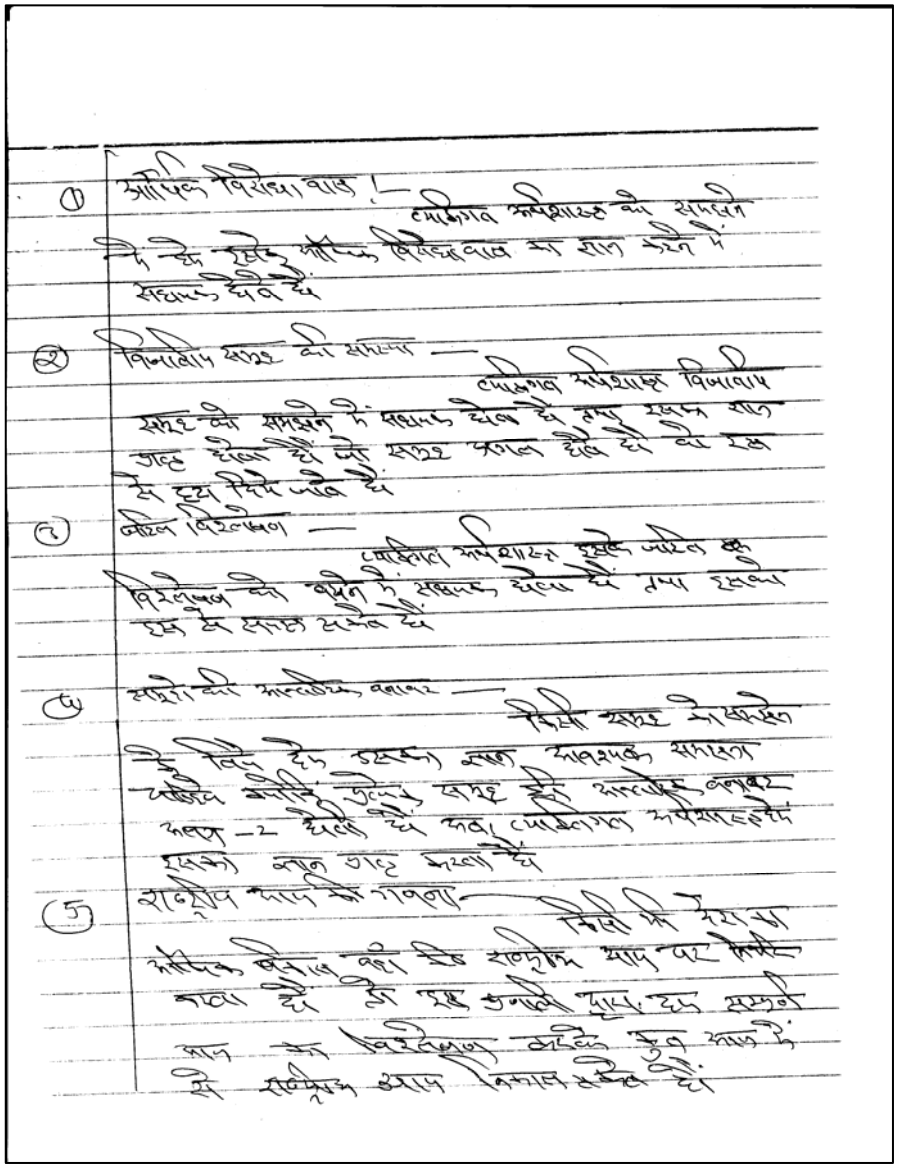
- Collection of Free form Handwriting
  - Paid upto $1 for pages of native handwriting
  - Languages: Arabic, Amharic, Chinese, Korean, Japanese, Greek, Cyrillic, Hebrew, Thai, Burmese, and Hindi
  - Up to 1000 pages of each

以先治网的遗憾

李网荟

赵紫阳不幸逝世了，亚讯传来仅人感到无比的悲痛：他为中网的改革开放和社会、经济的发展作出了不可磨灭的巨大贡献，网初人士都颂他，赞扬他，爱戴他！然而由于"六·四"事件仅他丧失了人身自由，十五年漫长的日日夜夜，但对他是：没有作一个爱国者不足羞愧的事使，仅他把握终身，无终不遗憾了！

现在不谈赵紫阳的"错误"性质为何；就在对他的处理方式而言，也是极端错误的，遗臂了以先治网的思想。仅人世为遗憾的是，力告有的领导人，一方面大谈特谈以先治网（当然也很必要），而另一方面却又限制以先治网，甚至践踏以先治网。事实，赵紫阳是建立网家的重要领导人，袁世凯窃网权，击败孙中山的总统击光，然而，未经任何先律手续对网作审判，仅残网元人的残向法，我把赵紫阳长期软禁起来，剥夺他的人身自由长达十五年之久。这雏道不是对以先治网的嘲讽和调刺？足对以先



ที่ อเมริกา

5 กันยายน 2521

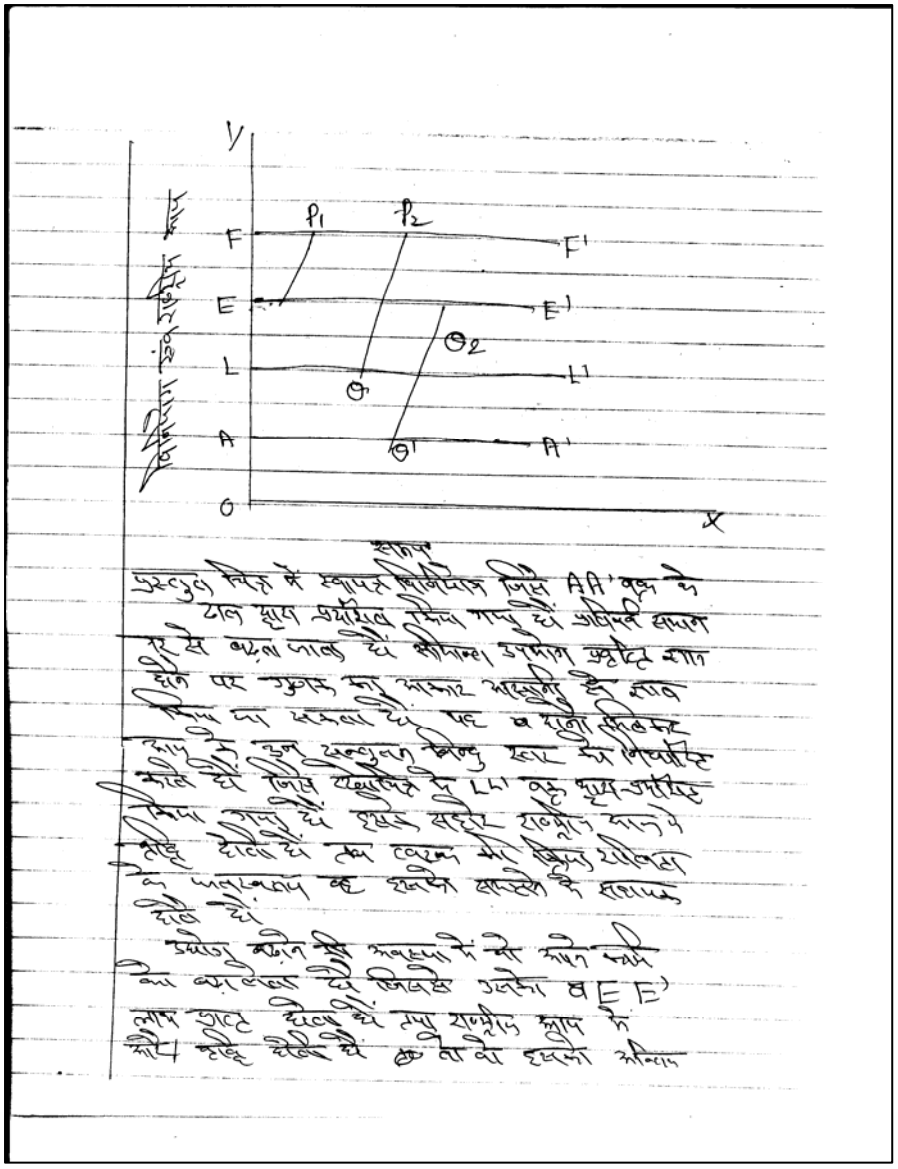
สวัสดีเพื่อนรัก

ตั้งแต่ฉันจากเมืองไทยไปอยู่อเมริกาฉันคงเขียนจดหมายถึงเพื่อนแทบจะไม่ครบสิบฉบับ แต่ฉันไม่เคยได้รับจดหมายตอบจากเพื่อนเลย จะเป็นเพราะว่าเพื่อนไม่ได้รับจดหมายของฉันหรือไงก็ไม่ทราบ ฉันคิดถึงเพื่อนมาก แต่เนื่องจากค่าครองชีพที่นี่สูงมาก และหนังสือเรียนก็หนักจริงๆ ฉันจึงไม่ค่อยมีเวลาเขียนจดหมายถึงเพื่อนเท่าที่ควร ฉันหวังว่าเพื่อนคงเข้าใจฉันดี

ฉันเข้าใจว่าเพื่อนคงมีเรื่องมากมายที่ต้องการจะเขียนถึงฉัน แต่ฉันก็ไม่อยากให้เพื่อนฉันมีภาระมากเกินไป ฉันจึงขอเพียงแค่ให้เพื่อนเขียนจดหมายมาหาฉันเดือนละครั้งก็พอแล้ว ฉันก็จะพอใจแล้ว ถึงแม้ว่าเพื่อนจะไม่มีเรื่องอะไรจะเขียนมาก ก็ขอให้เขียนมาสั้นๆ ก็ได้ ฉันก็จะดีใจมากแล้ว ฉันอยากรู้ข่าวคราวของเพื่อนๆ ทุกคนมาก ว่าตอนนี้เป็นอย่างไรกันบ้าง ใครแต่งงานแล้วบ้าง ใครมีลูกแล้วบ้าง ใครเรียนจบแล้วบ้าง ใครยังเรียนอยู่ ใครทำงานที่ไหนแล้วบ้าง ฉันอยากรู้ข่าวคราวของเพื่อนๆ ทุกคนมากจริงๆ ถึงแม้ว่าฉันจะไม่ได้อยู่เมืองไทยแล้ว ฉันก็ยังรักเพื่อนๆ ทุกคนเสมอ

ขอให้เพื่อนๆ ทุกคนมีความสุขมากๆ นะ

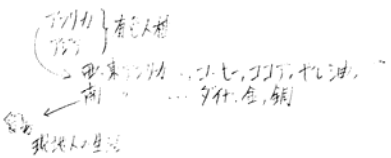
รักและคิดถึงเพื่อนเสมอ

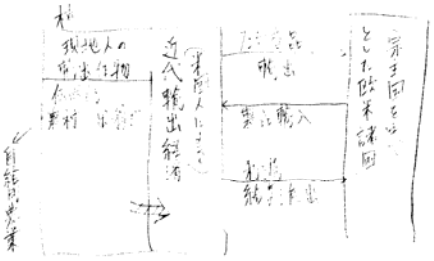


Ⅱ 遼上國の社会経済的特質

1. 基本的構造 ～ 従属型二重経済

今までは百人隊経済構造を作りました。これは植民地型二重経済の時はかりでなく、他の多くの植民地(米カナダ、NZ、カナダ、ブラジル)が資本の発展によって急速に創出されるのにこれら植民地は経済を余儀なくされた。





植民地二重経済とは、本国経済のみが決定的な主導的に特化し、その生産物をこの特定の外国へ、我地人が自給的農業と低賃金労働を生産する構造。

6. The Three Kingdoms and the Six Dynasty

p85 ferment	n.	불안, 동요, 정치적 동란
usurp	v.	강탈하다
incessant	a.	끊임없는
shrink	v.	움츠러들다, 줄어들다, 기가 죽다 (p.p. shrunken)
rashly	ad.	무모하게, 성급하게
massacre	v.	학살하다
flee	v.	달아나다, 도피하다 (p. fled)
tribe	n.	부족, 종족
abandon	v.	버리다, 포기하다
nomadic	a.	유목인의, 유목 생활을 하는
* Sinicize	v.	중국화하다, 중국식으로 만들다
cavalry	n.	기병대, 기마부대
refugee	n.	망명자, 피난자, 난민
perpetual	a.	영구한, 부단한
turmoil	n.	혼란, 동요, 불안
p87 undermind	v.	약화시키다, 서서히 무너지게 하다
* monastery	n.	수도원 산문 '사찰'(x)
vast	a.	광대한, 거대한, 엄청난
* proportion	n.	크기, 비율, 비중
realm /relm/	n.	왕국, 영역
bureaucracy	n.	관료제, 관료정치, 관료주의
* exert	v.	(힘·권력 등을) 행사하다, (영향을) 미치다
		< Taoism > 도교
Taoism	n.	도교
* calligraphy	n.	서예, 서도, 달필, 필체
conglomeration	n.	응집, 집합

HONG DANG MOO FANCY

И. Бродский.

Под вечер он видит...

— 1. —

Под вечер он видит, засунувши в дверях:
два всадника скачут в окрестных полях,
как будто по кругу, сквозь рощу и гать,
и дольше не могут друг друга догнать.
То бросив поводья, пошедших в галоп,
то снова в седле возбудившего привстав,
и быстро по светлому склону холма,
то в рощу опять, где сгущается тьма.

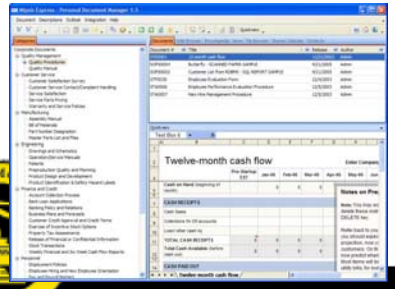
Два всадника скачут в вечерней глуши,
не только от дома, от сердца вдали,
друг друга они окликают, зовут,
небесные рати за рощу ушли.
И так никогда им на свете вдвоём
сквозь рощу и гать, сквозь пустой водоём,
не ехать в виду спящих полей,
как будто меж ними не сотня вёрст!

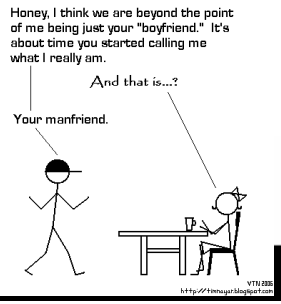
Стихи под эпиграфом. И. Бродский.

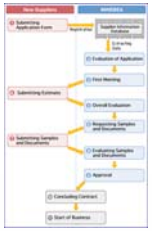
„То, что дозволено Юпитеру,
не дозволено быку...“

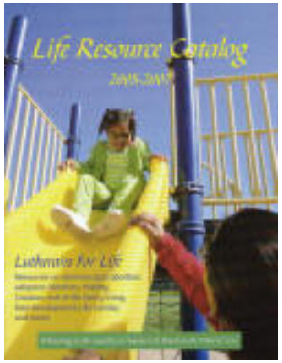
Каждый пред Богом наг.
Замок, наг и убог.
В каждой музыке Бах,
В каждом из нас Бог.
Ибо вечность — богам.
Брешься — удел быков...
Богово станет ^V каши
Сумерками богов.
И надо небом ^V дышать,
И, может быть, невпопад.
Ещё нас не раз распнут
И скажут потом: распад.

Other “Documents”









IBM Cross Pad Data

- 30 boxes, 30 writers producing 50-80 pages each
- 25000 pages total / 1 million words
- Most European Languages: German, French, Italian, English (UK), and Spanish
- Makeup: Characters (~8 boxes), Phrases, Freeform (1 box)
- Contracted with IBM to make the data public

FFX 115r

MELLO [redacted] NOT TO RUN FOR MAYOR OF FREMONT

Fremont Councilman Gary Mello closed the door Thursday on a mayoral bid, increasing the chances that Mayor Bill Ball will be re-elected to a second term in November.

Mello announced in March that he was considering a possible run for mayor, but he said subsequently on several occasions the chances were slim he would challenge Ball.

"After a long and difficult decision-making process, I have decided not to run for mayor of Fremont, at this time," Mello said in a written statement.

Mello's council seat does not expire until 1993.

In a telephone interview, the 42-year-old title insurance company executive said he did not want to devote more time to city business at the expense of his family and job. He said he currently spends about 30 hours a week on city-related [redacted] and that being mayor would mean at least 10 hours per week.

German - Basic

Bim, Bam, Bum - Ein Glockenton

Bim, Bam, Bum - Ein Glockenton

fliegt durch die Nacht, als

fliegt durch die Nacht, als

hätt' er Vogelflügel, er fliegt

hätt' er Vogelflügel, er fliegt

in römischer Kirchentracht

in römischer Kirchentracht

wohl über Tal und Hügel. Er

wohl über Tal und Hügel, Er

FFX 121J

GENEROSITY FINANCES MERCY MISSION TO FARMWORKERS

A silent disaster has stricken the people of California's Central valley, and two Fremont City Hall staffers are spearheading a drive to help.

The situation is stark.

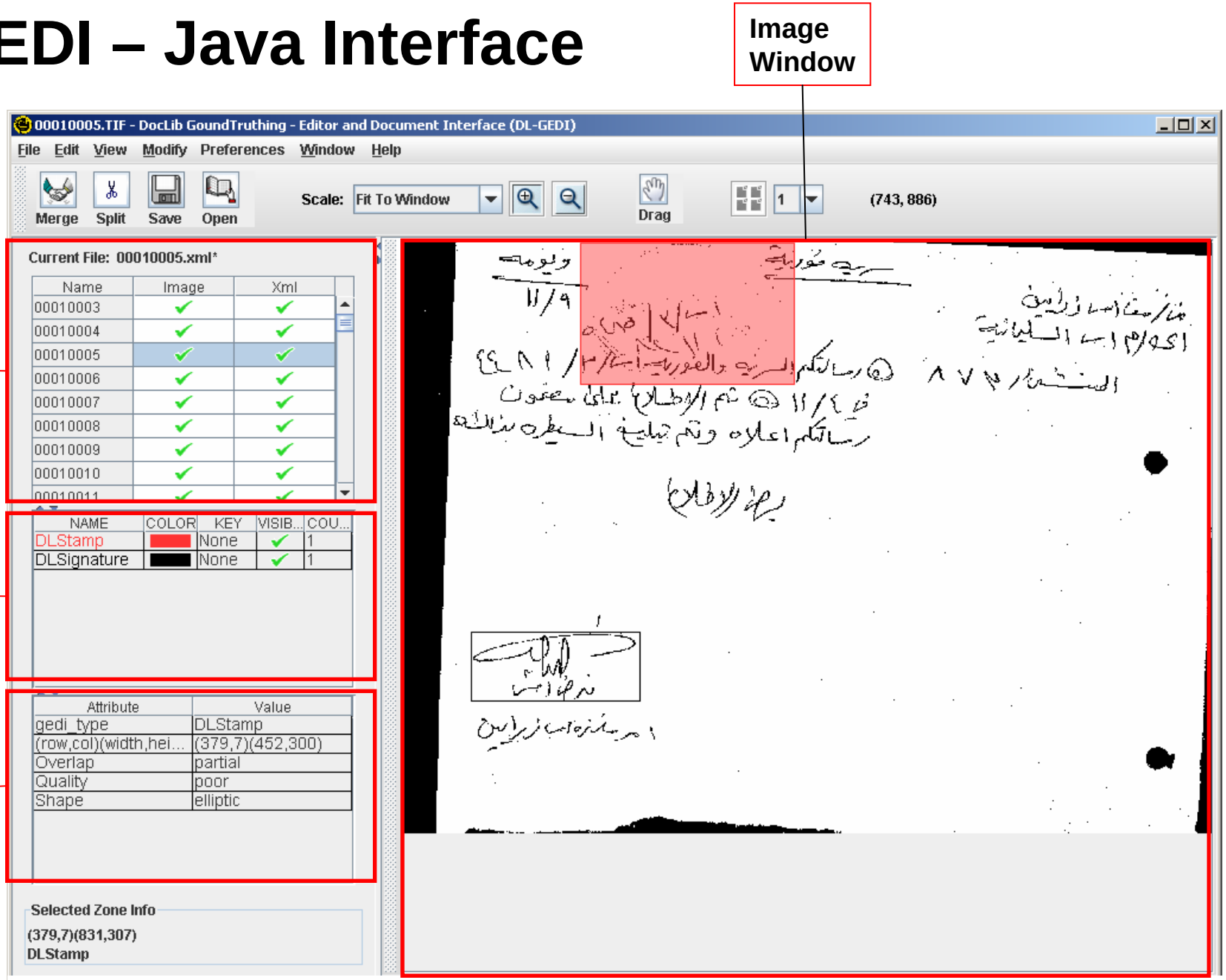
"People are going to starve to death," warned Fremont Mayor Bill Ball.

The Fremont effort is focused on the area around Firebaugh, in of California's largest olive processing plant and, according to city literature, the "Heart of the California Orange Area."

The severe winter wiped out the citrus crop throughout Firebaugh. Halfway between Fresno and Bakersfield, is running at 20% percent. Work will not be available in the area industry there until November.

On May 5, a Farmworker Relief Caravan from Fremont, San Francisco, El Cerrito, and Richmond delivered 12 tons of flour, two tons of rice, two tons of beans and hundreds of pounds of sugar, wood, cooking oil and used clothing.

GEDI – Java Interface



GEDI

- allows users to label and display rectangular zones in images
- supports user specified zone types
- handles type-specific attribute lists
- offers a graphical interface for editing and displaying zones
- enables users to create and distribute configuration files
- provides hotkeys for faster labeling
- can list multiple images in thumbnail views
- saves ground-truth and metadata as XML (compatible with DocLib)

Software Architeture

- **Efficient Technology Transfer**
  - **software compatibility**
  - **balance of academia, governemnt, and industry needs**
  - **common framework for document processing**

- **Scalability**
  - **rapid prototyping of new methods**
  - **simple algorithm comparison**

- **Robustness and Stability**
  - **high quality standards**
  - **platform-independence**
  - **accommodation of frequently changing requirements**

DocLib Status

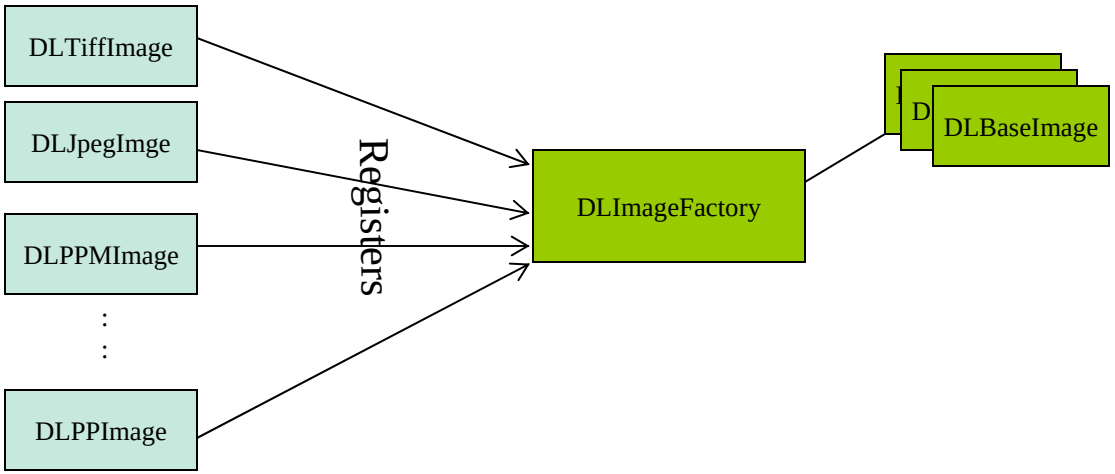
- Core DocLib components matured and stable (in use by a variety of government installations)\
- Addons being integrated/implemented, primarily by developers
- Freely available to government researchers
- Core supported on Solaris, Linux and Windows

Core vs Add-ons

- Core components are loosely defined as necessary building blocks for ANY document analysis process
- Addons are tools and applications for specific types of analysis

We try to put as few constraints on the representations as possible.

Image Factory



Design Factors:

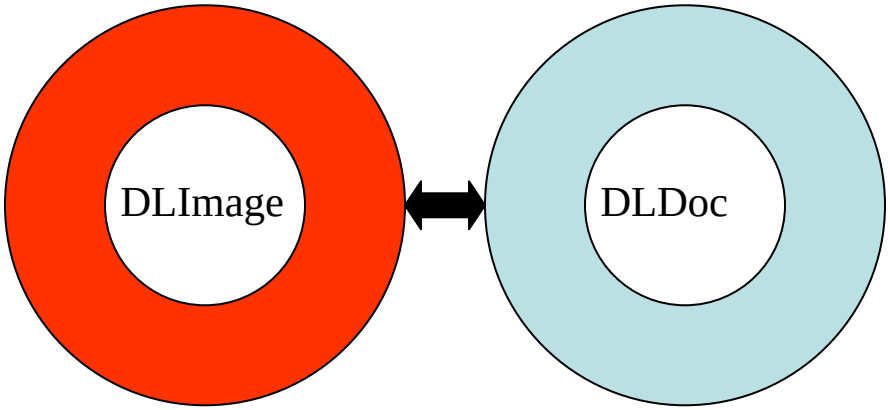
- **Image Type objects are static/singleton objects created on startup**
- **DLImageFactory is a static/singleton object**
- **Image Type objects registers itself with the DLImageFactory during startup**
- **DLImageFactory keeps a list of supported Image objects as each image type calls the register function**
- **Additional image types can be plugged into DOCLIB without modifying existing DOCLIB code.**

DocLib Architecture

DocLib´s architecture rests on two pillars:

DLImage:
- **Image Processing**

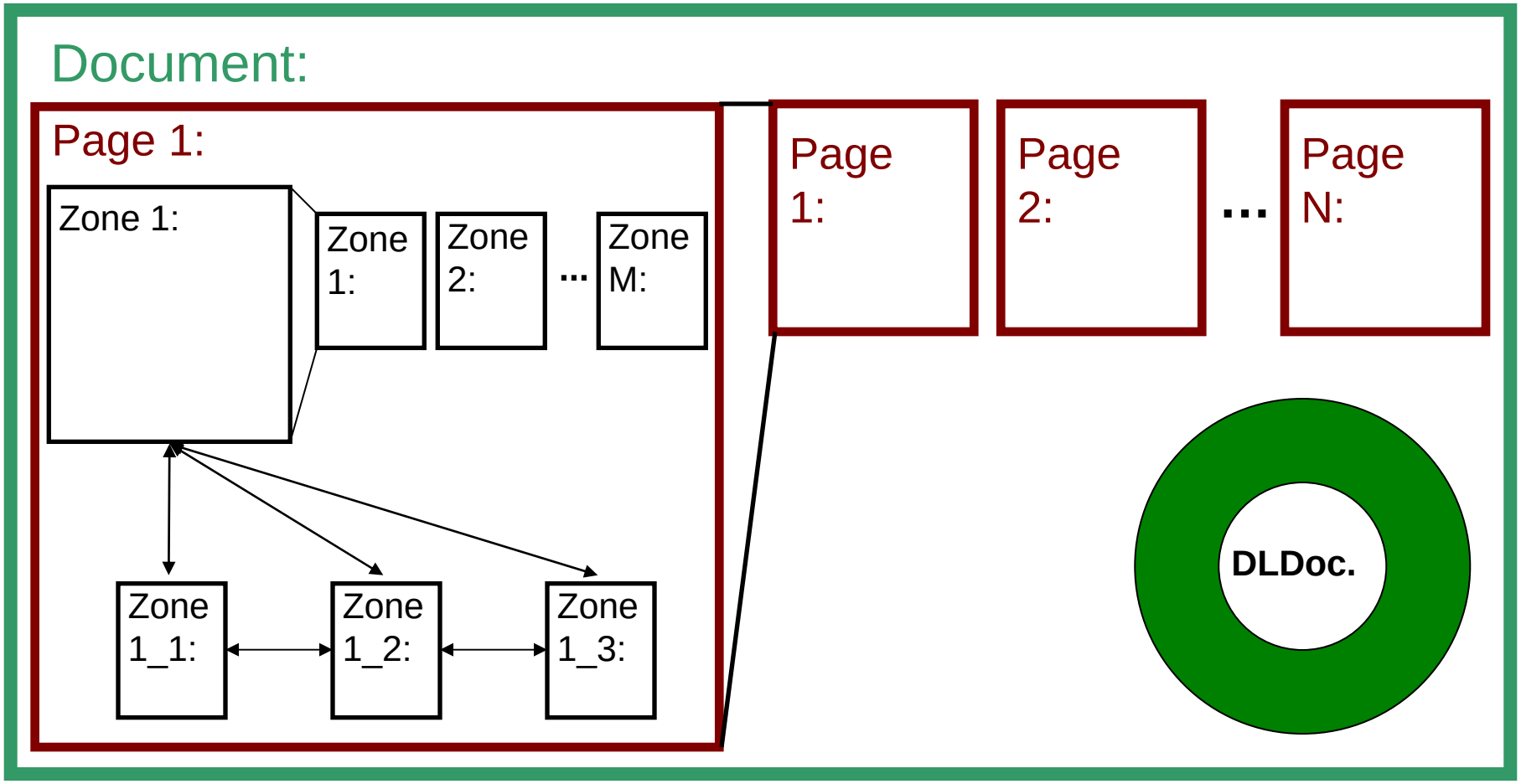
DLDocument:
- **Document Processing**



e.g.
- **image rotation**
- **image deskewing**
- **image conversions**
- **cc calculation**
- **shape drawing**

e.g.
- **page segmentation**
- **text line extraction**
- **logo detection**
- **XML input/output**
- **page layout analysis**

Document Hierarchy



Recent Modules

- Thinning
- Rotation
- Deskewing

- XML i/o
- Degradation
- OCR Scansoft interface (Windows)
- Docstrum

- Logo detection
- Signature processing

- LogoDetect
- TokenMatch
- Machine vs. Handwritten
- Jargon
- Text Line Detection

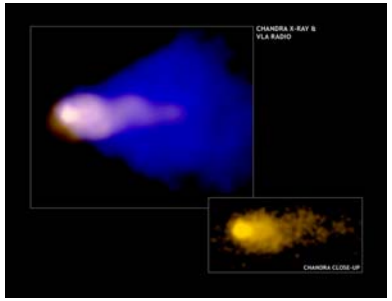
Logo and Stamp Detection

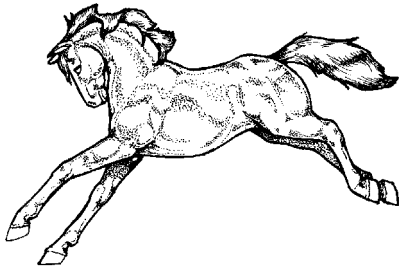
Document/Non-Document Discrimination

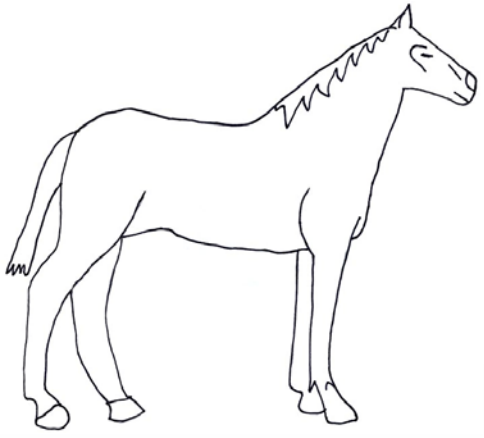
- Dataset
  - documents: 422
  - non-documents: 570

	Doc	Non-Doc
Doc	0.9031	0.0969
Non-Doc	0.0941	0.0941

Non-Documents





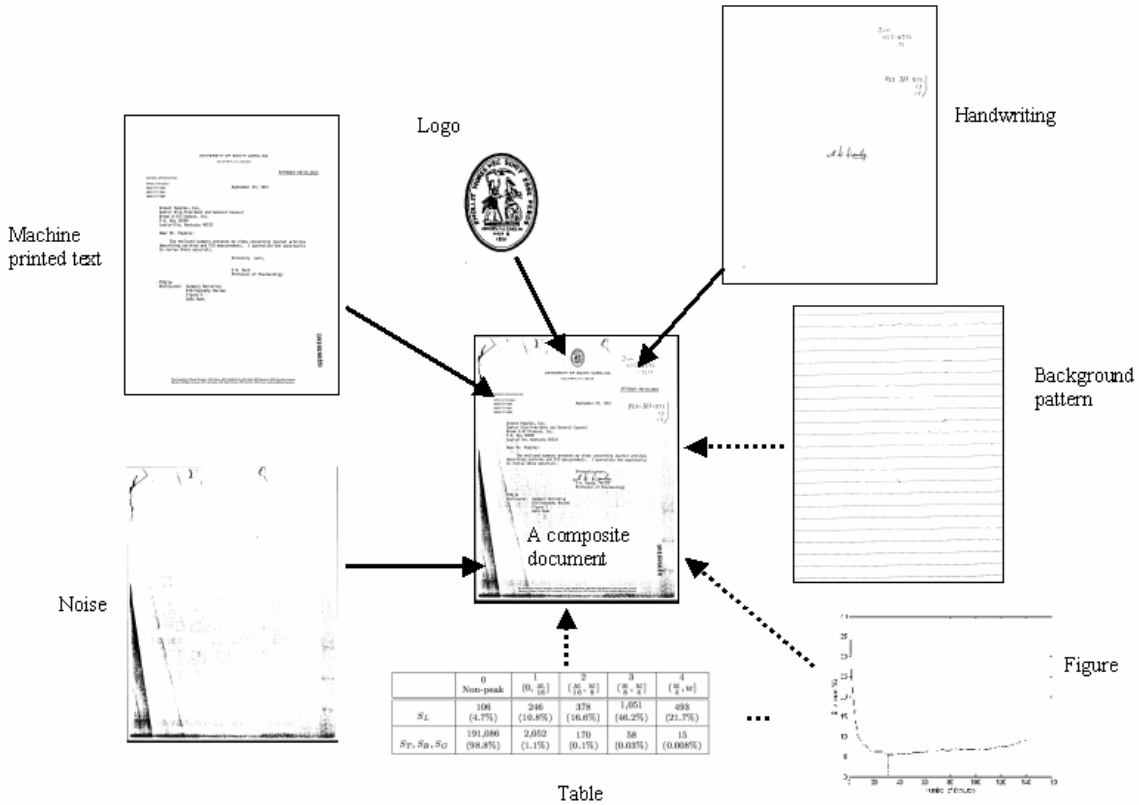


Original Challenges

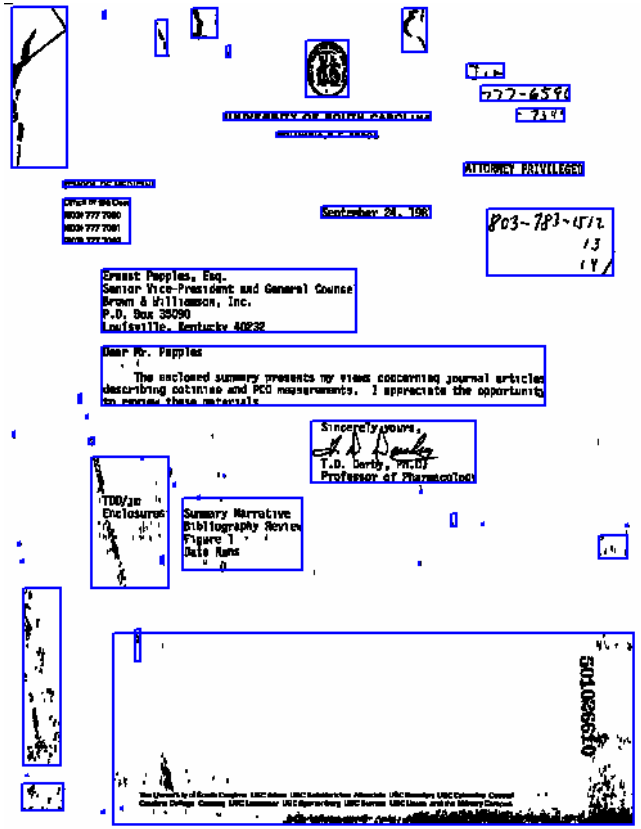
- Optical Character Recognition (OCR) does not work well in handwritten characters
- Word segmentation in handwritten is challenging
  - Words often touch each other
- Large variation exists even for the handwriting from the same person
- Handwriting often mixes with noise
  - Underground lines
  - Noise introduced by binarization

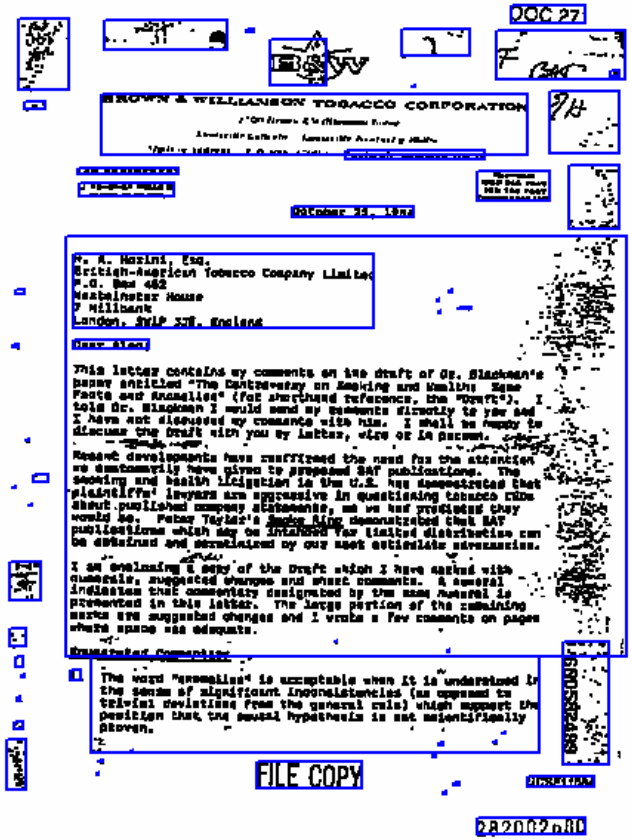
Introduction

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



Page Segmentation for Noisy Documents





* Docstrum page segmentation technique is used

Overview of Our Approach

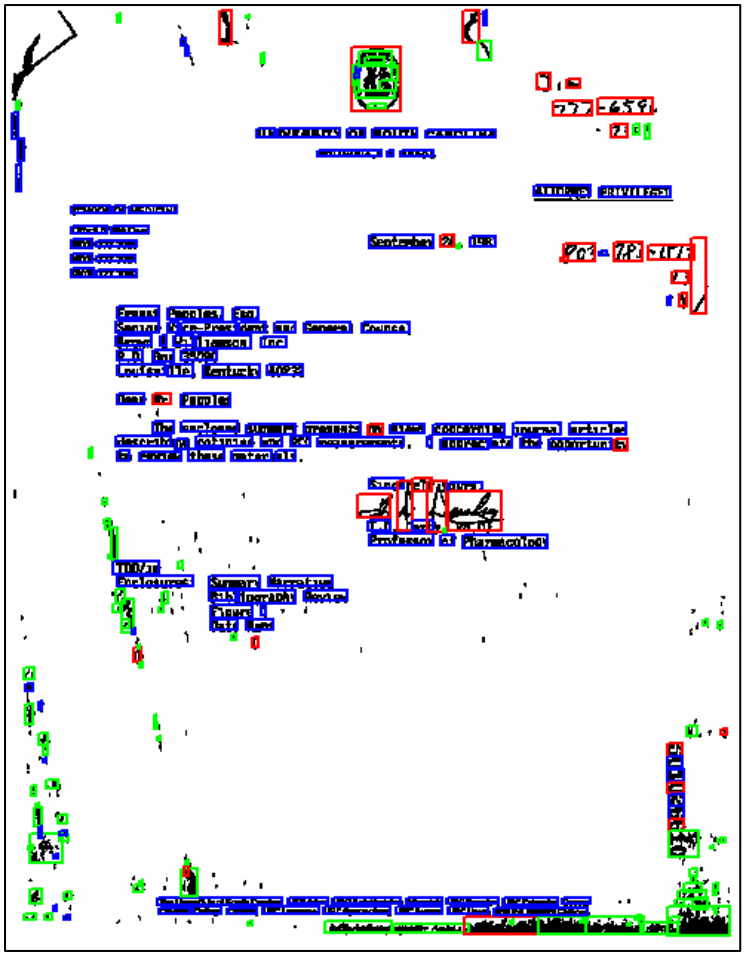
- Overview of our approach
  - Segment the document to word level using connected component based, bottom-up approach.
  - Classify each segmented block into noise, handwriting or printed text, based on extracted features and the Fisher classifier.
  - Using MRF (Markov Random Field) to refine the classification result.

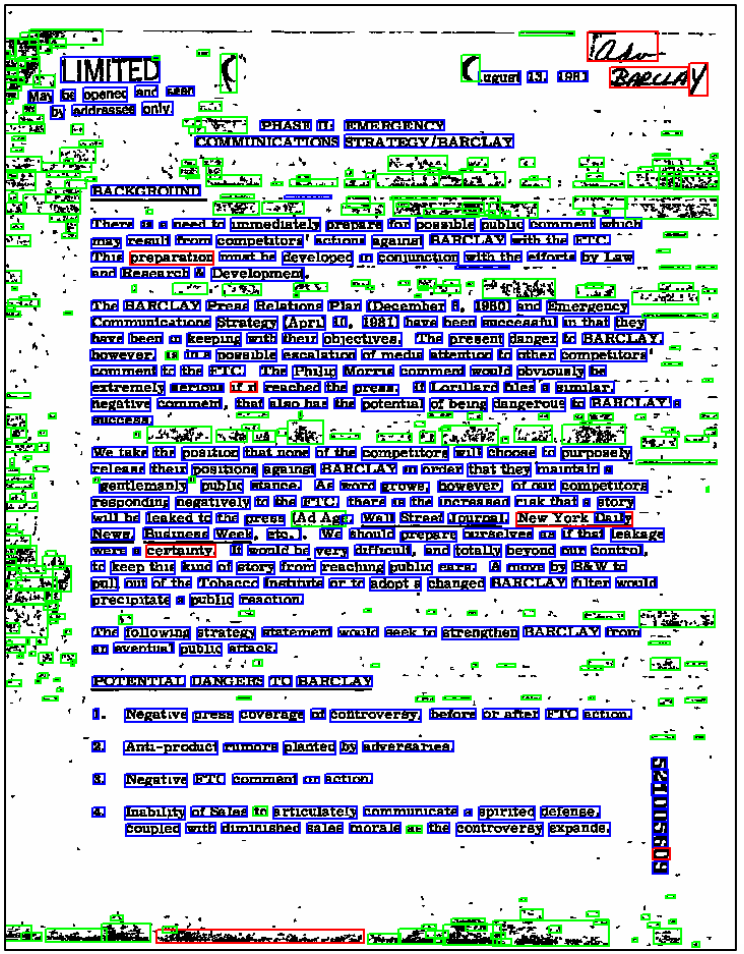
Classification Results with Fisher Classifier

Printed text

Handwriting

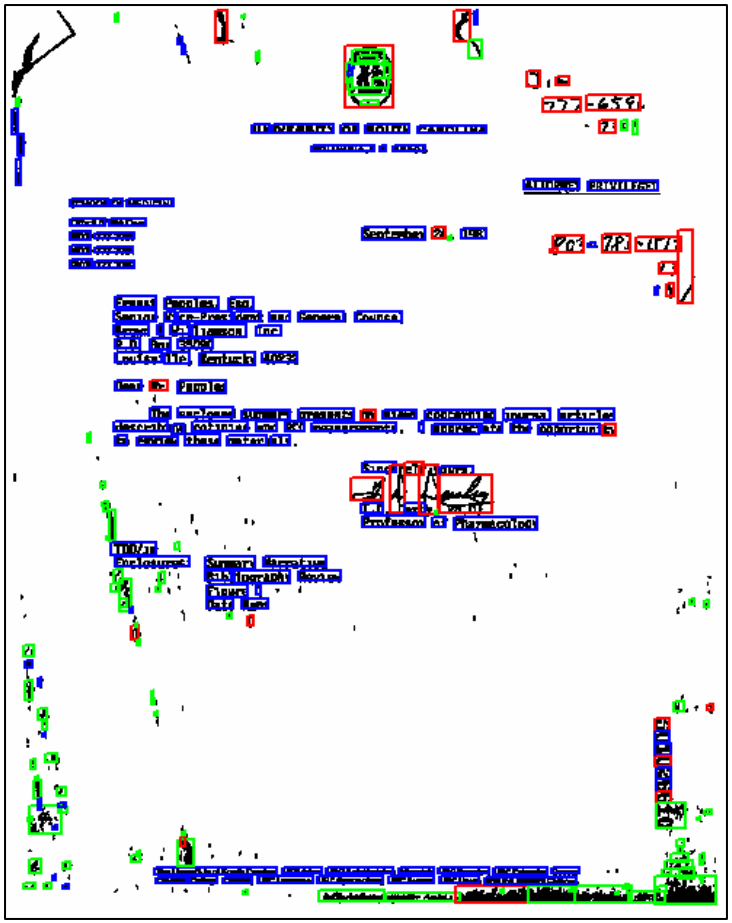
Noise



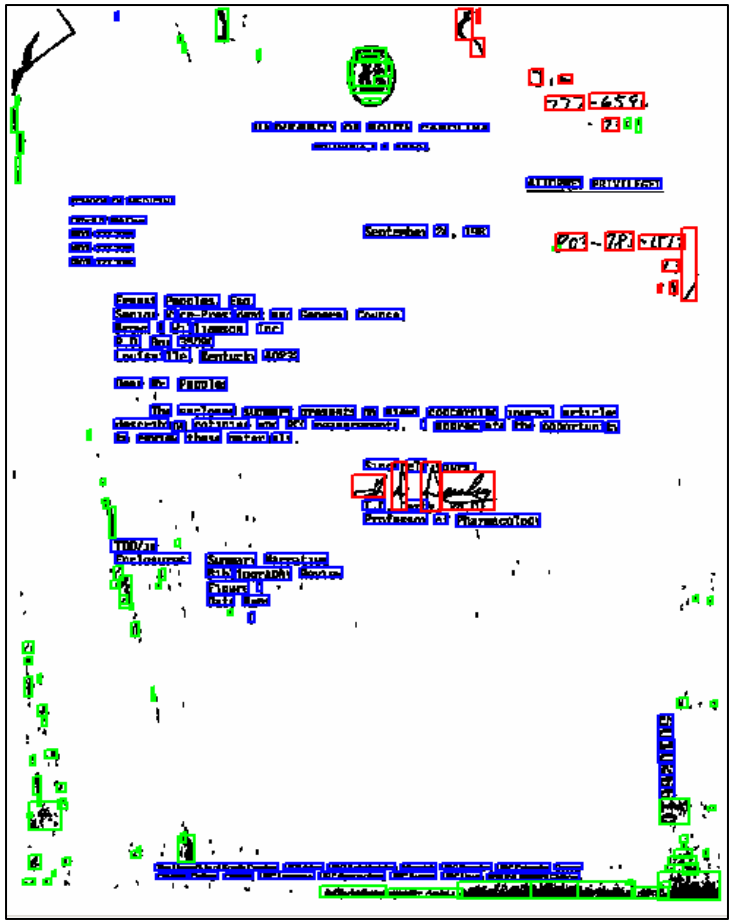


MRF Postprocessing Example

Printed text
Handwriting
Noise



Before MRF-based postprocessing



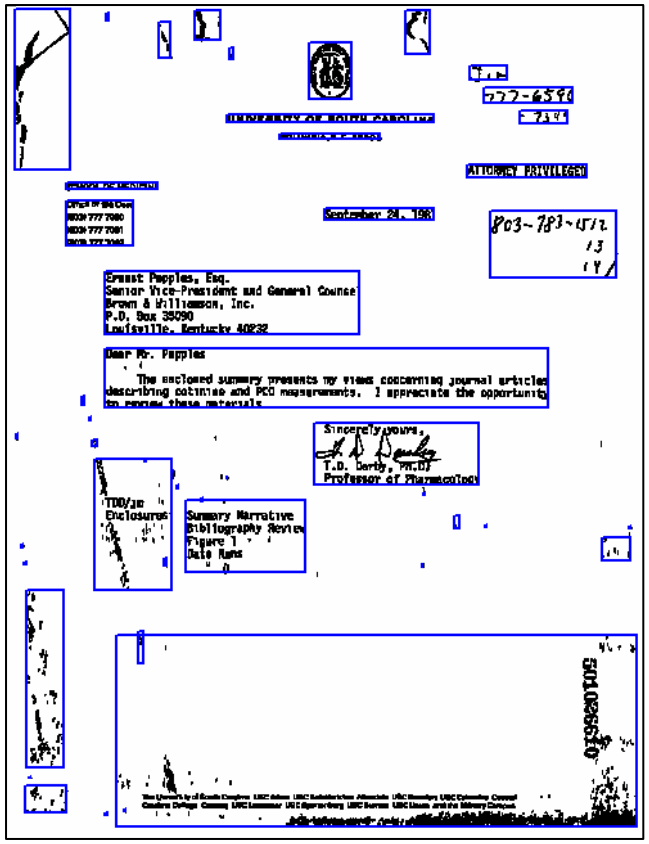
After MRF-based postprocessing

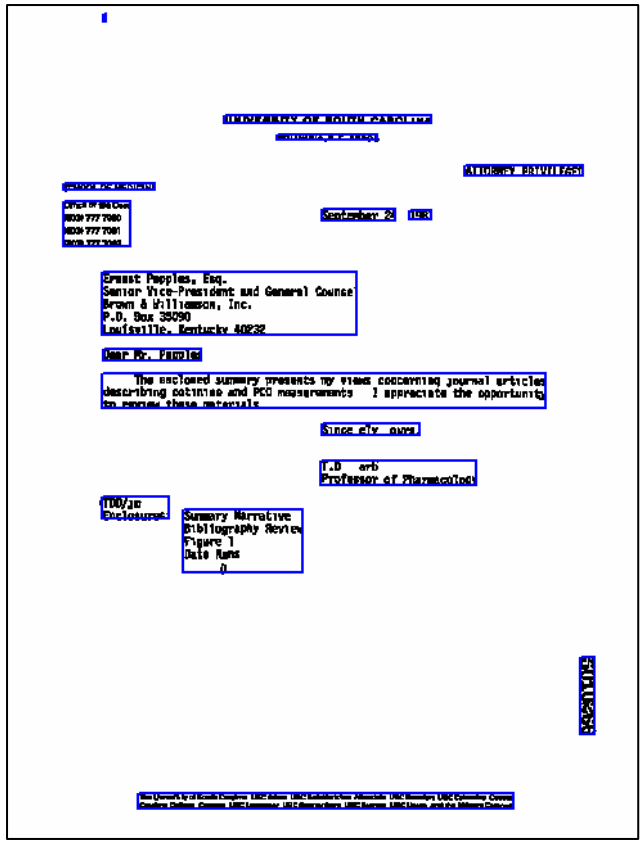
Evaluation

- Data Collection
  - 318 documents provided by the tobacco industry.
  - 94 documents of testing, the other for training.

	#Total	Percentage	Before Post-processing		After Post-processing	
			Accuracy	Precision	Accuracy	Precision
Printed Words	19,227	66.9%	95.9%	99.5%	98.0%	99.7%
Handwritten Words	701	2.4%	93.2%	62.9%	93.0%	83.3%
Noise Blocks	8,802	30.7%	96.8%	93.0%	98.6%	96.0%
Total	28,730	100%	96.1%	N/A	98.1%	N/A

Application to Page Segmentation

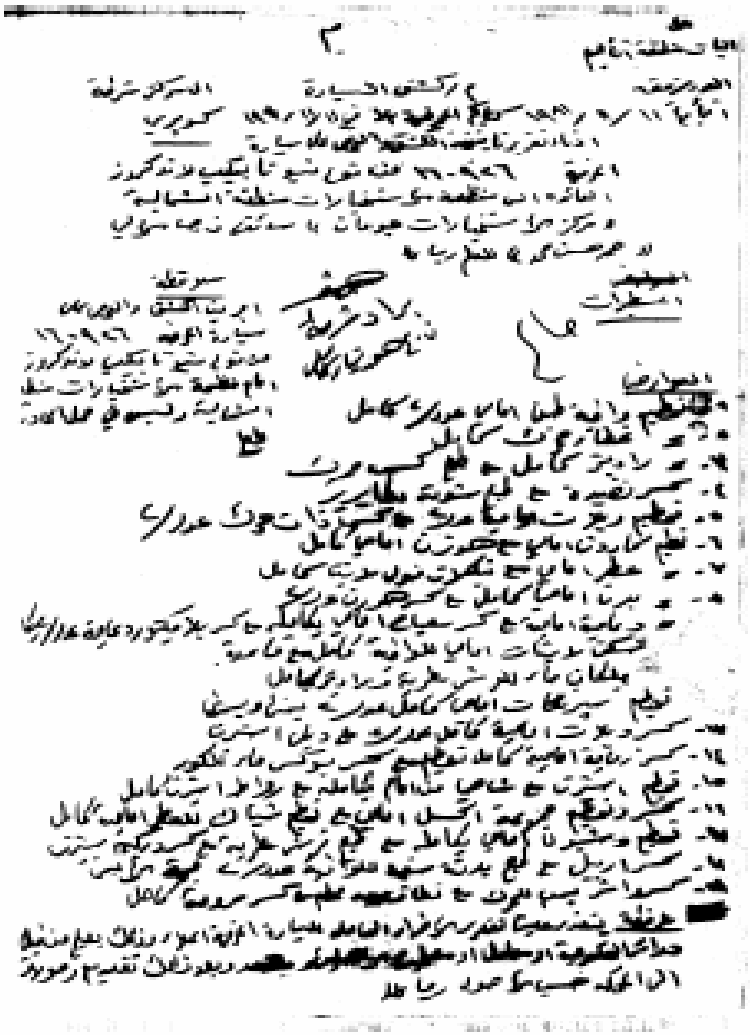


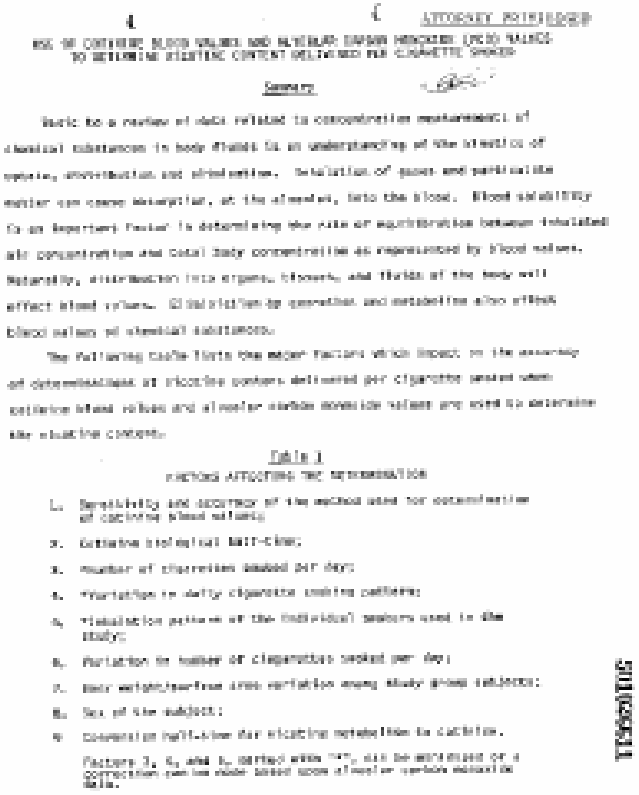


Before enhancement

After enhancement

Handwritten Line Detection





4 ATTORNEY DOCKET/DOCID

USE OF CERTAIN BLOOD VALUES AND ALVEOLAR CARBON MONOXIDE LEVELS TO DETERMINE SMOKING CONSTANT DELIVERED PER CIGARETTE SMOKED

Summary

Basic to a review of data relative to concentration mechanisms of chemical substances in body fluids is an understanding of the kinetics of uptake, distribution and elimination. Solution of gases and particulate matter can cause absorption, at the alveoli, into the blood. Blood solubility is an important factor in determining the rate of equilibration between inhaled air concentration and total body concentration as represented by blood values. Naturally, biotransformation into organs, tissues, and fluids of the body will affect blood values. Elimination by excretion and metabolism also affect blood values of chemical substances.

The following table lists the major factors which impact on the accuracy of determinations of nicotine content delivered per cigarette smoked when cotinine blood values and alveolar carbon monoxide values are used to determine the nicotine content.

Table I

FACTORS AFFECTING THE DETERMINATION

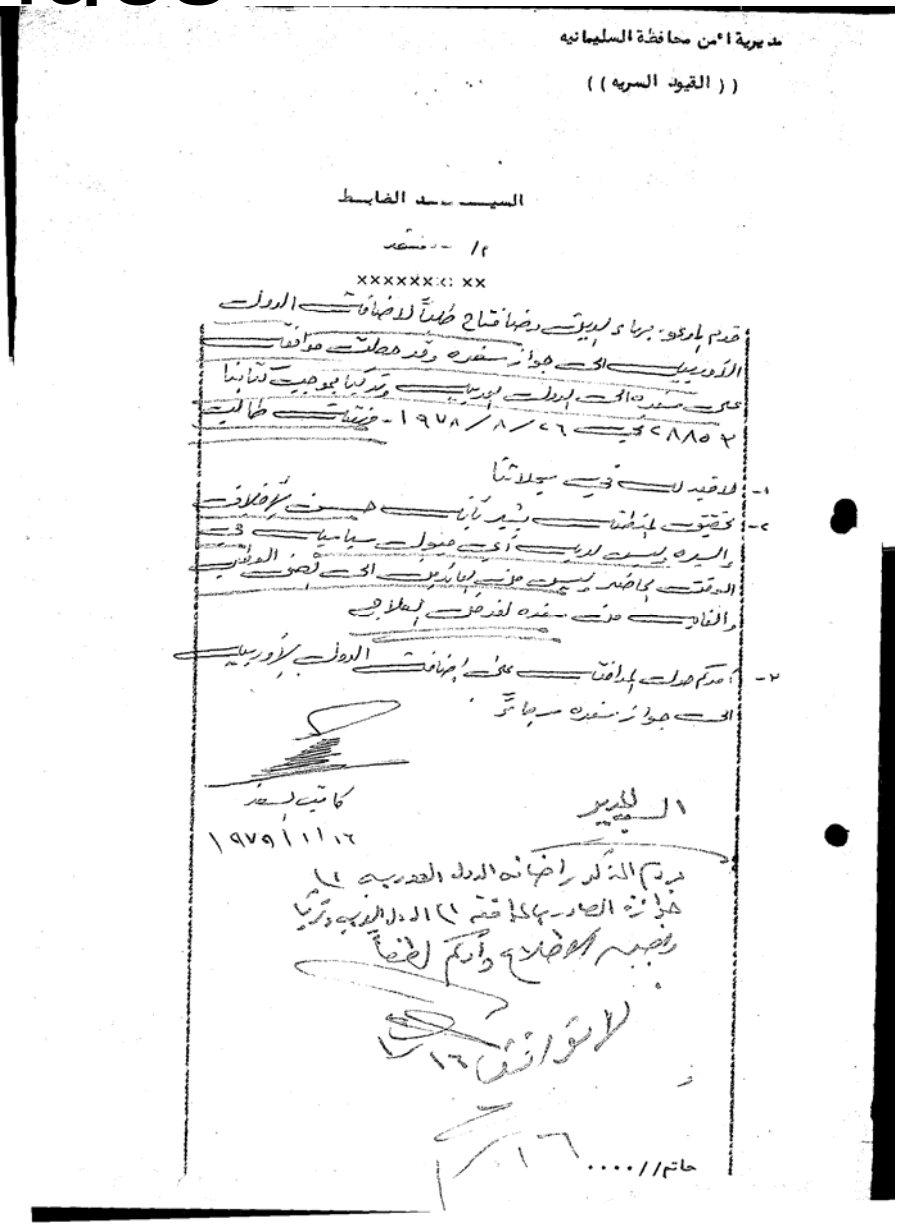
1. Sensitivity and accuracy of the method used for concentration of cotinine blood values;
2. Cotinine biological half-life;
3. Number of cigarettes smoked per day;
4. Variation in daily cigarette cotinine patterns;
5. Inhalation patterns of the individual smokers used in the study;
6. Variation in number of cigarettes smoked per day;
7. Body weight/surface area variation among study group subjects;
8. Sex of the subjects;
9. Conversion half-life for nicotine metabolism to cotinine.

Factors 3, 4, and 5, marked with "*", can be minimized or eliminated by correction factors made based upon alveolar carbon monoxide data.

5010696S1

Challenges

- Handwriting
  - Curvilinear
  - Touching lines, words, characters
  - Non-Manhattan layout
  - Multiple orientation
- Machine Printed
  - Straight
  - Significant Gap
  - Regular layout



Our Assumptions

- Text Lines are generally horizontal
- Text line orientation may vary slightly
- There are gaps between neighboring lines
  - Strong for machine printed
  - Weak for handwriting

Intuitive Approach

- Preprocessing
  - Gaussian Blurring using rectangular window
- Segmentation: The Level Set Methods
  - Enforce merging in horizontal direction
- Localization
  - Group nearby isolated connected components

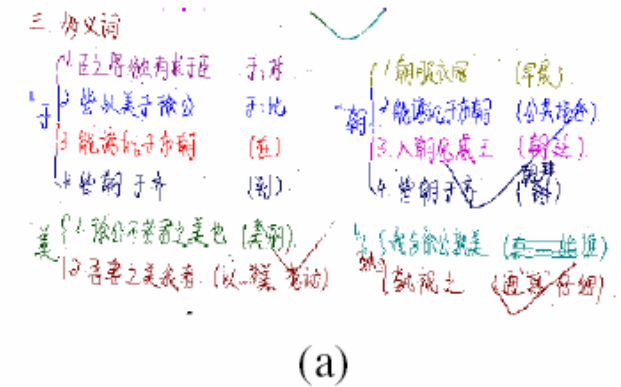
The level set method

$$\frac{\partial f}{\partial t} + \underbrace{\vec{S} \cdot \nabla f}_{\text{Vector Field Based}} + \underbrace{V_N \left|\nabla f\right|}_{\text{In Normal Direction}} = \underbrace{b\kappa \left|\nabla f\right|}_{\text{Curvature Based}}$$

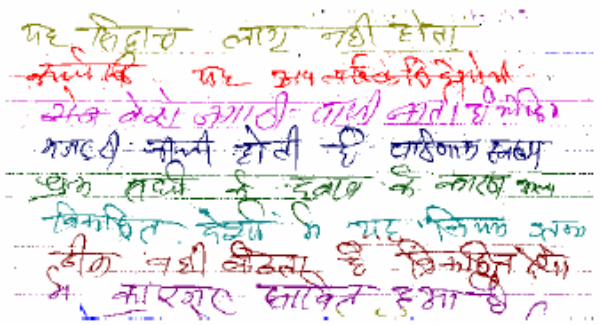
http://vision.ece.ucsb.edu/~sumengen/level_set_methods

- The moving speed df/dt is changed by forces based on curvature, normal direction or external vector field

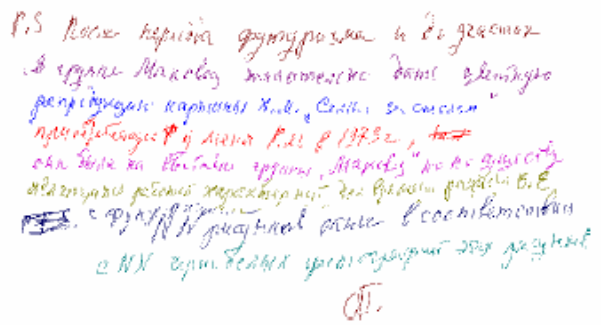
Results



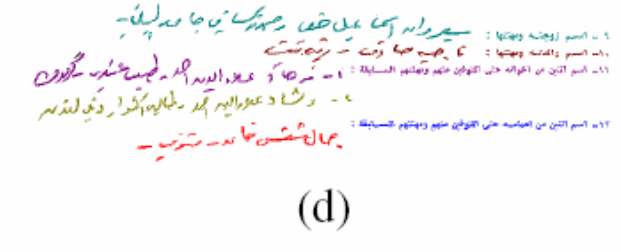
(a)



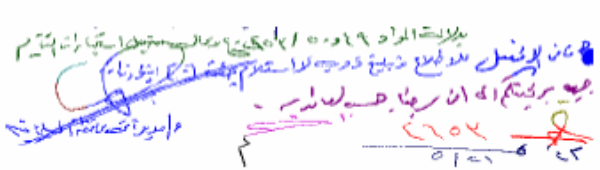
(b)



(c)



(d)



(e)

(f)

Evaluation

- Ground truth
  - VIPER Ground Truth Editor
  - Pixel level: use polygon to represent text lines
- Evaluation method
  - The Hungarian algorithm
  - One-to-one correspondence
- Statistics
  - Four scripts, 100 documents per script
  - 5000 text lines

Comparison

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100 HANDWRITTEN HINDI DOCUMENTS.

	Hit Rate	STD of Hit Rate	Detected Text Lines (Total: 1,365)
Our Method	97%	0.02	1,295 (95%)
Projection Profile based method	43%	0.37	673 (49%)
Improved connected component	78%	0.19	1,103 (80%)
Docstrum based method	74%	0.22	996 (72%)

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100 HANDWRITTEN CHINESE DOCUMENTS.

	Hit Rate	STD of Hit Rate	Detected Text Lines (Total: 1,672)
Our Method	98%	0.01	1,532 (92%)
Projection Profile based method	58%	0.24	965 (57%)
Improved connected component	64%	0.12	1,131 (67%)
Docstrum based method	94%	0.04	1,389 (83%)

Comparison (cont')

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100 HANDWRITTEN WITH BACKGROUND PARALLEL LINES.

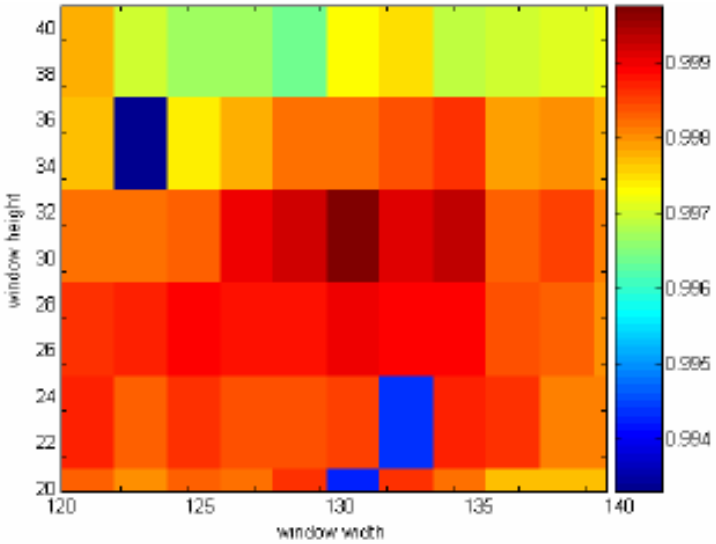
	Hit Rate	STD of Hit Rate	Detected Text Lines (Total: 1,589)
Our Method	98%	0.01	1,466 (93%)
Projection Profile	79%	0.10	1,178 (74%)
Improved connected component	54%	0.34	849 (53%)
Docstrum based method	62%	0.29	1,024 (64%)

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100 FREESTYLE HANDWRITTEN DOCUMENTS.

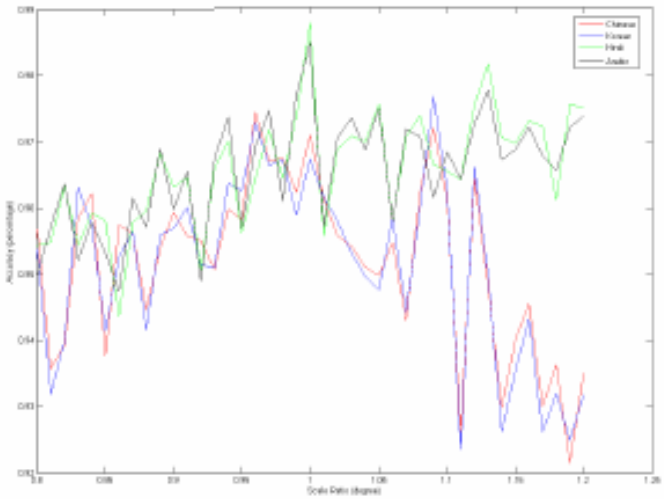
	Hit Rate	STD of Hit Rate	Detected Text Lines (Total: 2,178)
Our Method	93%	0.03	1,863 (85%)
Projection Profile	72%	0.16	1,537 (70%)
Improved connected component	81%	0.13	1,576 (73%)
Docstrum based method	82%	0.15	1,751 (80%)

Robustness Test

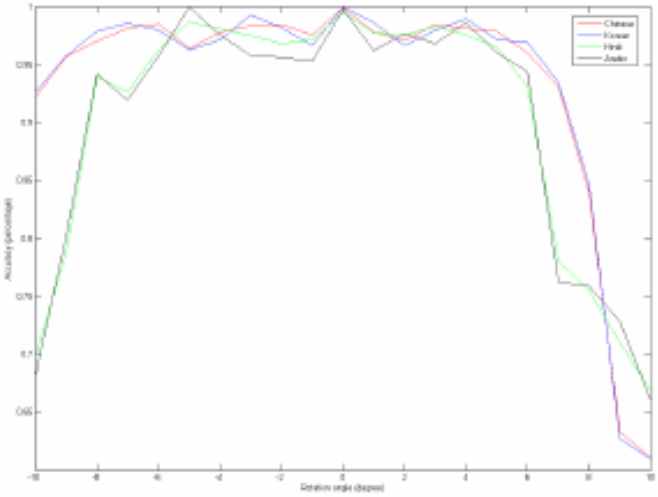
- Change of internal parameter
- Change of input scale
- Rotate input image
- Corrupt input with Noise



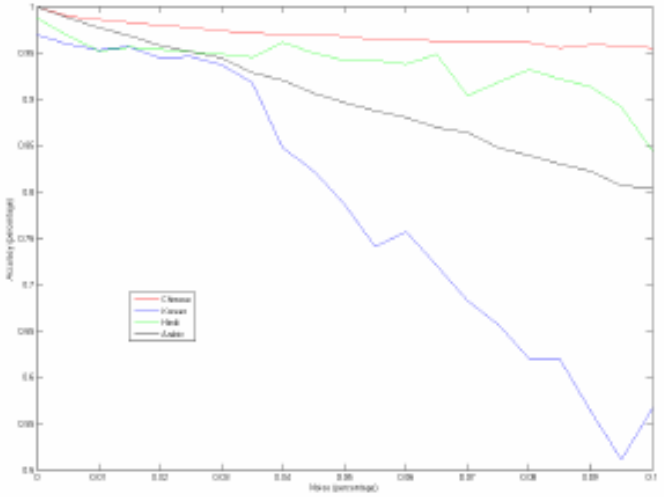
(a)



(b)



(c)



(d)

Fig. 6. Robustness test. (a) Accuracy of different parameters. (b) Accuracy of different scales. (d) Accuracy of different rotation angles. (d) Accuracy of different noises

Publications

- Yi Li, Yefeng Zheng and David Doermann. Detecting Text Line in Handwritten Documents. ICPR'06,
- Yi Li, David Doermann, Stefan Jaeger, and Yefeng Zheng . A new algorithm and its implementation in Detecting Text Line in Handwritten Documents. IWFHR'06, 2006.
- Yi Li, Yefeng Zheng, Stefan Jaeger, and David Doermann. A new algorithm and its evaluation in Detecting Text Line in Handwritten Documents. UMD Technical Report. (in preparation)

DVD

- Dataset without Ground Truth
- Slides from Today

Agenda

Project Overview
- Introduction
- Goals and Objectives

Data Collection and Ground Truth

GEDI and Evaluation Framework

Evaluation and Research Components
- Unconstrained Signature and Logo Detection and Matching for Off-line Document Image Retrieval
- Document/Non-Document Discrimination

Technical Presentations
- **Adaptive OCR**
- **Document Classification by Layout**
- **Learning from Web for Shape-based Object Recognition**