CLEAT:

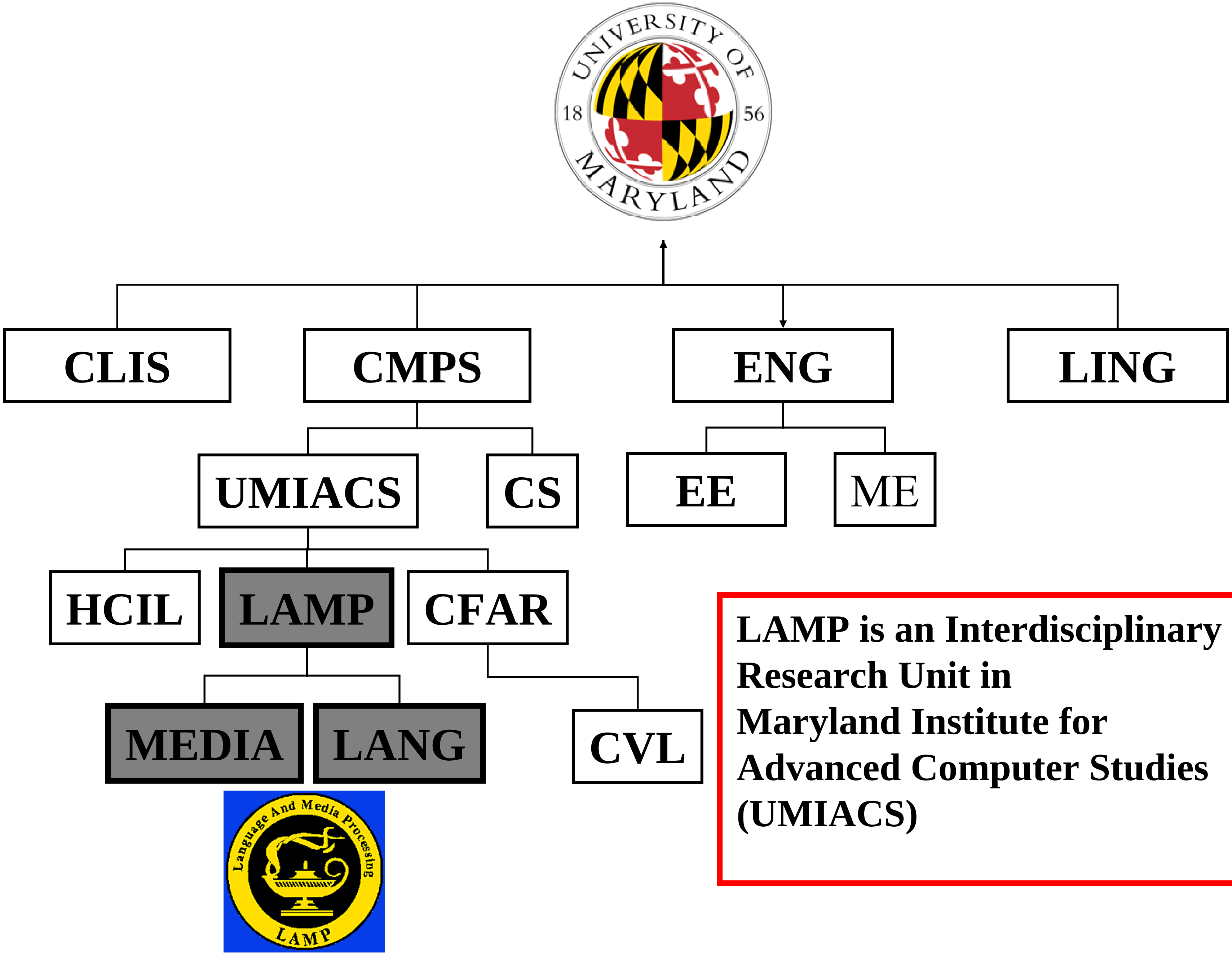
A CLassification, Enhancement and Analysis Toolkit for Heterogeneous Document Image Collections

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)

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

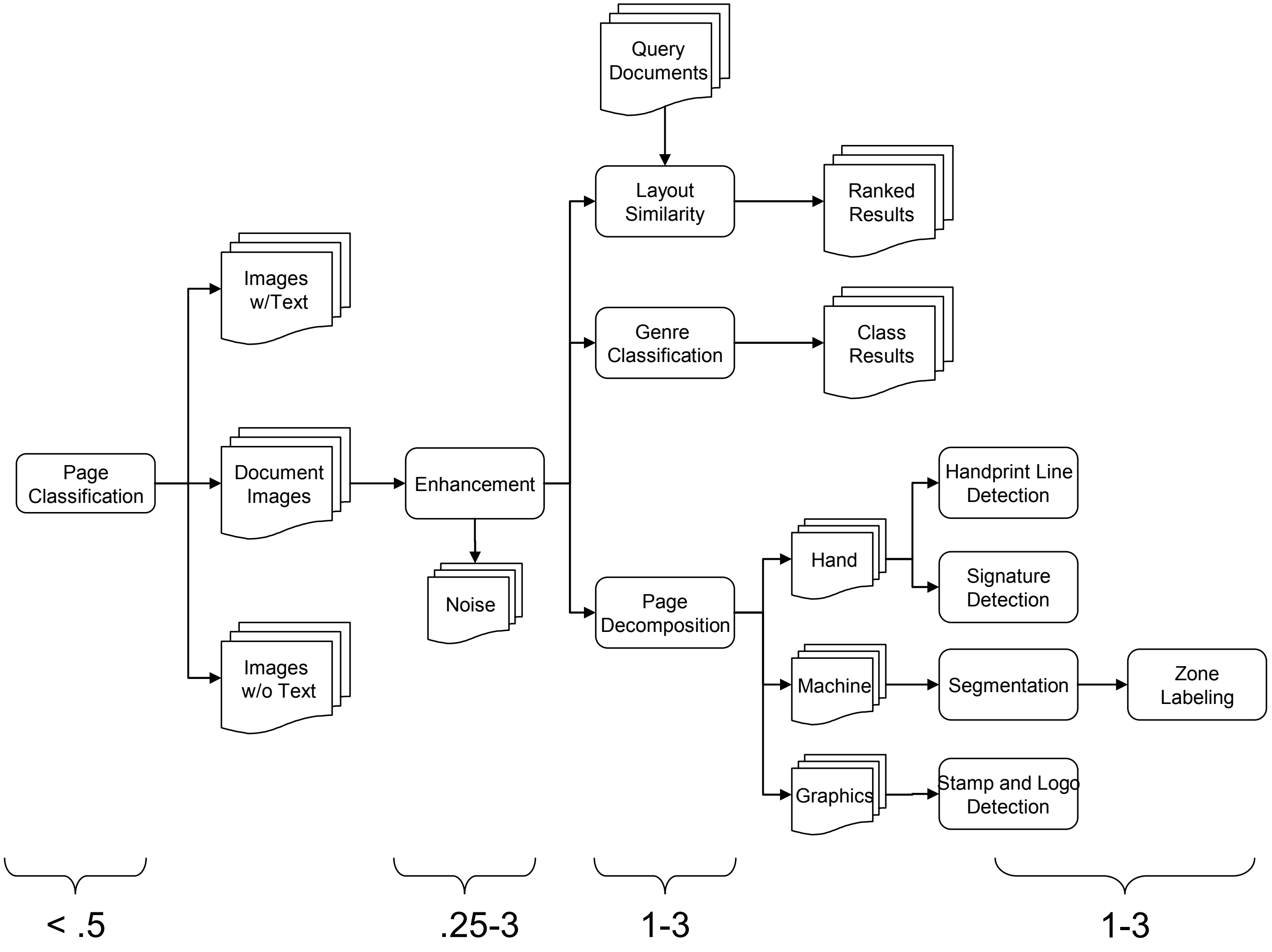
Schedule

- Tuesday
  - AM: Project Overview and Status
  - PM: Logo Detection/Recognition
- Wednesday
  - AM: Font OCR, Word Level ScriptID
    Vision Related Research
  - PM: Review and Feedback

Agenda
Tuesday AM

- Project Overview
  - Introduction
  - Goals and Objectives
- Tools
  - GEDI Display Environment
  - Datasets
- DocLib and Algorithms
  - Technical Presentations

Project Overview



Page Classification

Images w/Text

Document Images

Images w/o Text

Enhancement

Noise

Query Documents

Layout Similarity

Ranked Results

Genre Classification

Class Results

Page Decomposition

Hand

Machine

Graphics

Handprint Line Detection

Signature Detection

Segmentation

Zone Labeling

Stamp and Logo Detection

< .5

.25-3

1-3

1-3

Target Processing Speed in Seconds

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 I – March 2007

- Delivered complete CLEAT data collection.
- Provide ground truth for subset of data including signatures, stamps, logos, handwritten, and machine printed text.
- Provide document describing evaluation framework.

Phase II – July 2007

- Deliver completed ground truthing and visualization tool for CLEAT metadata.
- Deliver Prototype version of CLEAT Software API Modules:
  - Document Image Enhancement,
  - Document Text/Image Text/Non-Text Discrimination,
  - Page Layout Similarity Ranking on CLEAT data,
  - Page Layer Segmentation and Zone Labeling, and
  - Content Labeling of Signatures, annotations, Stamps and Logos.
- Provide results of CLEAT API run on CLEAT datasets.
- Provide preliminary evaluation report.
- Provide basic API documentation

Phase III

- Deliver Final version of CLEAT API.
- Provide training on use of CLEAT.
- Provide complete evaluation results on CLEAT data.
- Provide complete documentation of API.
- Provide feasibility report for system extensions.
- Provide a list of publications generated and planned as a result of this effort.

WWW

- lamp.cfar.umd.edu/media/projects/cleat
- Contains
  - Summary
  - Proposal
  - Reports
  - Presentations
  - Milestones and Deliverables
  - Software
  - Data

Agenda
Tuesday AM

- Project Overview
  - Introduction
  - Goals and Objectives
- Tools
  - GEDI Display Environment
  - Datasets
- DocLib and Algorithms
  - Technical Presentations

GEDI – Java Interface

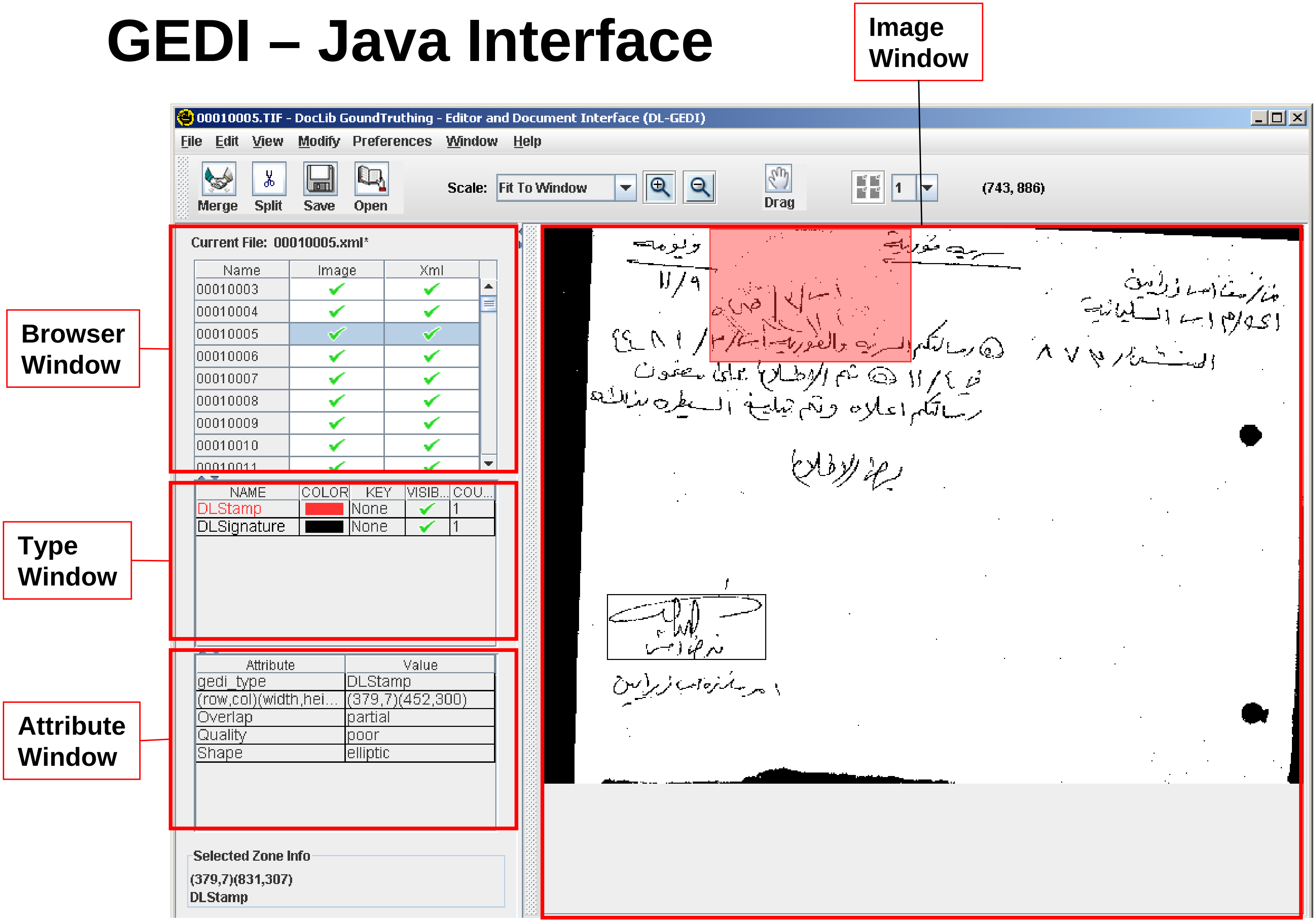


Image Window

Browser Window

Type Window

Attribute Window

Current File: 00010005.xml*

Name	Image	Xml
00010003	✓	✓
00010004	✓	✓
00010005	✓	✓
00010006	✓	✓
00010007	✓	✓
00010008	✓	✓
00010009	✓	✓
00010010	✓	✓
00010011	✓	✓

NAME	COLOR	KEY	VISIB...	COU...
DLStamp	🟥	None	✓	1
DLSignature	⬛	None	✓	1

Attribute	Value
gedi_type	DLStamp
(row,col)(width,hei...	(379,7)(452,300)
Overlap	partial
Quality	poor
Shape	elliptic

Selected Zone Info

(379,7)(831,307)
DLStamp

GEDI Features

- 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)

New Features

- Polygon and Oriented Boxes
- Scripting
- Text Alignment
- Multilingual support
- Additional Function Keys
- Bug Fixes

Data Collection

Datasets and Ground Truth – A dataset containing examples of each class of document we process will be included. The dataset will contain a minimum of 5000 documents and be collected from a variety of sources, including the internet, existing training and testing datasets, public collections, project collections, and scanning. All ground truth will be provided in GEDI format and accompany the images

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

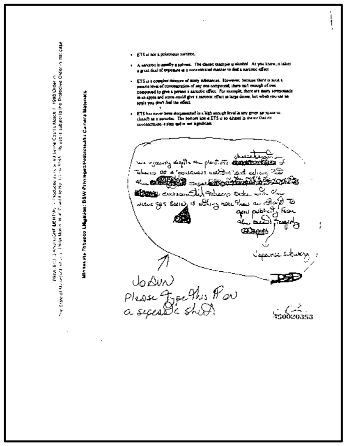
Document Image Acquisition

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

Document Genres

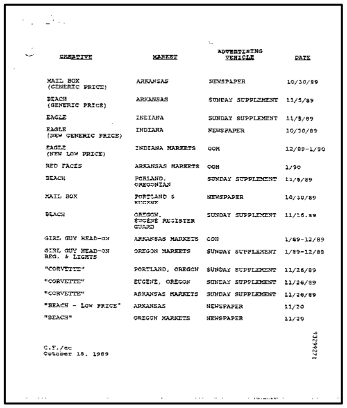
Genre	
Forms, Drawing, Tables et at.	
Forms	650
Drawing	80
Tables	100
Chemistry formulae	25
Math equations	165
Figures	40
Total	1060
Business documents and Memo letters	
Business documents	50
Business documents degraded	2700
Business documents with annotations	160
Memo letters	900
Total	3810
Journal 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



Jo Dun
Please type this Rev
a separate sheet

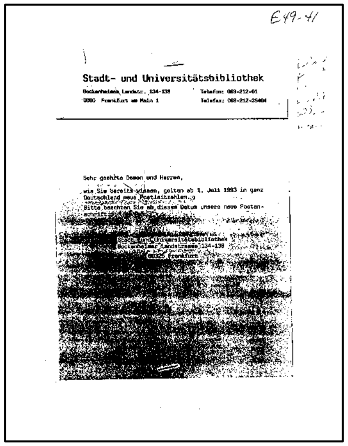
2500020353



CREATIVE	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	PORLAND, OREGONIAN	SUNDAY SUPPLEMENT	11/5/89
MAIL BOX	PORTLAND & EUGENE	NEWSPAPER	10/30/89
BEACH	OREGON, EUGENE REGISTER GUARD	SUNDAY SUPPLEMENT	11/15.89
GIRL GUY HEAD-ON	ARKANSAS MARKETS	OOH	1/89-12/89
GIRL GUY HEAD-ON REG. & LIGHTS	OREGON MARKETS	SUNDAY SUPPLEMENT	1/89-12/88
"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.F./ec
October 18, 1989

93296271



E49-41

Stadt- und Universitätsbibliothek

Bockenheimer Landstr. 134-138
6000 Frankfurt am Main 1

Telefon: 069-212-01
Telefax: 069-212-25404

Sehr geehrte Damen und Herren,

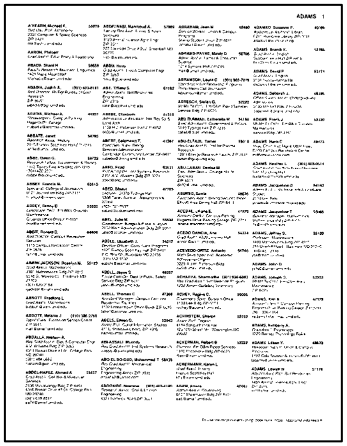
wie Sie bereits wissen, gelten ab 1. Juli 1993 in ganz Deutschland neue Postleitzahlen.

Bitte beachten Sie ab diesem Datum unsere neue Postanschrift:

Stadt- und Universitätsbibliothek
Bockenheimer Landstraße 134-138
Frankfurt

A a

aardvark



ADAMS 1

The Washington Post

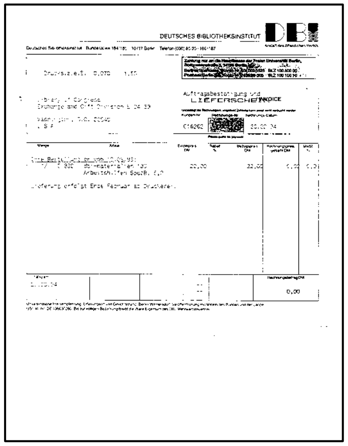
After Zarqawi, No Clear Path In Weary Iraq

The Young Apprentice

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

Kaine Delays Execution Of Inmate for 6 Months

FDA Approves Vaccine That Should Prevent Most Cervical Cancers



DEUTSCHES BIBLIOTHEKSINSTITUT

DBI

LIEFERSCHEIN

THE WORLD'S RICHEST PEOPLE

SPECIAL ISSUE

Forbes

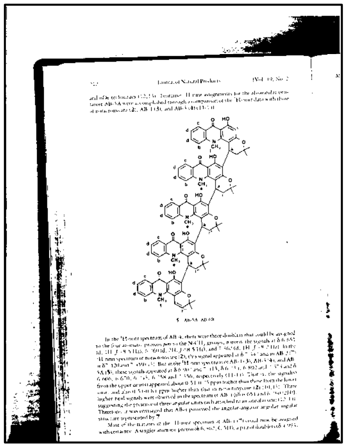
BILLIONAIRES

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

ALSO 946 BILLIONAIRES

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

INDIA UPSTAGES JAPAN



Journal of Natural Products

FINANCIAL TIMES

Collateral damage

Iraq braced as Saddam rejects Bush ultimatum

Fed holds rates but signals uncertain future

US probes Ahold collusion claims

Chopard

Language And Media Processing

LAMP

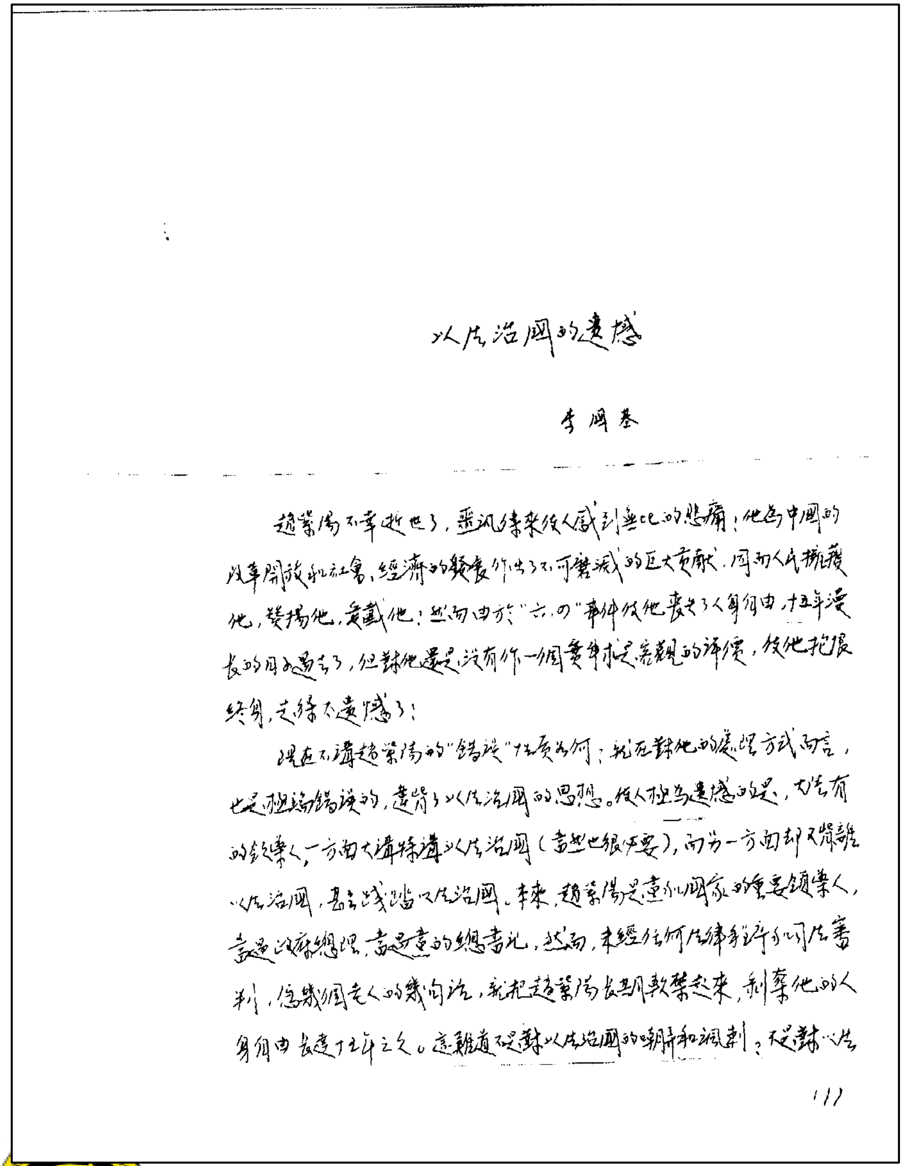
Internet Downloads

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



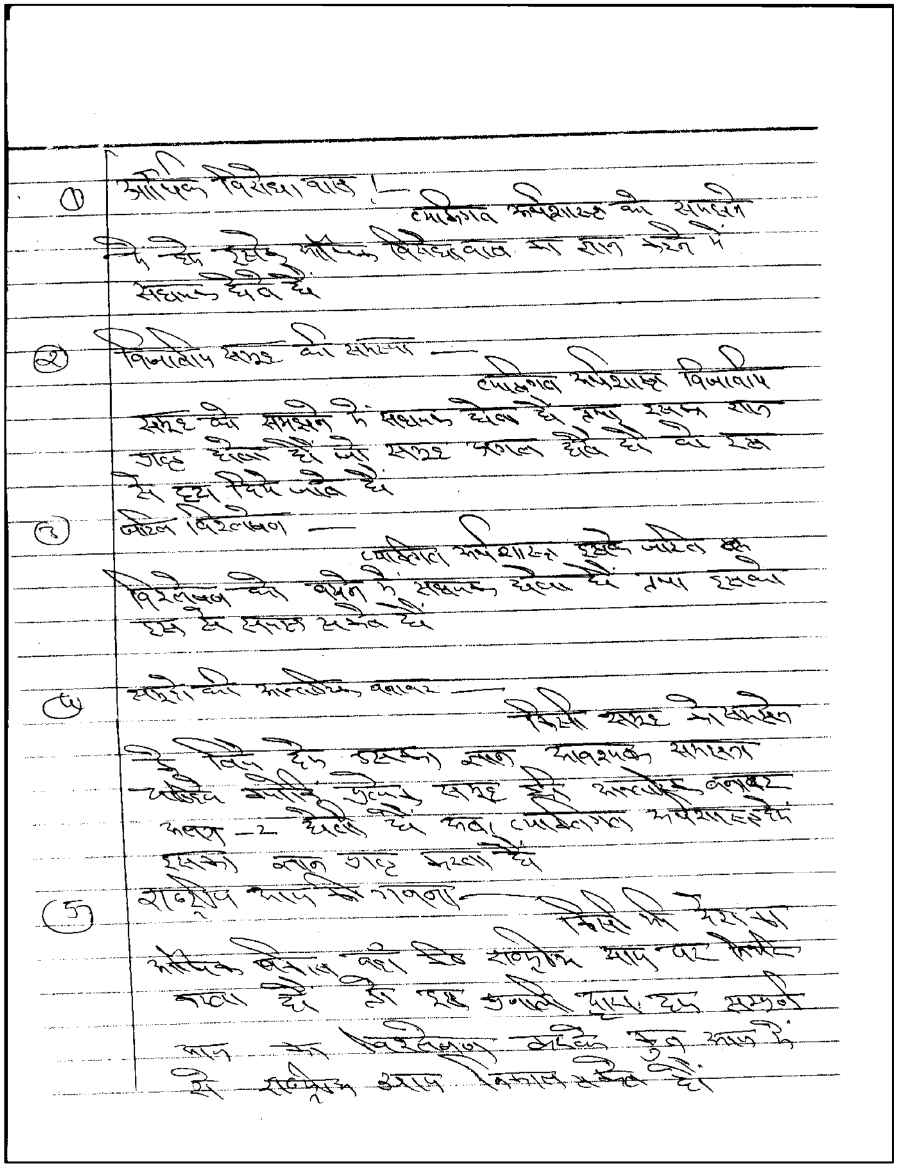
以法治國的遺憾

李國基

趙紫陽不幸逝世了，噩訊傳來使人感到無比的悲痛！他為中國的改革開放和社會、經濟的發展作出了不可磨滅的巨大貢獻，因而人民擁護他，贊揚他，愛戴他！然而由於"六四"事件使他喪失了人身自由，十五年漫長的日子過去了，但對他還是沒有作一個實事求是客觀的評價，使他抱恨終身，真係太遺憾了！

現在不講趙紫陽的"錯誤"性質如何；就在對他的處理方式而言，也是極端錯誤的，違背了以法治國的思想。使人極為遺憾的是，有些有的領導人，一方面大講特講以法治國（當然也很必要），而另一方面卻又背離以法治國，甚至踐踏以法治國。本來，趙紫陽是黨和國家的重要領導人，當過政府總理，當過黨的總書記，然而，未經任何法律手續和司法審判，憑幾個老人的幾句話，就把趙紫陽長期軟禁起來，剝奪他的人身自由長達十五年之久。這難道不是對以法治國的嘲諷和諷刺？不是對以法

111



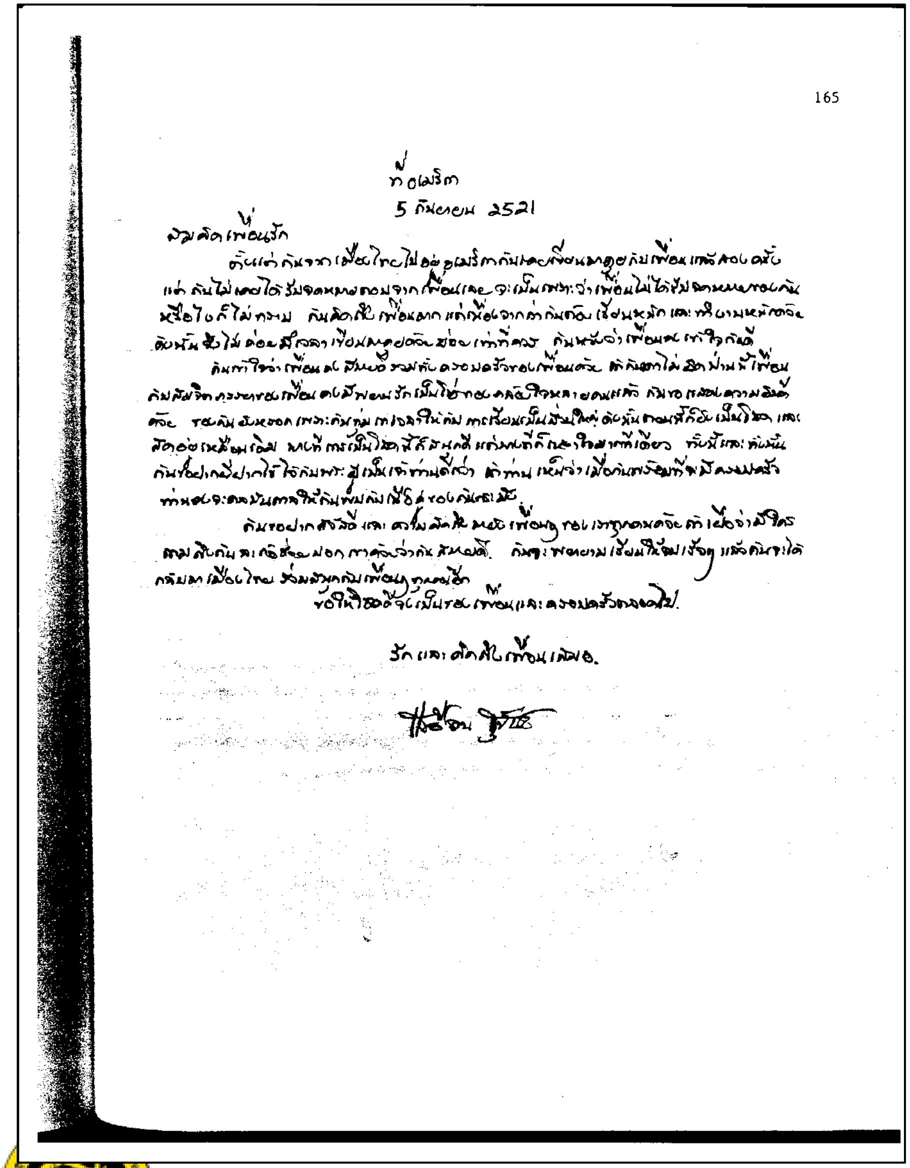
① आर्थिक विरोधाभास — व्यष्टिगत अर्थशास्त्र की समष्टिगत
के रूप इसके आर्थिक विरोधाभास का ज्ञान करने में
सहायक होता है

② विभागीय स्तर की समस्या — व्यष्टिगत अर्थशास्त्र विभागीय
स्तर की समस्या के समाधान में सहायक होता है तथा इसका ज्ञान
उद्योग होता है जो स्तर [illegible] होते हैं या इस
से दूर किये जाते हैं

③ मॉडल निर्माण — व्यष्टिगत अर्थशास्त्र इकाई मॉडल के
निर्माण को प्रयोग में सहायक होता है तथा इसका
रूप से स्पष्ट करता है

④ साधनों का आवंटन बटवारा — किसी स्तर पर साधन
के लिए है इसका साथ आवश्यक साधन
[illegible] क्योंकि उत्पादन स्तर हैं आवंटन बटवारा
अलग-2 होते हैं तथा व्यष्टिगत अर्थशास्त्र हमें
इसका ज्ञान प्राप्त करता है

⑤ राष्ट्रीय आय की गणना — किसी भी देश का
आर्थिक विकास वहाँ के राष्ट्रीय आय पर निर्भर
करता है और यह उत्पादों द्वारा इस प्रकार
आय का विश्लेषण करके कुल आय से
है राष्ट्रीय आय निकाल सकते हैं



165

ที่อเมริกา
5 กันยายน 2521

สมศักดิ์เพื่อนรัก

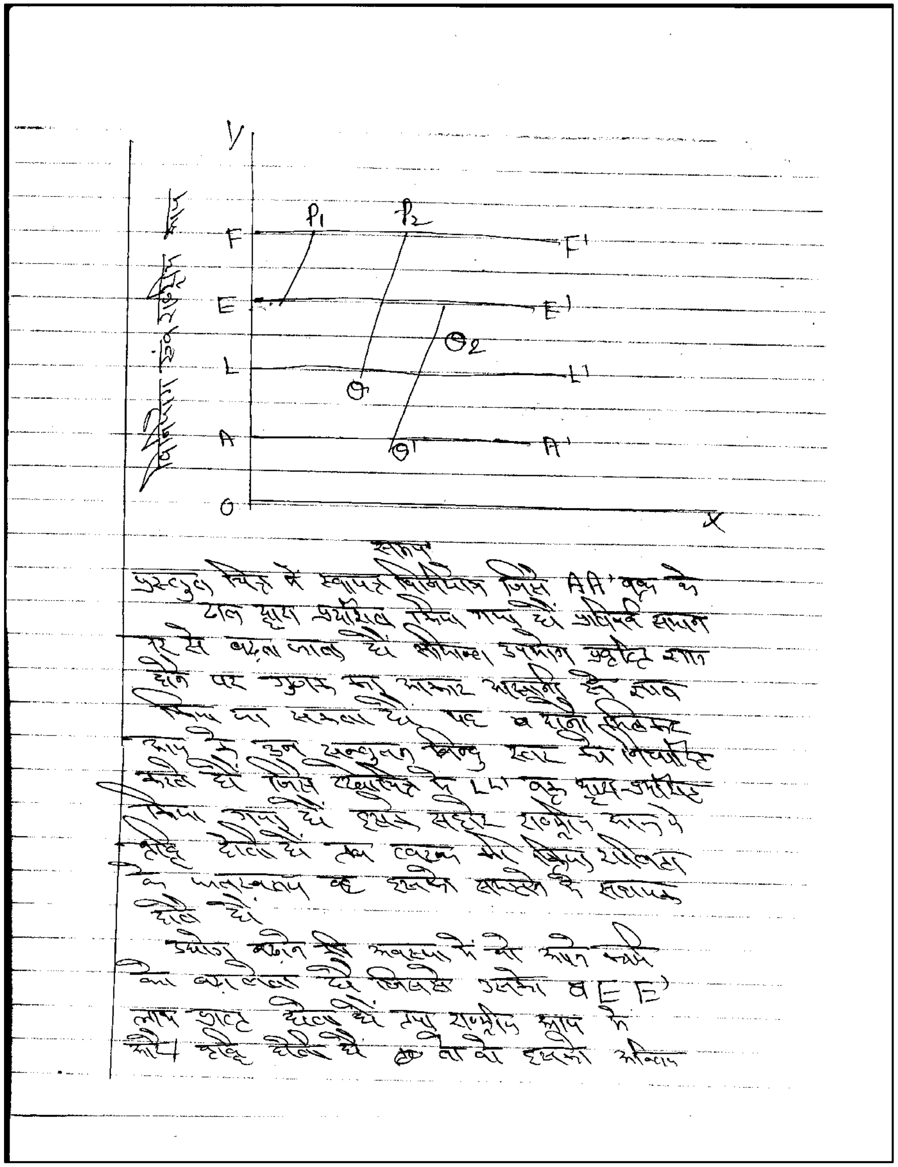
ตั้งแต่ฉันจากเมืองไทยไปอยู่อเมริกาฉันเคยเขียนมาคุยกับเพื่อนหลายครั้ง แต่ฉันไม่เคยได้รับจดหมายตอบจากเพื่อนเลย จะเป็นเพราะว่าเพื่อนไม่ได้รับจดหมายของฉันหรือไงก็ไม่ทราบ ฉันคิดถึงเพื่อนมาก แต่เนื่องจากว่าฉันต้องเรียนหนัก และทำงานหาเงินด้วย ดังนั้นจึงไม่ค่อยมีเวลาเขียนจดหมายถึงบ่อยเท่าที่ควร ฉันหวังว่าเพื่อนคงเข้าใจฉันดี

ฉันทราบว่าเพื่อนคงสบายดี รวมทั้งครอบครัวของเพื่อนด้วย ถ้าฉันจำไม่ผิดป่านนี้เพื่อนคงสมจิตความหวังของเพื่อน คงมีหลานรักเป็นโขยงคงหลายคนแล้ว ฉันขอแสดงความยินดีด้วย ของฉันยังหรอก เพราะฉันมุ่งทางด้านการศึกษาเป็นส่วนใหญ่ ดังนั้นตอนนี้ก็ยังเป็นโสดและสดอยู่เหมือนเดิม แท้ที่จริงการเป็นโสดมีก็ส่วนดี แต่บางทีก็เหงาใจมากทีเดียว ทุกวันนี้และฉันเห็น ฉันชอบเมืองไทยไม่ได้ไปไหน สู่เมืองท่านดีกว่า ถ้าท่านเห็นว่าเมื่อฉันพร้อมที่จะมีครอบครัว ท่านคงจะคอยมันคงให้ฉันพบกับผู้ที่ดีของฉันเองนะ

ฉันขอฝากความคิดถึงและความรักมาสู่เพื่อนๆ ของเราทุกคนด้วย ถ้าเมื่อว่ามีใครถามถึงฉัน ก็ฝากบอกด้วยว่าฉันสบายดี ฉันจะพยายามเขียนมาให้บ่อยๆ แล้วฉันจะได้กลับมาเมืองไทย ร่วมสนุกกับเพื่อนๆ อีก

ขอให้โชคดีจงเป็นของเพื่อนและครอบครัวตลอดไป.

รักและคิดถึงเพื่อนเสมอ.


y
F
E
L
A
O
x
P_1
P_2
F'
E'
L'
A'
θ_2
θ
θ'

व्याख्या

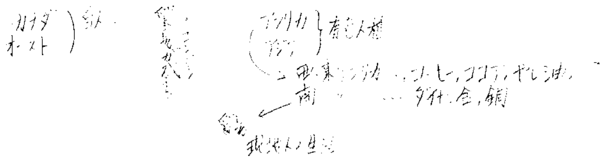
प्रस्तुत चित्र में स्थापित विभिन्न बिंदु AA' वक्र के टल द्वारा प्रदर्शित किया गया है। प्रतिदर्श समान रूप से बदल जाता है। अर्थात् प्रत्येक स्तर होने पर उत्पाद का आकार अपरिवर्तित है। स्थिर किया जा सकता है पर a घटना लेकर आय में इन उत्पादन बिन्दु स्तर से विचलित करते हैं जिससे उत्पादित के LL' वक्र पूर्व-निर्धारित किया गया है। इसमें सर्वोच्च उत्पादन स्तर से होते होते हैं तब व्यक्ति को निश्चित का व्यवहार के इसको सम्पूर्ण के सहायक होते हैं।

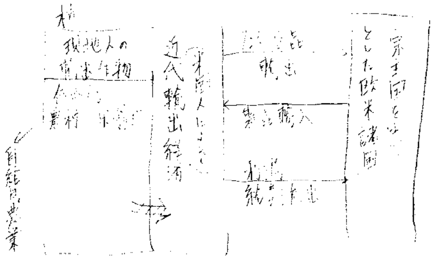
प्रत्येक बिंदु की अवस्था में जो अपना स्थान का आकलन है जिसके उत्पाद a E E' लाभ उठाते होते हैं तथा सम्पूर्ण आय से कार्य करते होते हैं वा वो इसको अधिकांश

Ⅱ 途上国の社会経済的特質

1.基本的構造 ～ 従属型二重経済

今までは歪んだ経済構造を作りだした。これは植民地型二重経済と呼ばれるもので、他の白人居住植民地(オーストラリア、NZ、カナダ、US)が貿易の発展により急速に伸びたのにこれら植民地は停滞を余儀なくされた。





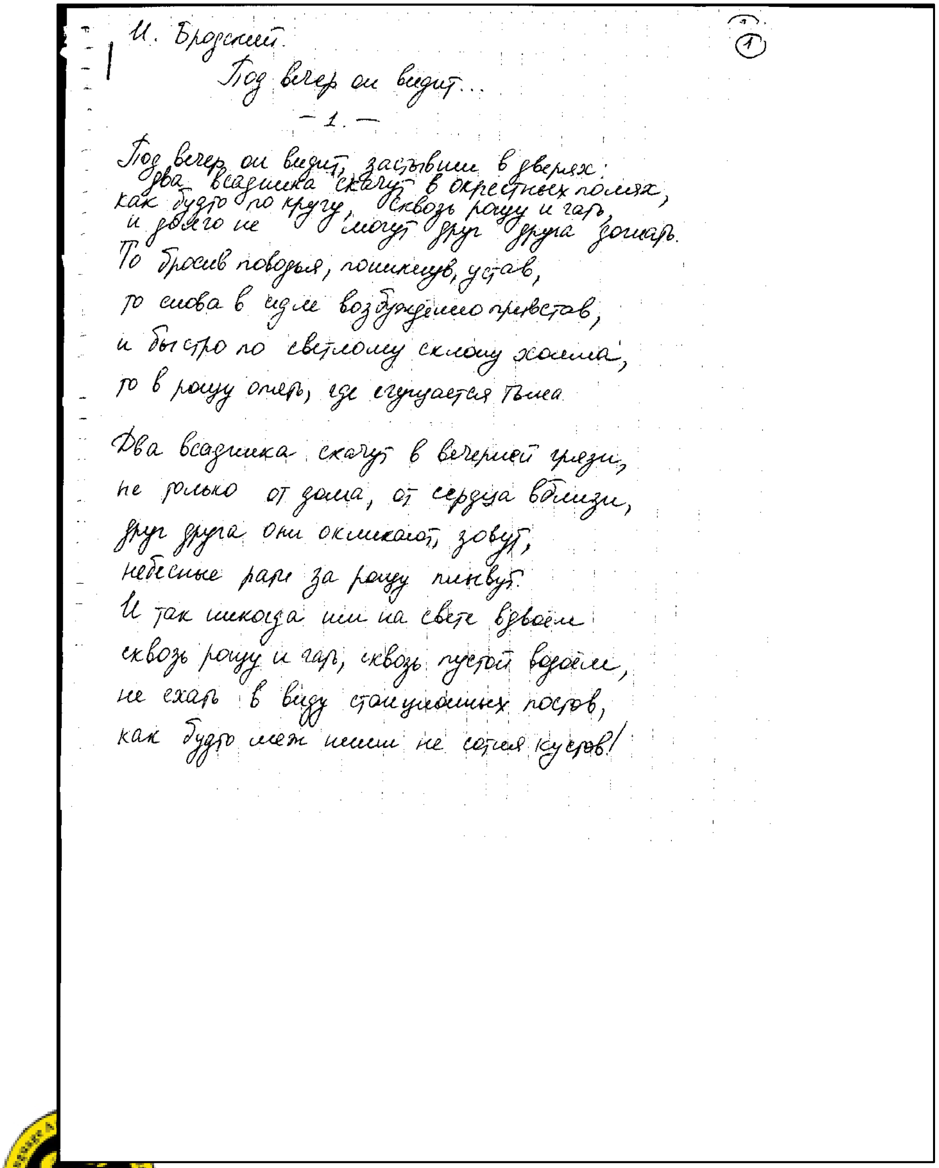
植民地二重経済は輸出向けの一次産品生産に特化し、その生産からは特定の外国人と[illegible]現地人の多くは自給農業に依存し生活水準は低い。
現地で[illegible]利潤の多くは[illegible]海外へ流出

6. The Three Kingdoms and the Six Dynasty

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

〈 Taoism 〉 도교

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



И. Бродский

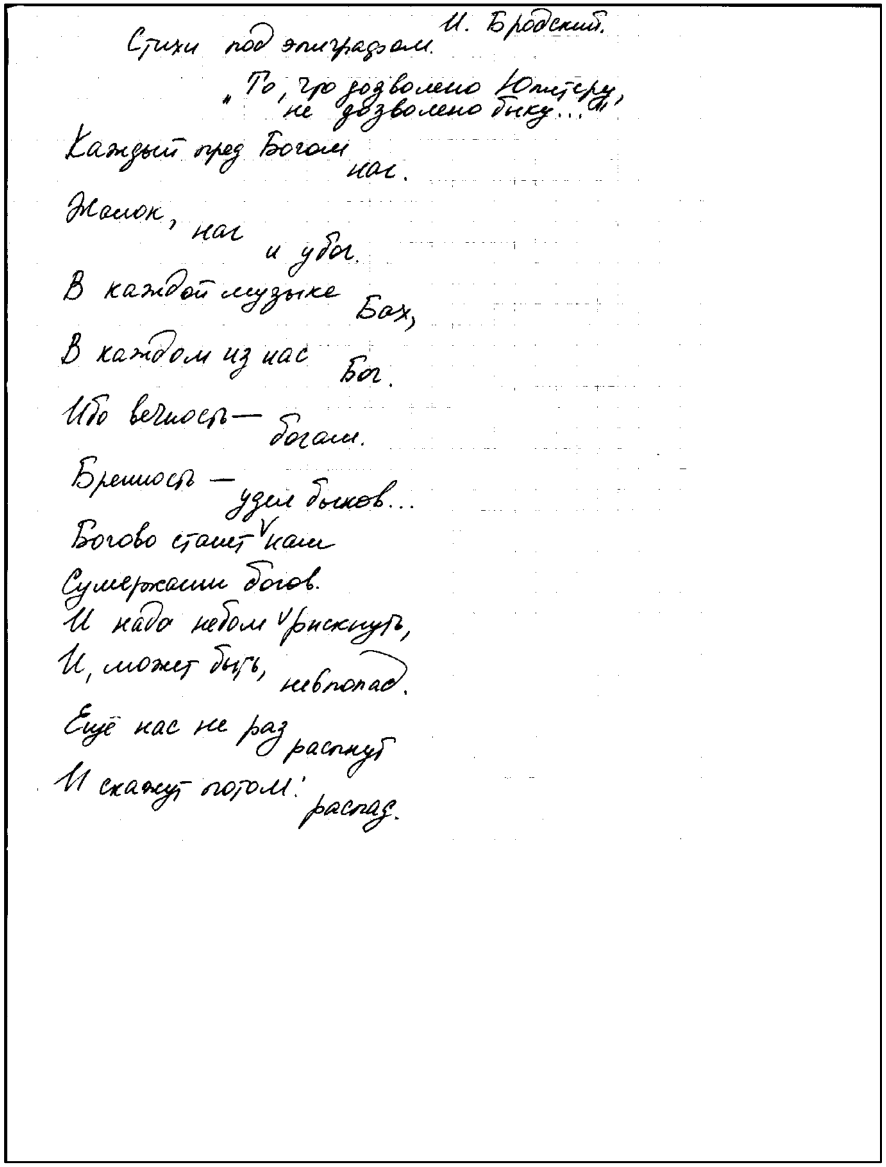
Под вечер он видит...

1

— 1. —

Под вечер он видит, застывши в дверях:
два всадника скачут в окрестных полях,
как будто по кругу, сквозь рощу и гарь,
и долго не могут друг друга догнать.
То бросив поводья, поникнув, устав,
то снова в седле возбужденно привстав,
и быстро по светлому склону холма,
то в рощу опять, где сгущается тьма.

Два всадника скачут в вечерней грязи,
не только от дома, от сердца вблизи,
друг друга они окликают, зовут,
небесные рати за рощу плывут.
И так никогда им на свете вдвоем
сквозь рощу и гарь, сквозь пустой водоем,
не ехать в виду станционных постов,
как будто меж ними не сотня кустов!



Стихи под эпиграфом. И. Бродский.

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

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

Other “Documents”

Showroom for:
Norsk Wood Works

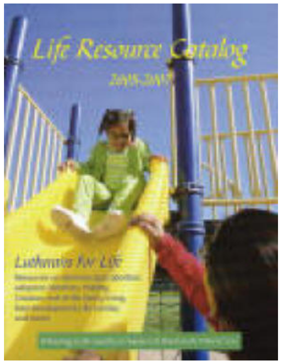
For Wood Carving &Furniture

Please call us directly at:

Phone: 715 468 2780

Web: www.norskwoodworks.com

Norwegian Fjord Horses!



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

Poker Tables from Poker-Wear

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

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

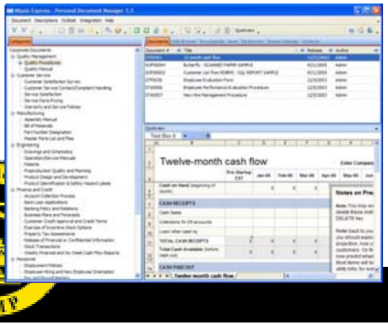
Poker Table

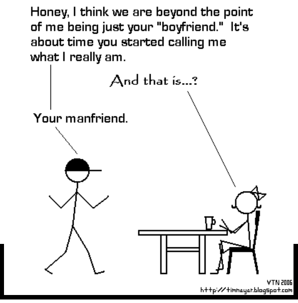
Poker Table

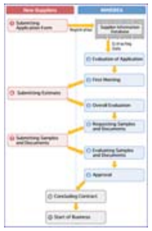
Poker Table

Poker Table

Poker-Wear.com 3121 Potts Crossing Smyrna, TN 37167 USA	Phone 615.566.9112	Hours Monday-Friday 9 AM to 5 PM Central

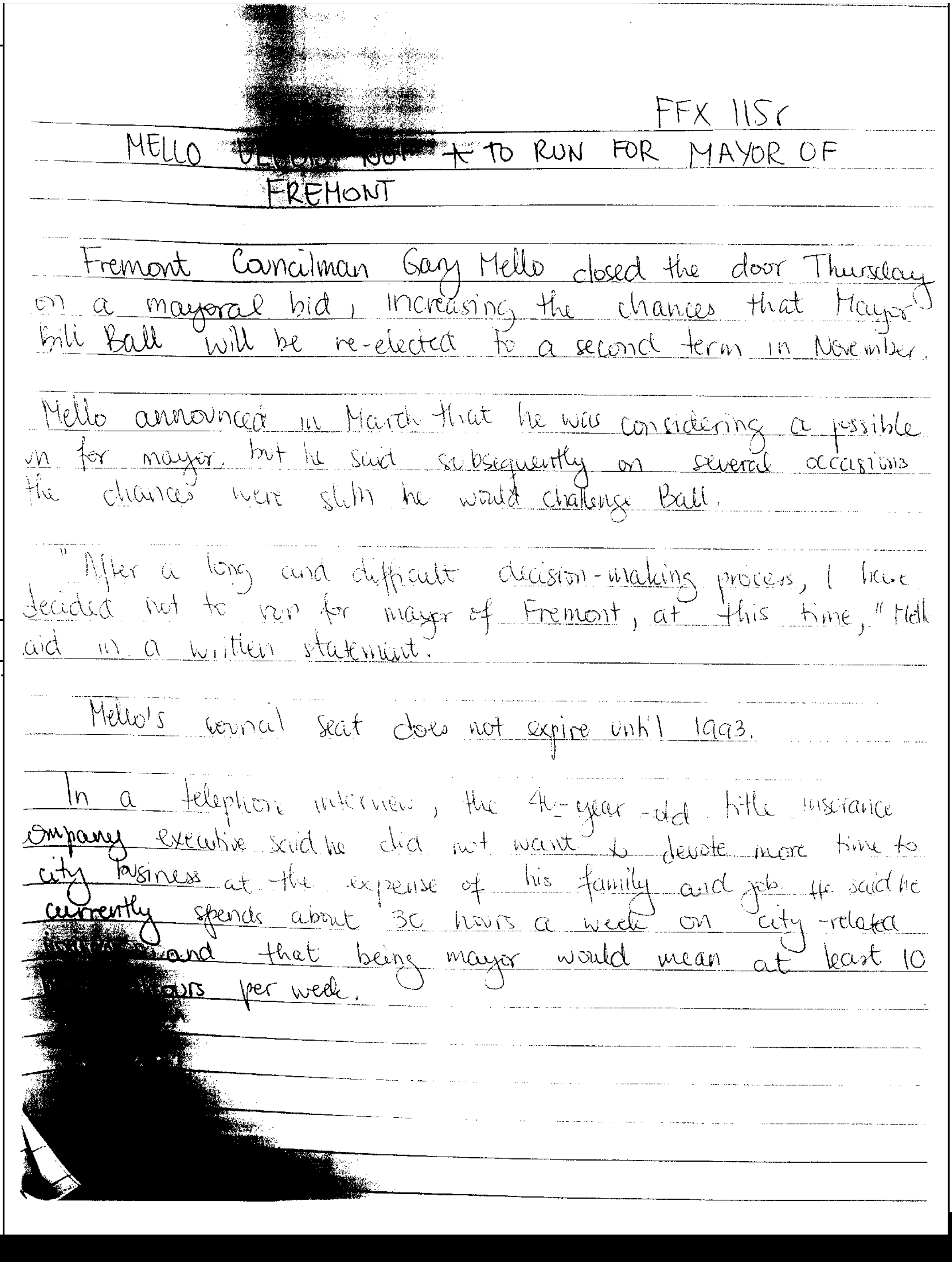



Honey, I think we are beyond the point of me being just your "boyfriend." It's about time you started calling me what I really am.
And that is...?
Your manfriend.
VTN 2006
http://timmeyer.blogspot.com



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 [illegible] + TO RUN FOR MAYOR OF FREMONT

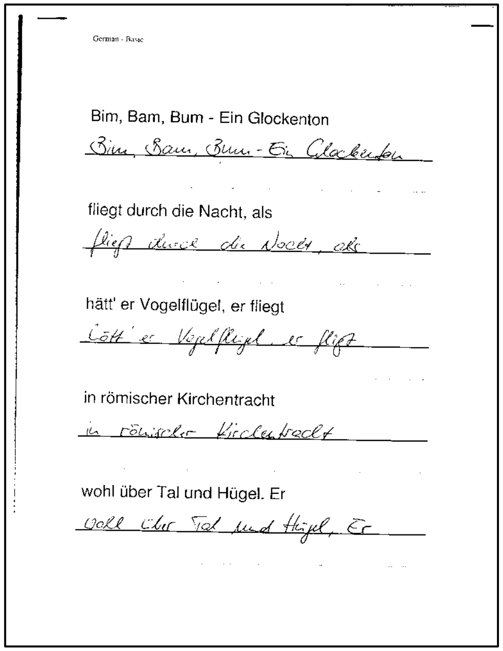
Fremont Councilman Gary Mello closed the door Thursday on a mayoral bid, increasing the chances that Mayor Bill Ball will be re-elected to a second term in November.

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," Mell aid in a written statement.

Mello's council seat does not expire until 1993.

In a telephone interview, the 4[illegible]-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 [illegible] and that being mayor would mean at least 10 [illegible] 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

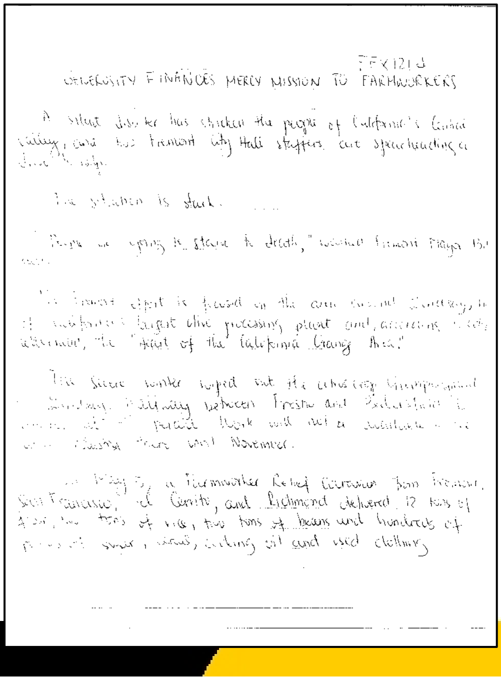
hätt' er Vogelflügel, er fliegt

in römischer Kirchentracht

in römischer Kirchentracht

wohl über Tal und Hügel. Er

wohl über Tal und Hügel, Er



FFX121d

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 Lindsay, [illegible] of California's largest olive processing plant and, according to city [illegible], the "heart of the California Orange Area."

The severe winter wiped out the citrus crop [illegible] Sunday. Halfway between Fresno and Bakersfield [illegible] harvest work will not be available [illegible] picking season 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, [illegible], cooking oil and used clothing

New Data

- 25,000 pages ground truthed to the zone level
- Sampled from the Tobacco Litigation Corpus of 49 Million pages

Distribution (docs, pages)

dt_calendar	44	90
dt_photograph	227	461
dt_questionnaire	188	461
dt_bibliography	175	530
dt_periodical	479	693
dt_list	405	710
dt_advertisement	519	894
dt_newspaper	688	921
co_fax	830	1150
co_drawings	638	1150
dt_email	973	1151
co_tables	1049	1980
dt_form	1582	2265
co_foreign	1669	2300
dt_notes	2288	2925
co_illegible	2598	3983
dt_graphic	2061	4307
dt_letter	3145	4601
dt_report	2213	4604
dt_memo	2762	4611
co_handwritten	4894	6903
co_marginalia	10665	17251

Agenda
Tuesday AM

- Project Overview
  - Introduction
  - Goals and Objectives
- Tools
  - GEDI Display Environment
  - Datasets
- DocLib and Algorithms
  - Technical Presentations

DocLib Architecture

- **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

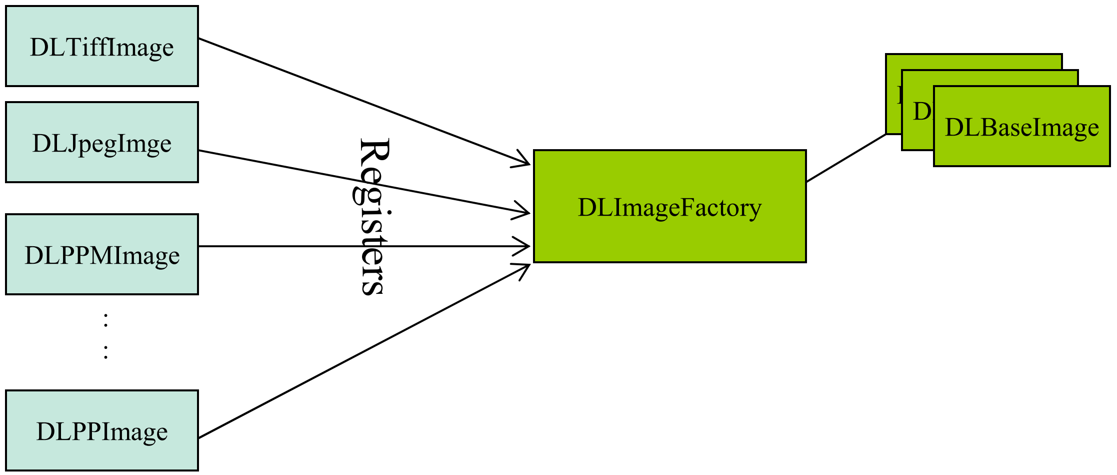
- Core DocLib components matured and stable (in use by a variety of government installations)\
- Addons being integrated/implemented, primarily by developers
- Freely available to government researchers
- Core supported on Solaris, Linux and Windows

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



<u>Design Factors:</u>

- **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

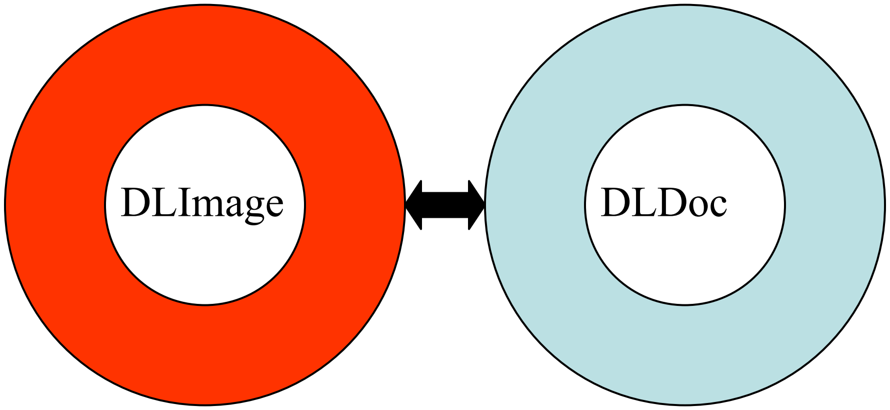
DocLib´s architecture rests on two pillars:

DLImage:

- Image Processing

DLDocument:

- Document Processing



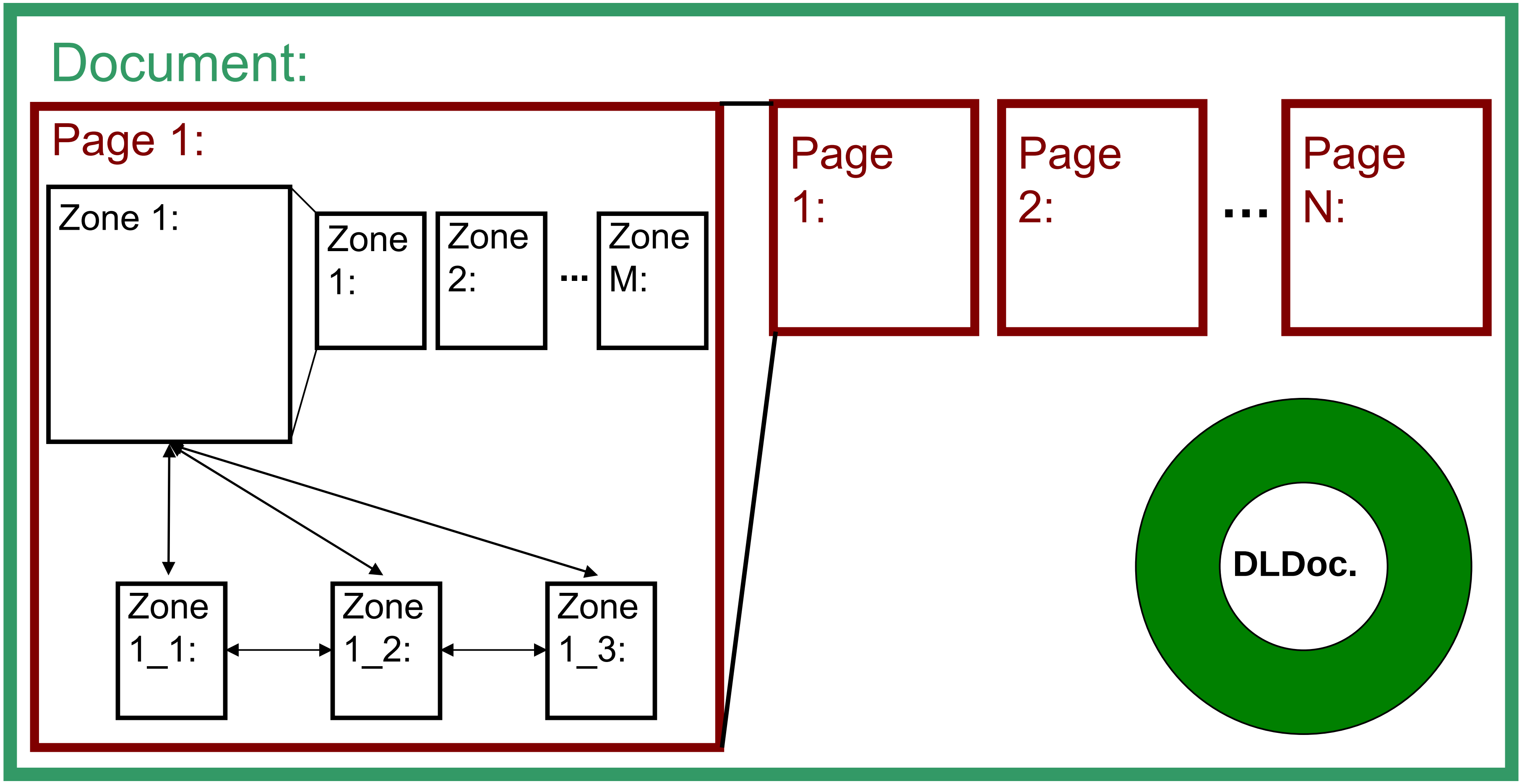
e.g.

- 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

XML Output Extension

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- GEDI is developed at Language and Media Processing Laboratory,
     University of Maryland.   -->
<GEDI xmlns="http://lamp.cfar.umd.edu/GEDI" version="1.0">
  <USER name="Elena" date="Sun, 14 Oct 2007 8:28 PM" />
  <DL_DOCUMENT src="aaa27e00.tif" docTag="xml" NrOfPages="2">
    <DL_PAGE gedi_type="DL_PAGE" src="aaa27e00.tif" pageID="1«
            width="2560" height="3296">
            <DL_ZONE gedi_type="STAMP" id="None" col="1174" row="495“
                    width="447" height="132" />
            <DL_ZONE gedi_type="LOGO" id="None" col="274" row="569"
                     width="346" height="159" contents="" />
            <DL_ZONE gedi_type="MACHINEPRINT" id="None" col="647"
                     row="626" width="1372" height="105" contents="" />
             <DL_ZONE gedi_type="MACHINEPRINT" id="None" col="2410"
                     row="2479" width="511" height="110" orientation="-
                     1.6295521495106193" contents="" />
     </DL_PAGE>
   </DL_DOCUMENT>
```

Agenda
Tuesday AM

- Project Overview
  - Introduction
  - Goals and Objectives
- Tools
  - GEDI Display Environment
  - Datasets
- DocLib and Algorithms
  - Technical Presentations

Technical Presentations

- Page Segmentation (and rule line separation)
- Page Layout Similarity
- Document ID/Script ID

This afternoon

- Logo Detection and Recognition
- Signature Detection
- Font OCR

Technical Presentation Outline

- Overview of Problem
- Technical Approach
- Datasets
- Results
- Implementation and Software

Examples of Drivers

- ScriptID [-x] [-l] filename

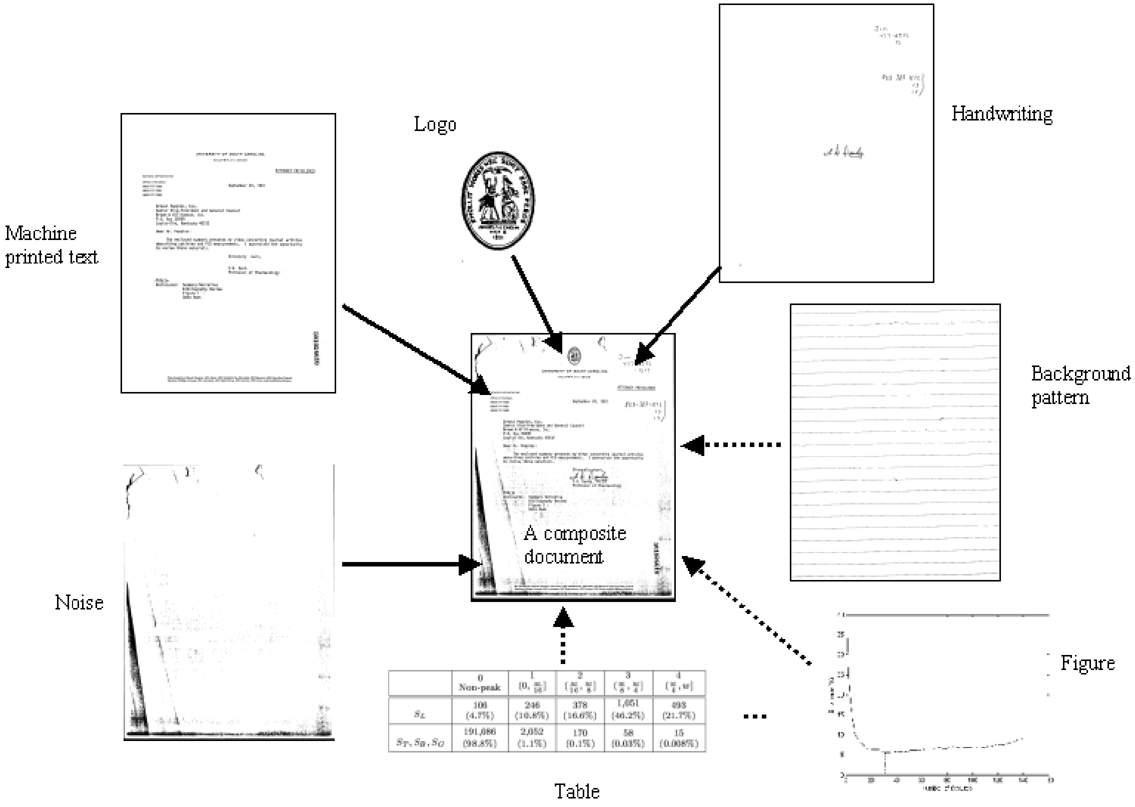
```
-x  ---  Write classification results into an xml file for each input image.
         It creates a new xml file if no associated xml file exists.
-l  ---  File containing the list of input images to execute
-h  ---  Show help at command line
```

- DocID [-x] [-l] filename

```
-x  ---  Write classification results into an xml file for each input image.
         It creates a new xml file if no associated xml file exists.
-l  ---  File containing the list of input images to execute
-h  ---  Show help at command line
```

Page Layer Segmentation

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



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

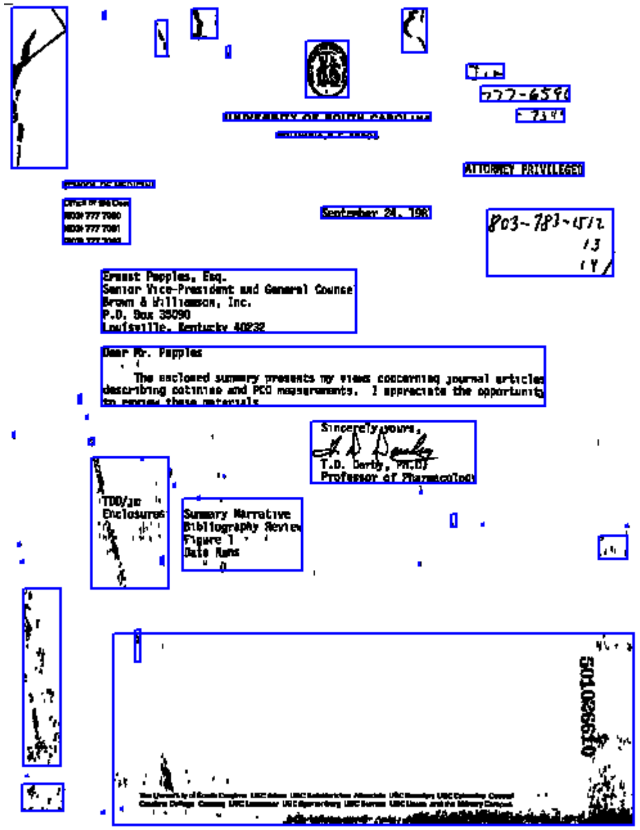
Motivation

- Document analysis has been viewed as a solved problem in clean, well-constrained documents.
- However, the performance degrades significantly when a small amount of noise is introduced.
- We further separate handwriting from machine printed text.

Motivation

- Layer analysis and separation for general, heterogeneous documents, is a very hard problem.
- Handwritten documents are very important
  - Handwriting was developed a long time ago as a means to expand human memory and to facilitate communication.
  - We are continuing to produce handwritten documents.

Page Segmentation for Noisy Documents



777-6590

7397

UNIVERSITY OF SOUTH CAROLINA

ATTORNEY PRIVILEGED

September 24, 198[illegible]

803-783-1512
/3
/4

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

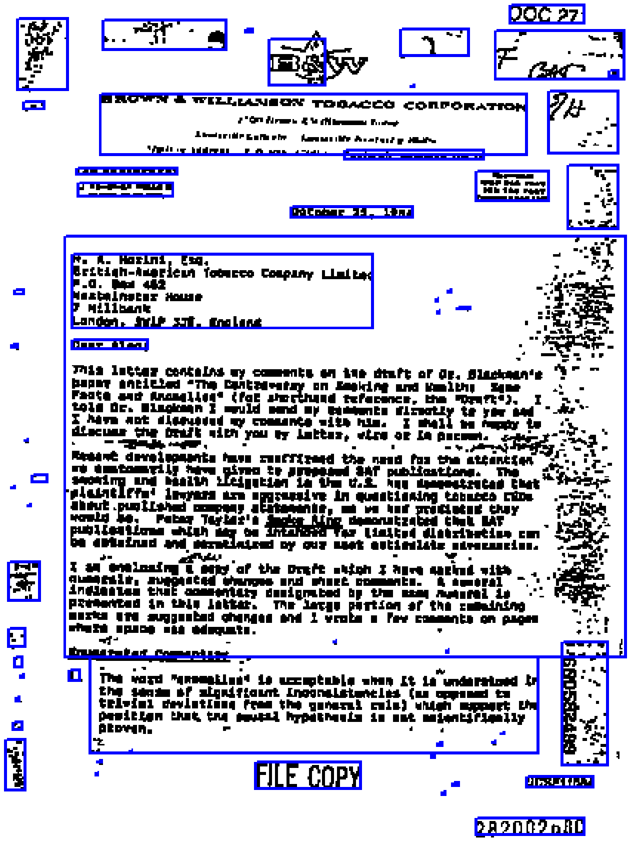
Dear Mr. Pepples

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

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

TDD/jw
Enclosures

Summary Narrative
Bibliography Review
Figure 1
Data Runs



DOC 27

BROWN & WILLIAMSON TOBACCO CORPORATION

October 2[illegible], 198[illegible]

R. A. Morini, Esq.
British-American Tobacco Company Limited
P.O. Box 482
Westminster House
7 Millbank
London, SW1P 3JE, England

Dear Alan:

This letter contains my comments on the draft of Dr. Blackman's paper entitled "The Controversy on Smoking and Health: Some Facts and Anomalies" (for shorthand reference, the "Draft"). I told Dr. Blackman I would send my comments directly to you and I have not discussed my comments with him. I shall be happy to discuss the Draft with you by letter, wire or in person.

Recent developments have reaffirmed the need for the attention we customarily have given to proposed BAT publications. The smoking and health litigation in the U.S. has demonstrated that plaintiffs' lawyers are aggressive in questioning tobacco CEOs about published company statements, as we had predicted they would be. Peter Taylor's Smoke Ring demonstrated that BAT publications which may be intended for limited distribution can be obtained and scrutinized by our most articulate adversaries.

I am enclosing a copy of the Draft which I have marked with numerals, suggested changes and short comments. A general indication that commentary designated by the same numeral is presented in this letter. The large portion of the remaining marks are suggested changes and I wrote a few comments on pages where space was adequate.

Enumerated Comments:

1. The word "anomalies" is acceptable when it is understood in the sense of significant inconsistencies (as opposed to trivial deviations from the general rule) which support the position that the causal hypothesis is not scientifically proven.

FILE COPY

* Docstrum page segmentation technique is used

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.

Feature Extraction and Selection

- We extracted 140 features and 31 of them are selected to train the

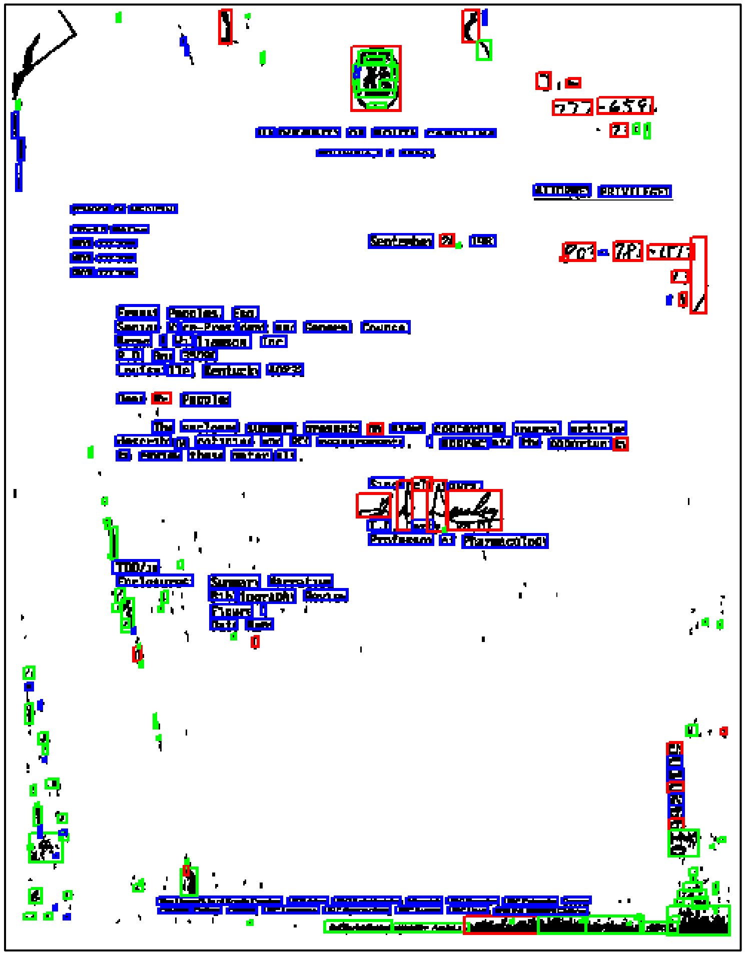
	Usage description	Dimensio	Selected
Structural	Region size, connected components	18	9
Gabor filter	Stroke orientation	16	4
Run-length histogram	Stroke length	20	5
Crossing counts histogram	Stroke complexity	10	6
Co-occurrence	Texture	16	2
2×2 gram	Texture	60	5
Total		140	31

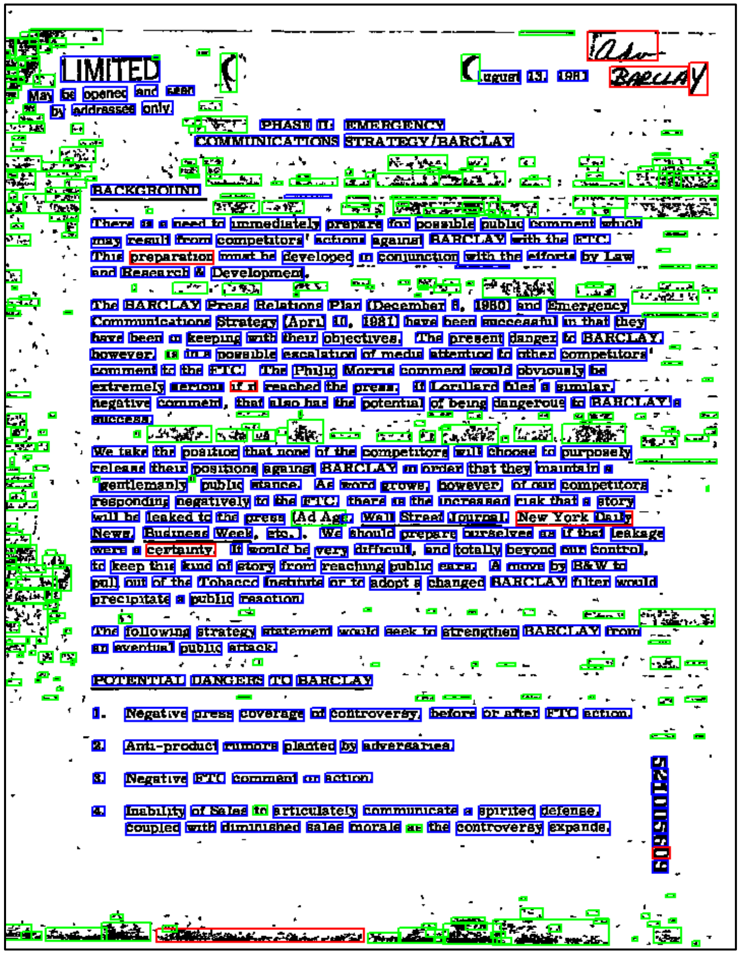
Classification Results with Fisher Classifier

Printed text

Handwriting

Noise





LIMITED — May be opened and seen by addressees only.

August 13, 1981

Barclay

PHASE II: EMERGENCY COMMUNICATIONS STRATEGY/BARCLAY

BACKGROUND

There is a need to immediately prepare for possible public comment which may result from competitors' actions against BARCLAY with the FTC. This preparation must be developed in conjunction with the efforts by Law and Research & Development.

The BARCLAY Press Relations Plan (December 8, 1980) and Emergency Communications Strategy (April 10, 1981) have been successful in that they have been in keeping with their objectives. The present danger to BARCLAY, however, is in a possible escalation of media attention to other competitors' comment to the FTC. The Philip Morris comment would obviously be extremely serious if it reached the press. If Lorillard files a similar, negative comment, that also has the potential of being dangerous to BARCLAY's success.

We take the position that none of the competitors will choose to purposely release their positions against BARCLAY in order that they maintain a "gentlemanly" public stance. As word grows, however, of our competitors responding negatively to the FTC, there is the increased risk that a story will be leaked to the press (Ad Age, Wall Street Journal, New York Daily News, Business Week, etc.). We should prepare ourselves as if that leakage were a certainty. It would be very difficult, and totally beyond our control, to keep this kind of story from reaching public ears. A move by B&W to pull out of the Tobacco Institute or to adopt a changed BARCLAY filter would precipitate a public reaction.

The following strategy statement would seek to strengthen BARCLAY from an eventual public attack.

POTENTIAL DANGERS TO BARCLAY

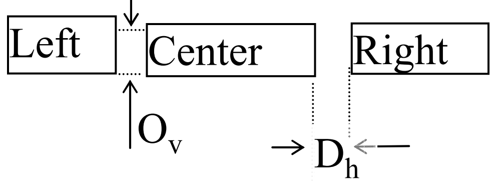
1. Negative press coverage of controversy, before or after FTC action.
2. Anti-product rumors planted by adversaries.
3. Negative FTC comment on action.
4. Inability of Sales to articulately communicate a spirited defense, coupled with diminished sales morale as the controversy expands.

Using Context

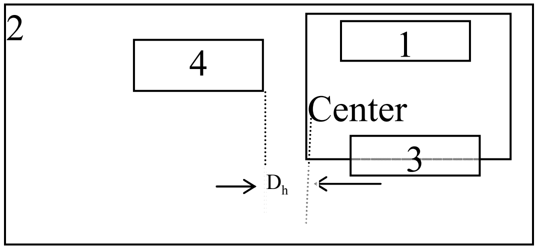
- The results are reasonable with a few mis-classification due to the overlapping of different classes in the feature space.
- Context can be used to refine classification results further
  - Words of printed text tend to lie on the same line.
  - Noise block are likely to overlap each other.
- This kind of local dependency among neighboring components can be described with the Markov Random Field (MRF).

Clique Definition

- Low level MRF is defined on regular lattice (pixel)
- Our high level MRF is defined on a graph.
  - After defining the connection between word blocks, a graph is generated.
  - Neighborhood of MRF is defined on the graph.
- Clique C_p for printed text



- Clique C_v for Noise



Clique Potential

- Clique potential:

$$V_p(c) = -\frac{P(x_l, x_c, x_r)}{\left(P(x_l)P(x_c)P(x_r)\right)^w} \qquad V_n(c) = -\frac{P(x_c, x_1, x_2, x_3, x_4)}{\left(P(x_c)P(x_1)P(x_2)P(x_3)P(x_4)\right)^w}$$

Probabilities are estimated from ground truth.

- Total energy of Gibbs distribution:

$$U(\underline{X} / \underline{Y}) = -w_s \sum_{s \in \Omega} P(x_s / y_s) + w_p \sum_{c \in C_p} V_p(c) + w_n \sum_{c \in C_n} V_n(c)$$

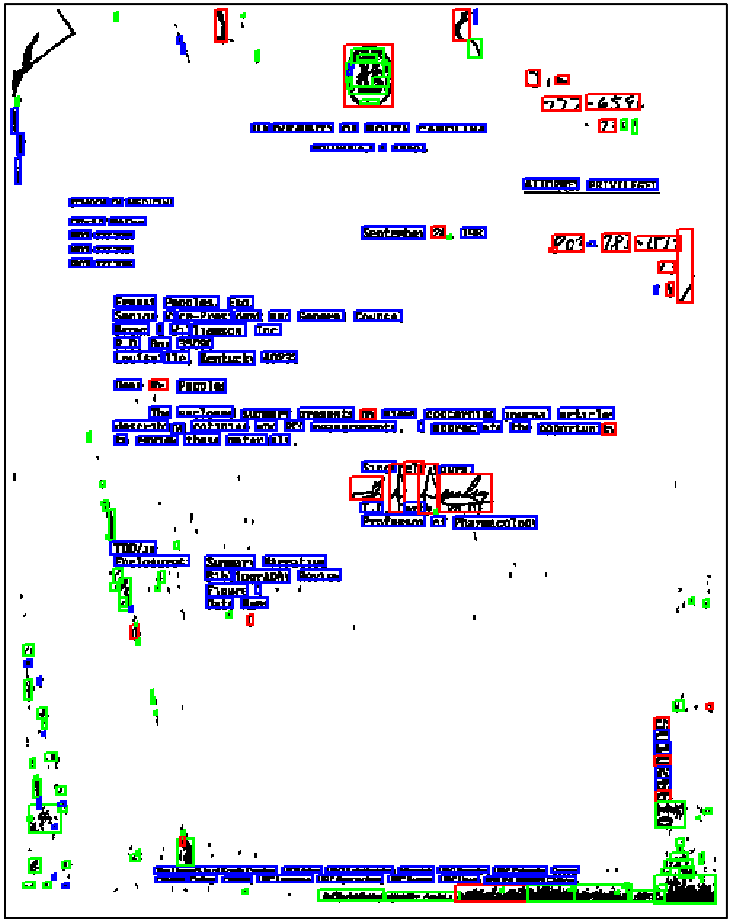
HCF (Highest Confidence First) method is used to minimize the energy function.

MRF Postprocessing Example

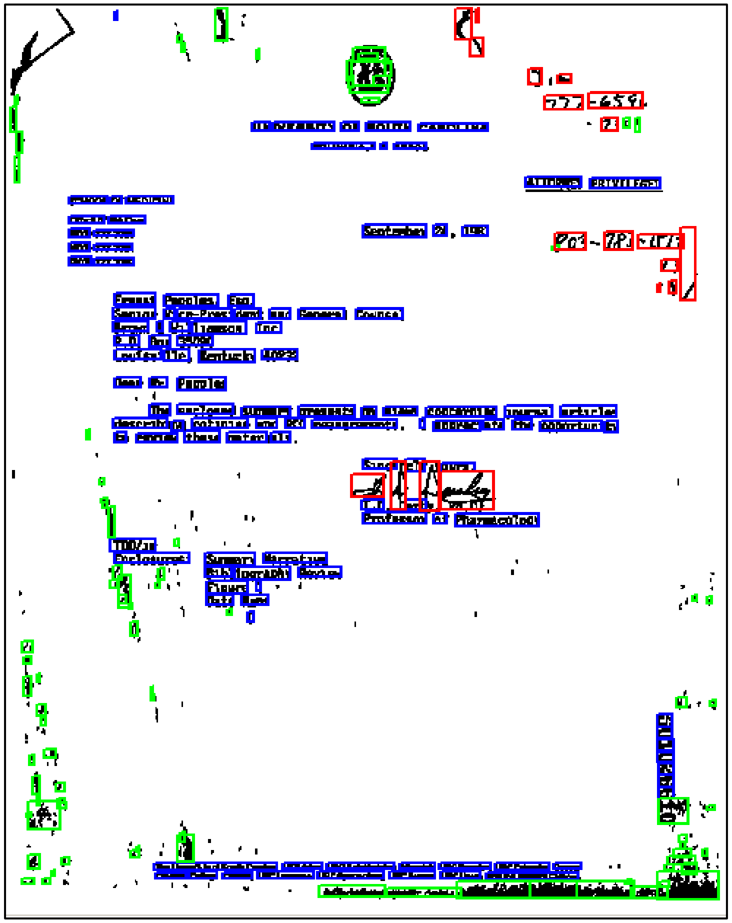
Printed text

Handwriting

Noise



Before MRF-based postprocessing



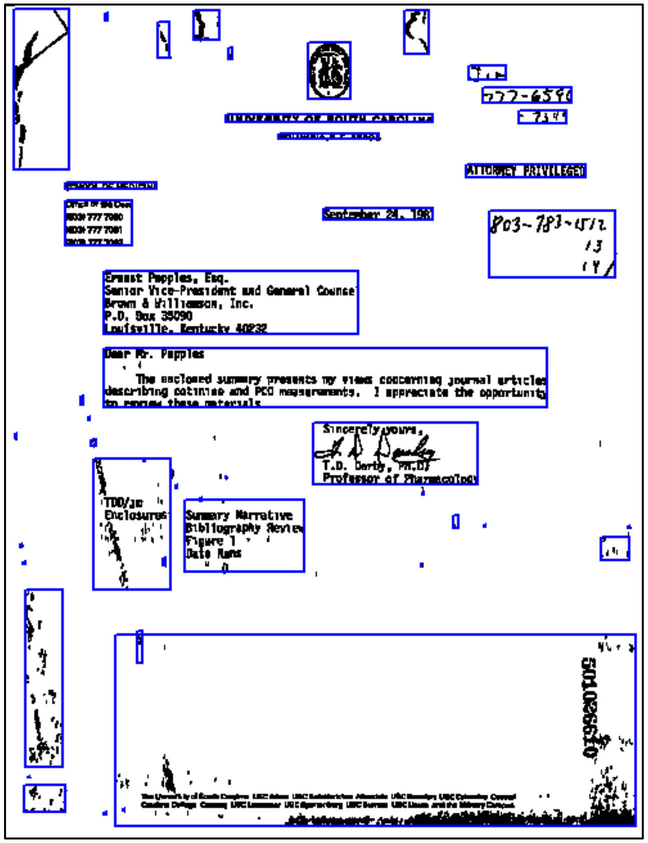
After MRF-based postprocessing

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



777-6596

7391

UNIVERSITY OF SOUTH CAROLINA

ATTORNEY PRIVILEGED

September 24, 198

803-783-1512

13

14

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

Dear Mr. Pepples

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

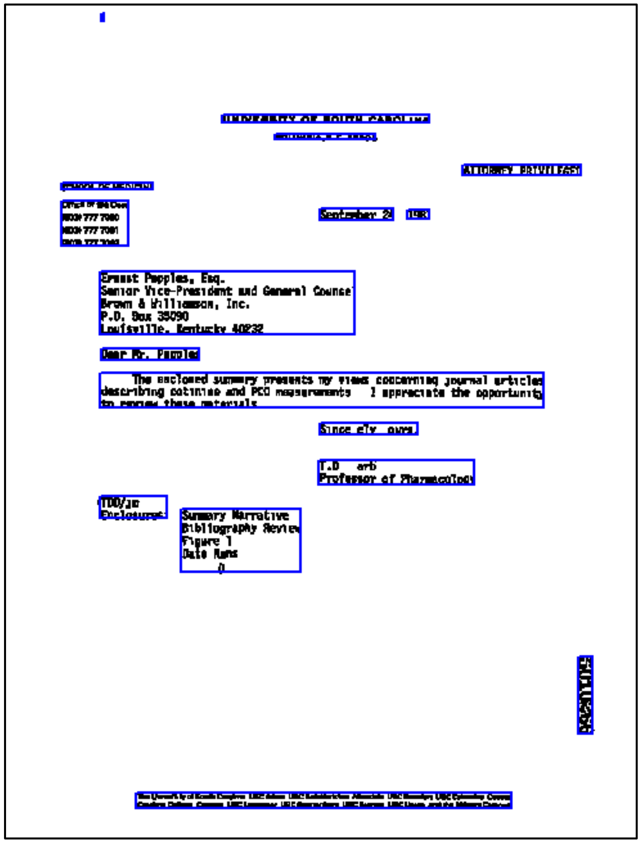
Sincerely yours,

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

TDD/jr
Enclosures:

Summary Narrative
Bibliography Review
Figure 1
Date Runs

Before enhancement



UNIVERSITY OF SOUTH CAROLINA

ATTORNEY PRIVILEGED

September 2 198

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

Dear Mr. Pepples

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

Since ely ours

T.D arb
Professor of Pharmacology

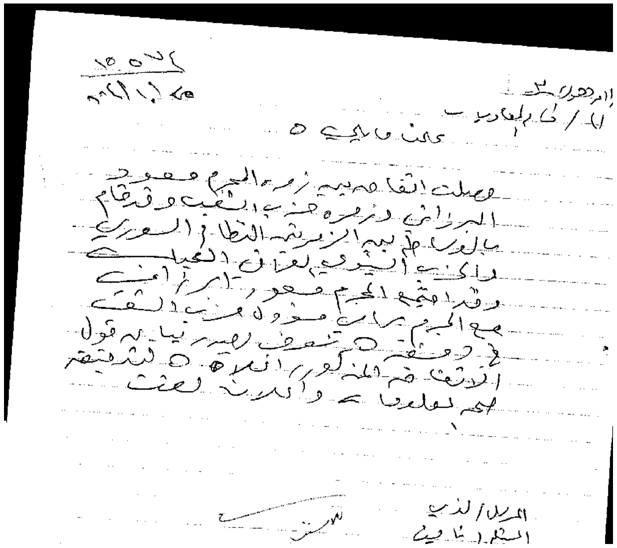
TDD/jr
Enclosures:

Summary Narrative
Bibliography Review
Figure 1
Date Runs

After enhancement

Background Pattern (Rule Line) Separation

- Many handwritten documents are produced on rule lined paper
- These lines should be detected and removed before we feed the text to an Optical Character Recognition (OCR) engine.



٢٧٥٠١٥

١٩٩١/٤/٢٥

ج [illegible]

م / [illegible]

[illegible]

حصلت اتفاقية بين زمرة المجرم مسعود
البرزاني وزمرة حزب البعث وقد قام
[illegible] النظام السوري
[illegible] العراق [illegible]
وقد اجتمع المجرم مسعود [illegible]
مع المجرم [illegible] مسؤول [illegible]
في [illegible] قول
[illegible]
[illegible] للعلم [illegible]

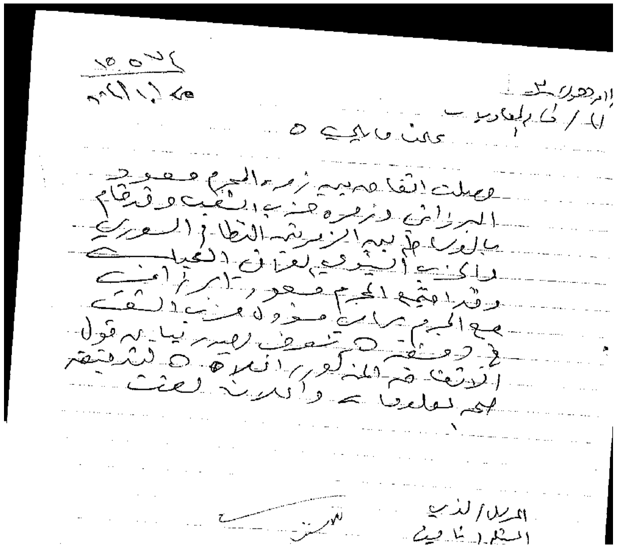
[illegible]

An Arabic handwritten document on rule lined paper

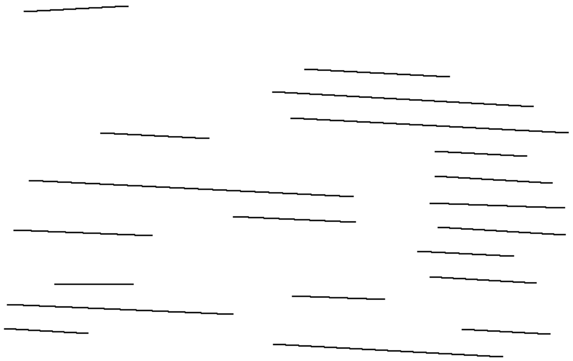
Challenges

- Previous work
  - Hough transform
  - Projection based (Strip Projection [Chen98] and Skew Projection [Liu95])
  - Vectorization based (BAG [Jain96], SPV [Dori99], and DSCC [Zheng01])
- Challenges
  - The documents may be degraded with severely broken lines.
  - High accuracy and low false alarm rate are demanded.

Challenges



Original document

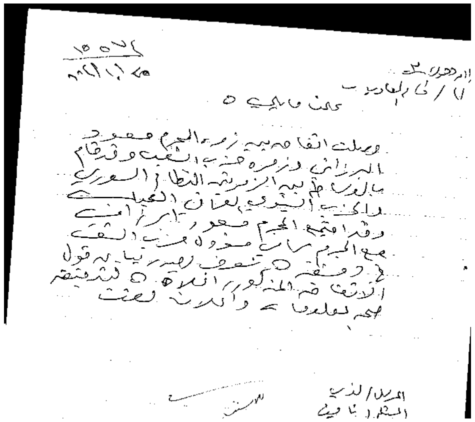


Line detection results using DSCC method

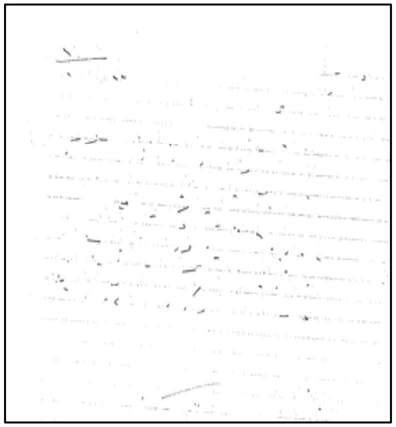
- We propose a model-based method
  - Model the horizontal projection profile with an HMM model.
  - Under the model, lines are detected simultaneously.

Preprocessing

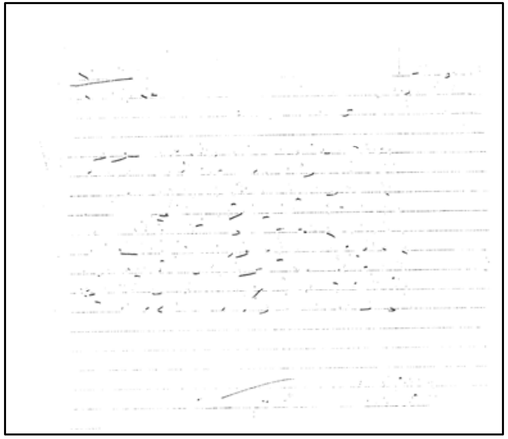
- Text filtering
- Skew estimation and correction



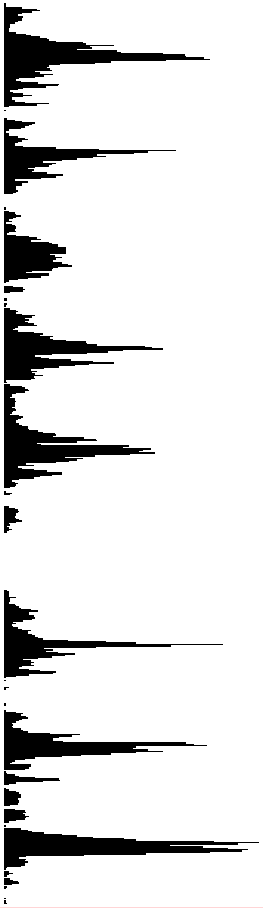
Original image



Text filtering



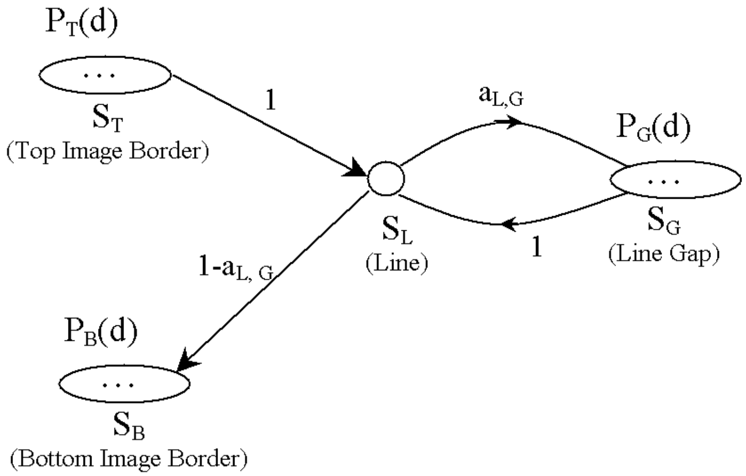
Skew correction



Horizontal projection profile

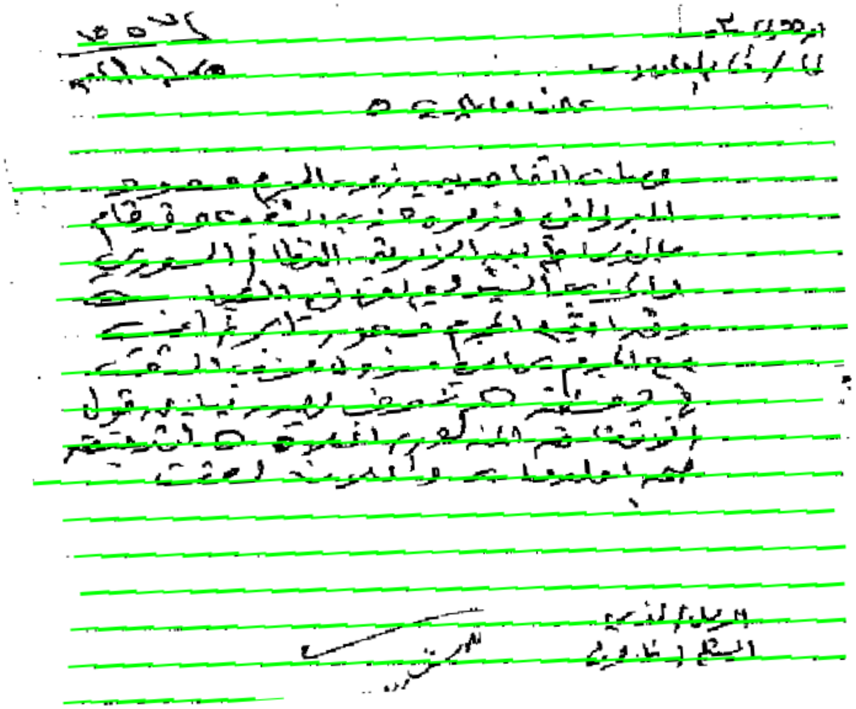
HMM Model for Parallel Lines

- Model the projection profile with an HMM model
  - The vertical position of lines $\{Y_i\}$ form a Markov Chain
  - We can not observe $\{Y_i\}$ directly, but projection profile
  - The gaps between neighboring lines are consistent on the same page

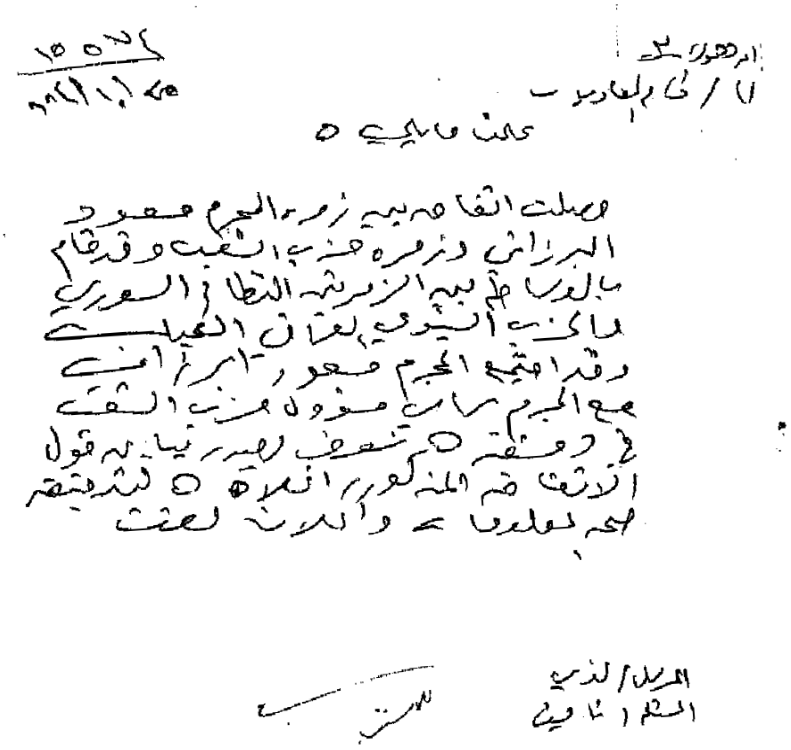


- Parameters of the model are estimated from ground truth.
- Viterbi algorithm is used to decode the model.

Rule Line Detection Example



Model-based line detection result

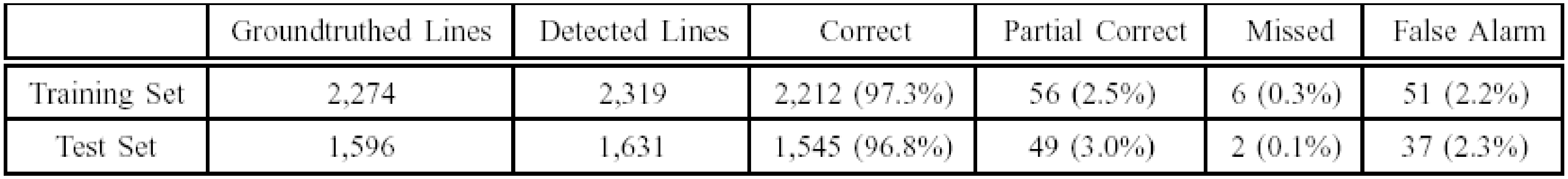


After rule line removal

Evaluation

- Database
  - 168 Arabic documents with a total of 3,870 groundtruthed lines.
  - 100 images for the training of the HMM model, 68 images for the testing.
- Quantitative evaluation (evaluation metrics are discussed in the paper in detail).

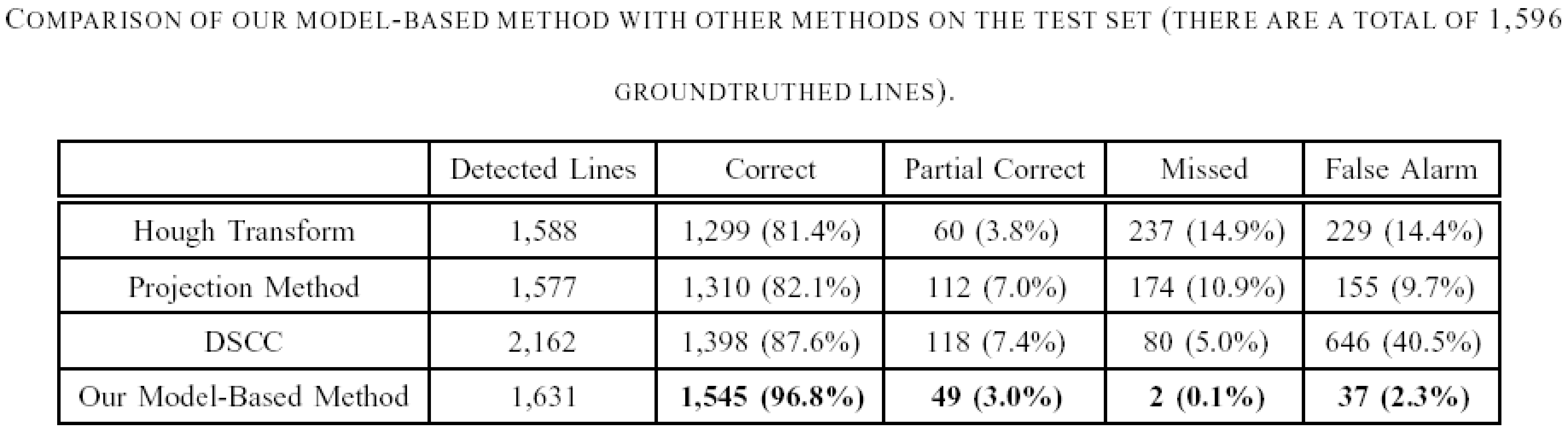
QUANTITATIVE EVALUATION OF THE RULE LINE DETECTION RESULT.



	Groundtruthed Lines	Detected Lines	Correct	Partial Correct	Missed	False Alarm
Training Set	2,274	2,319	2,212 (97.3%)	56 (2.5%)	6 (0.3%)	51 (2.2%)
Test Set	1,596	1,631	1,545 (96.8%)	49 (3.0%)	2 (0.1%)	37 (2.3%)

Comparison with Other Methods

- Hough transform
- DSCC
- Projection based methods



COMPARISON OF OUR MODEL-BASED METHOD WITH OTHER METHODS ON THE TEST SET (THERE ARE A TOTAL OF 1,596 GROUNDTRUTHED LINES).

	Detected Lines	Correct	Partial Correct	Missed	False Alarm
Hough Transform	1,588	1,299 (81.4%)	60 (3.8%)	237 (14.9%)	229 (14.4%)
Projection Method	1,577	1,310 (82.1%)	112 (7.0%)	174 (10.9%)	155 (9.7%)
DSCC	2,162	1,398 (87.6%)	118 (7.4%)	80 (5.0%)	646 (40.5%)
Our Model-Based Method	1,631	**1,545 (96.8%)**	**49 (3.0%)**	**2 (0.1%)**	**37 (2.3%)**

Software

- Implemented as a set of Libraries
- Trainable with new data

Technical Presentations

- Page Segmentation (and rule line separation)
- Page Layout Similarity
- Document ID/Script ID

This afternoon

- Logo Detection and Recognition
- Signature Detection
- Font OCR

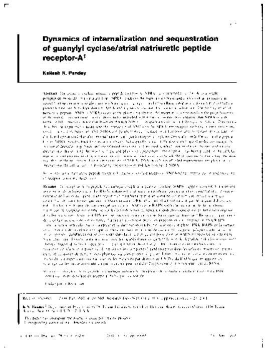
Multi-Class Classification using Document Layout

- Motivation
- Document Representation
- Random Chopping
- Feature Selection
- Score Function
- Experiments
- Summary and Future work

Motivation

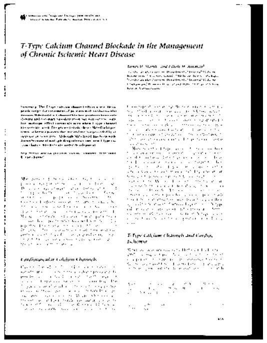
- *In a large collection of documents (forms, academic papers, handwritten letters, checks, receipts, etc.), most times people need to handle only those with some specific layout.*
- ***Drawback** of our previous system for document ranking based on layout : training is restarted from beginning each time a new layout comes*
- **Reason**: *we do not give an explicit definition of layout, the system learns no concept of layout, but image content.*
- ***Proposal***: Let the system itself figure out important dissimilarities for layout classification.

Layout Examples

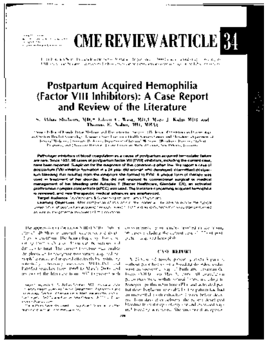

Dynamics of internalization and sequestration of guanylyl cyclase/atrial natriuretic peptide receptor-A

Kailash N. Pandey

1C

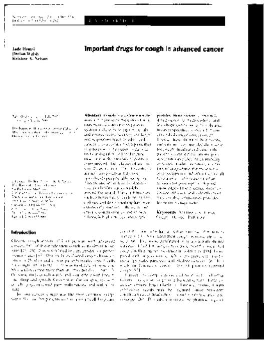

T-Type Calcium Channel Blockade in the Management of Chronic Ischemic Heart Disease

2C

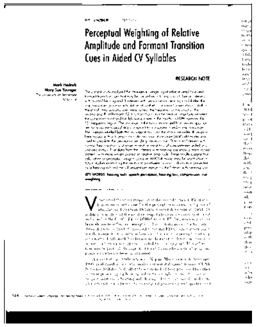

CME REVIEW ARTICLE 34

Postpartum Acquired Hemophilia (Factor VIII Inhibitors): A Case Report and Review of the Literature

1r2C

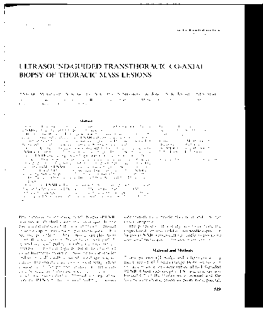

Important drugs for cough in advanced cancer

3C

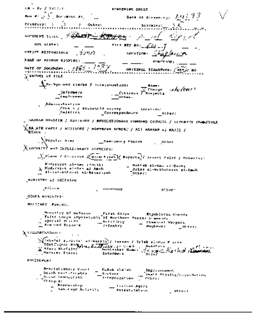

Perceptual Weighting of Relative Amplitude and Formant Transition Cues in Aided CV Syllables

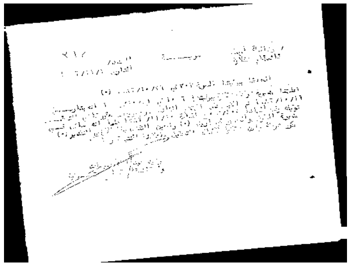
RESEARCH NOTE

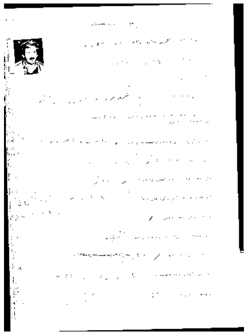
2c_asym

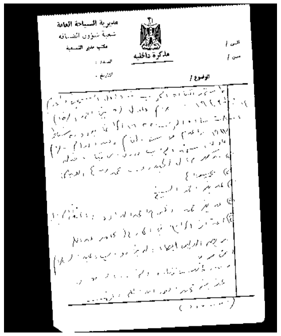

ULTRASOUND-GUIDED TRANSTHORACIC CO-AXIAL BIOPSY OF THORACIC MASS LESIONS

2c2c_asym


class1


class2

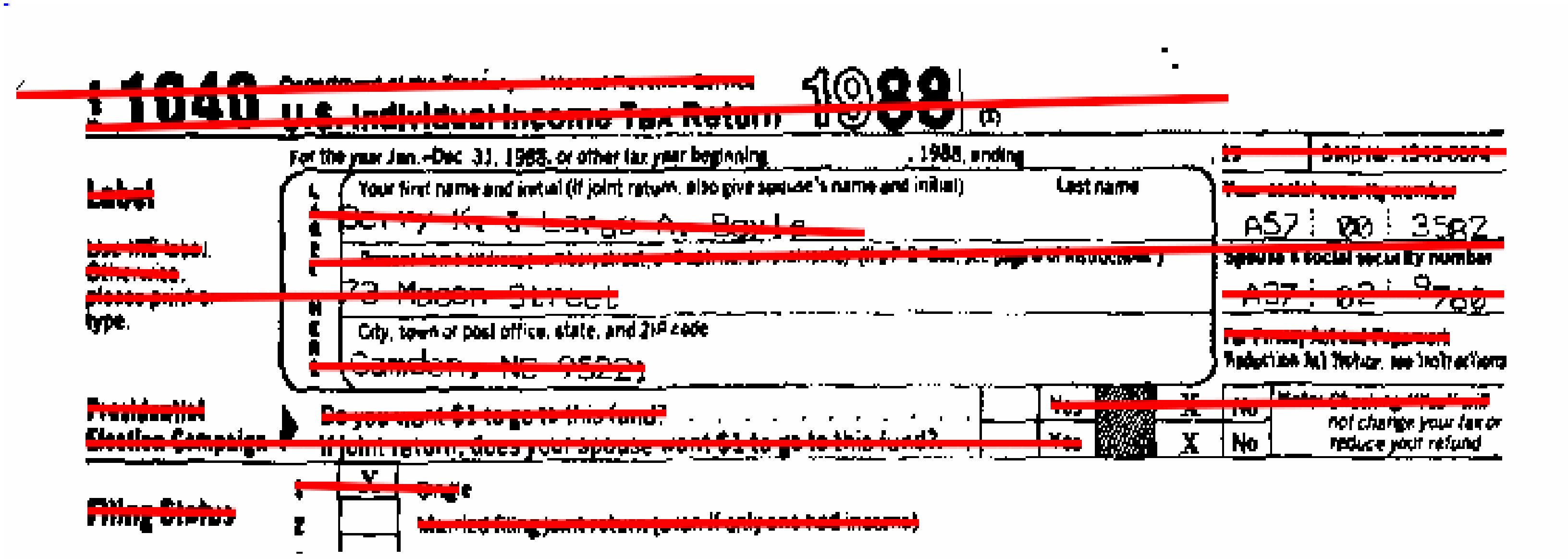

class3


class5

Document Representation

-- Building blocks

- Text lines extracted by TB library (endpoint coordinates, line orientations)



1040 U.S. Individual Income Tax Return 1988

For the year Jan.–Dec. 31, 1988, or other tax year beginning , 1988, ending

Label

Your first name and initial (if joint return, also give spouse's name and initial) Last name

Use IRS label. Otherwise, please print or type.

73 Mason Street

City, town or post office, state, and ZIP code

Your social security number A57 00 3582

Spouse's social security number

For Privacy Act and Paperwork Reduction Act Notice, see Instructions

Presidential Election Campaign

Do you want $1 to go to this fund? Yes X No

If joint return, does your spouse want $1 to go to this fund? Yes X No

Note: Checking "Yes" will not change your tax or reduce your refund

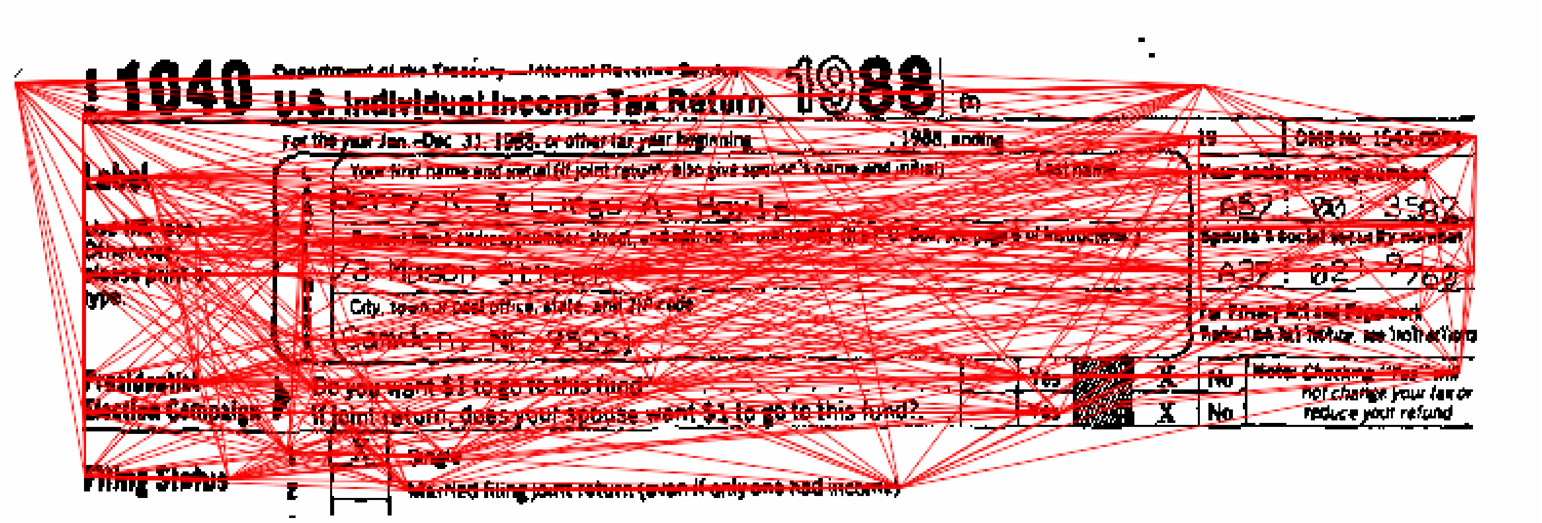
Filing Status

1 X Single

2 Married filing joint return (even if only one had income)

Quadrilaterals from text line pairs

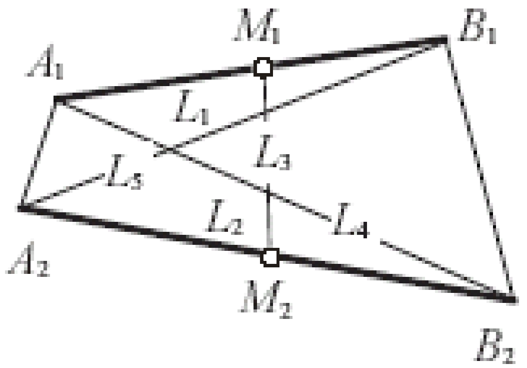
- A document := {Quadrilaterals}


1040
Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return
1988
For the year Jan.–Dec. 31, 1988, or other tax year beginning
, 1988, ending
19
OMB No. 1545-0074
Label
Your first name and initial (if joint return, also give spouse's name and initial)
Last name
Your social security number
Spouse's social security number
City, town or post office, state, and ZIP code
For Privacy Act and Paperwork Reduction Act Notice, see Instructions
Presidential Election Campaign
Do you want $1 to go to this fund?
If joint return, does your spouse want $1 to go to this fund?
Yes
No
X
Note: Checking "Yes" will not change your tax or reduce your refund
Filing Status
Single
Married filing joint return (even if only one had income)

- Merits:
  - Text line properties (length, orientation) are defined implicitly by their relative contribution to the quadrilateral shape
- Drawbacks:
  - O(n) → O(n^2)

Quadrilateral Shape Vector

- 5D shape vector



L_1, L_2: text lines

L_4, L_5: diagonals

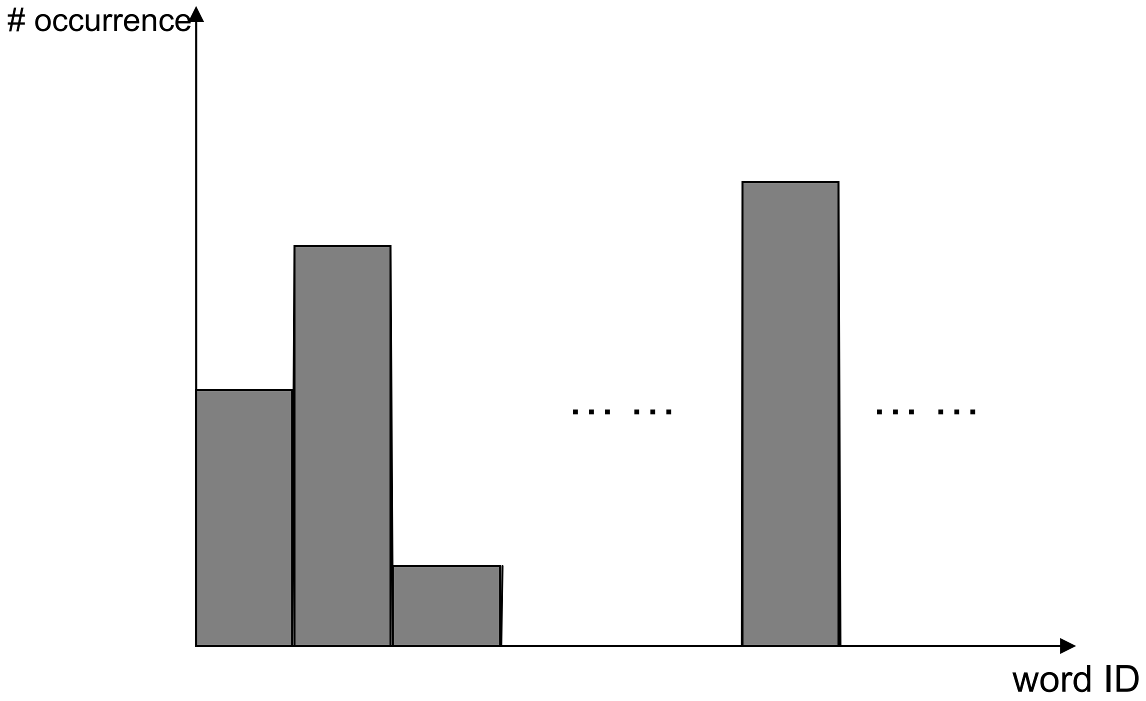
L_3: midpoints connection line

  - Vector uniquely defines the quadrilateral shape
  - Text line correspondence guaranteed
  - Efficient clustering
- Document represented this way is translation and 180° rotation invariant

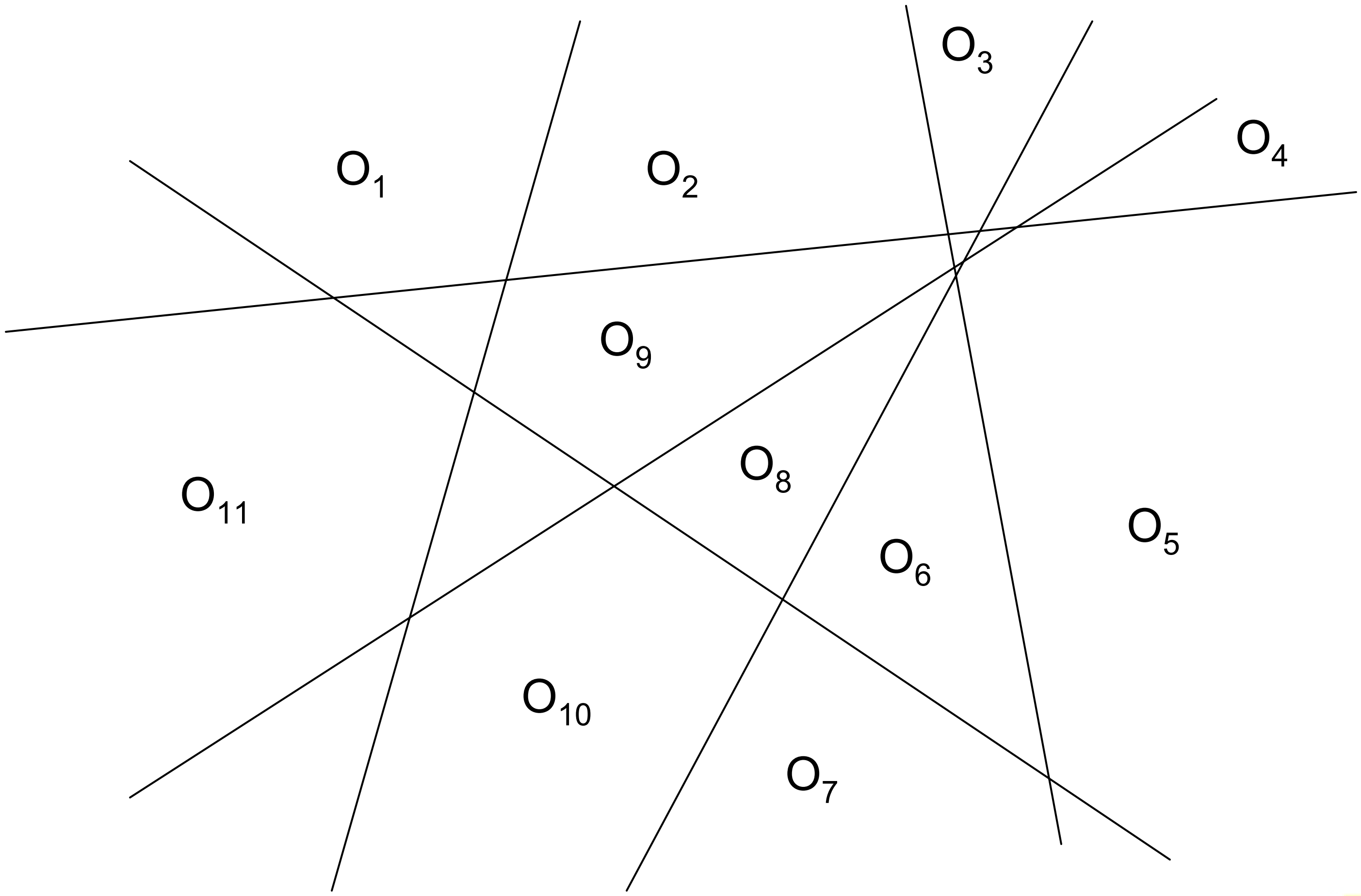
Dictionary of Quadrilaterals

- We need to establish correspondences between quadrilaterals so that documents comparison can break down into quadrilateral comparison.
- Clustering in 5D space using range search, each quadrilateral cluster is regarded as a word in the dictionary
- Need a rich dictionary to avoid too many unknowns in a query
- From 101 documents, we built a dictionary with 976 words

Document Representation



Random Chop – the idea



Specifically

- For each layout class, we choose some training samples
- For i= 1 to NUM_CHOPS
  - Randomly chop layout classes into two classes
  - Validity checking of current chop
  - Feature Selection
  - Train a binary discriminative classifier using Logistic Regression on training samples
  - Evaluate the classifier on a validating set

Feature Selection

- Document histogram vector lies in a very high dimension space
- Select subset of features that is relevant to the chopping in consideration
- CMIM criterion : Conditional Mutual Information Maximization

  $v(1) = \text{argmax}\ I\ (Y; X_i)\ \ 1 \leq i \leq N$

  $v(k+1) = \text{argmax}\ \{\min I(Y;X_i|X_{v(l)})\}\ ,\ 1 \leq k \leq K,\ l \leq k$
- Stopping criteria:
  - Maximum number of selected features is reached
  - Information gain is lower than a threshold

Score a query document

- Each document has a signature *S* like

1	0	0	1	0	… …	1	1

- Each layout class has a relaxed signature *RS* averaged from training samples. (consistency)

0.9	0.1	0.12	1	0.07	… …	0.875

- Each classifier has a performance value *P* on validation set. (discriminativity)

0.75	0.8	0.66	0.55	0.7	… …	0.6

- Score of a query against layout class i

$$\text{Score}_i = \sum_k F(S_k , RS_{i,k}) * P_k$$

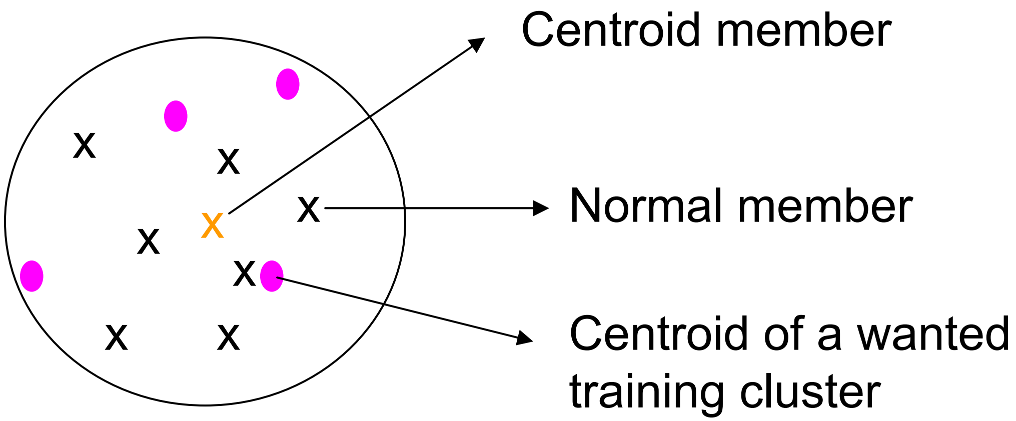
$$C = \text{argmax}_i \text{ Score}_i$$

Scoring a testing document

- $S = (\sum_i N_i * W_i) / (\sum_i N_i)$

  W_i : weight of the wanted training cluster which is the nearest neighbor within fixed range of testing cluster i.

  N_i : size of cluster i.



Centroid member

Normal member

Centroid of a wanted training cluster

Original: neighbor with the highest weight;

Lamp: nearest neighbor

Evaluation Scheme

- Mean Average Precision (MAP)
  - $P_i = (\sum_{i \leq j} P_j) / (\sum_{i \leq j} 1)$
- Average Relevance Rank (ARR)
  - $\boldsymbol{I} = (\sum(R_i - (N_t+1)/2)) / (N * N_t)$

    R_i : rank of one wanted testing document.

    N : testing size

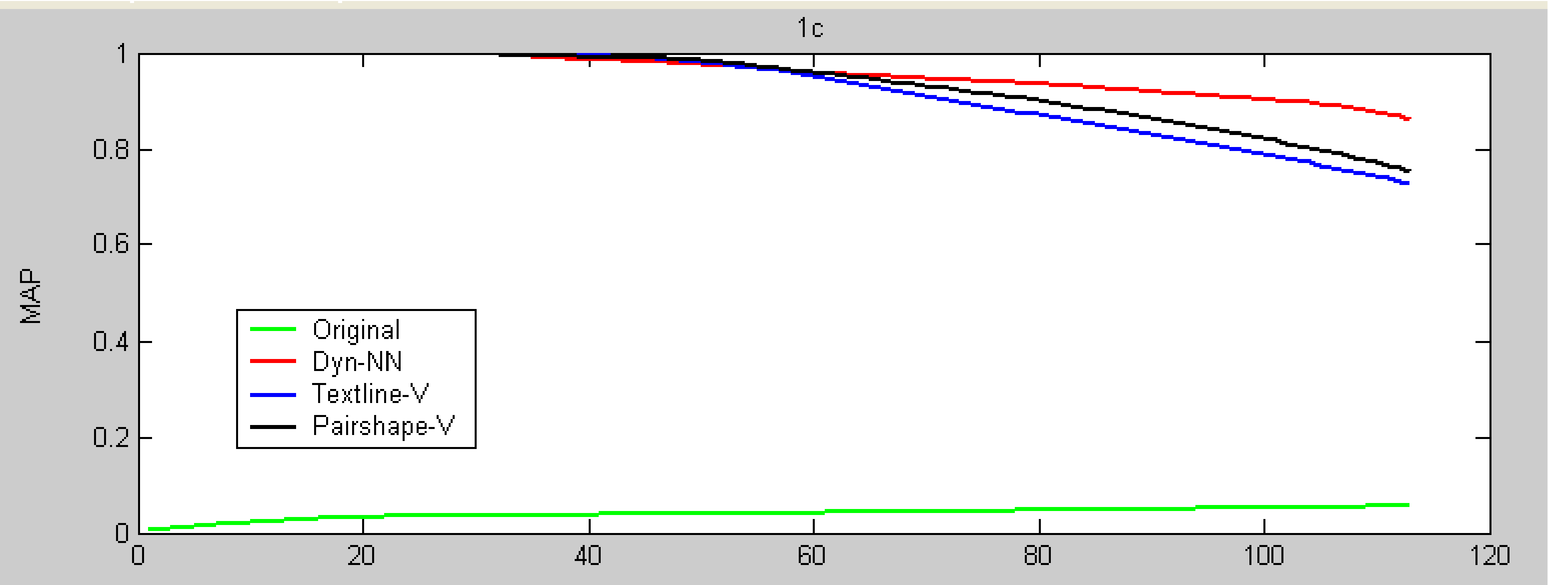
    N_t: wanted testing size
  - $\boldsymbol{I}$ Є $[0, 1-N_t/N)$, the lower the better

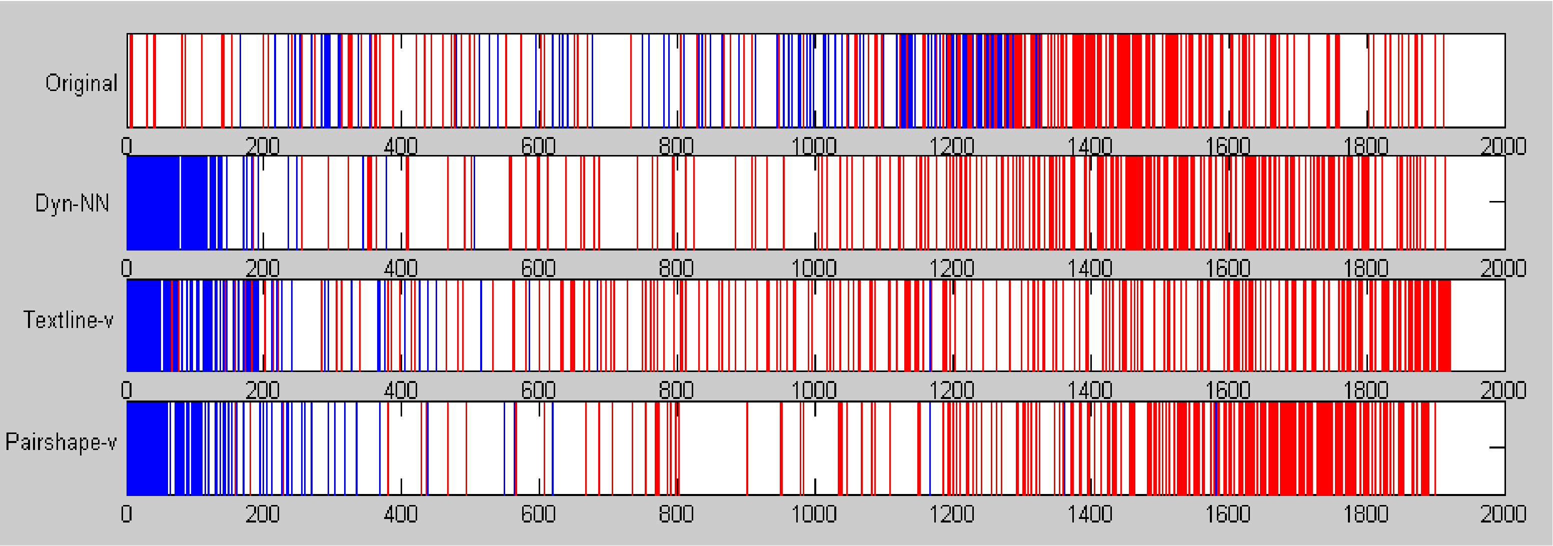
Experimental Results
--Confusion Matrix

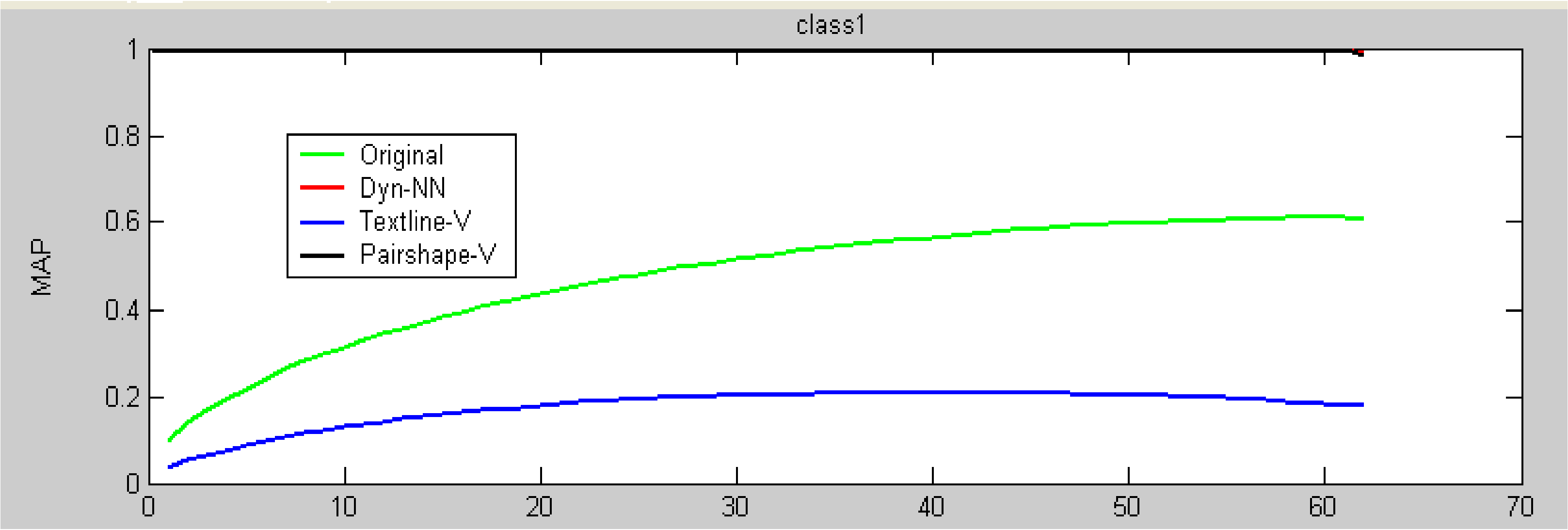
	1c	**2c**	**1r2c**	**3c**	**2c_asym**	**2c2c_asym**	**class 1**	**class 2**	**class3**	**class4**
1c (113)	87	8	16		2					
2c (144)		133	4	1		5	1			
1r2c (431)	9	168	246			8				
3c (23)				23						
2c_asym (6)					3	3				
2c2c_asym (45)		1				44				
Class1 (62)							62			
Class2 (264)	3					2	3	230	2	24
Class3 (121)	1			1			13	2	101	3
Class4 (95)				1		1	17	27	7	52

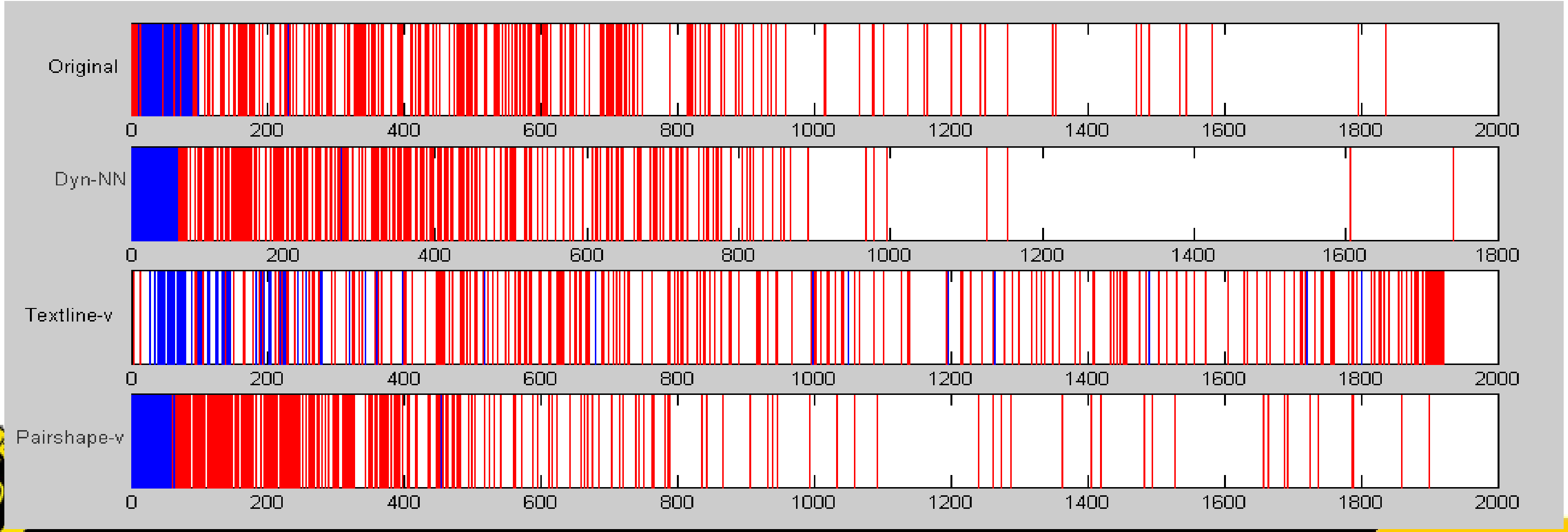
Experiments – ARR Results

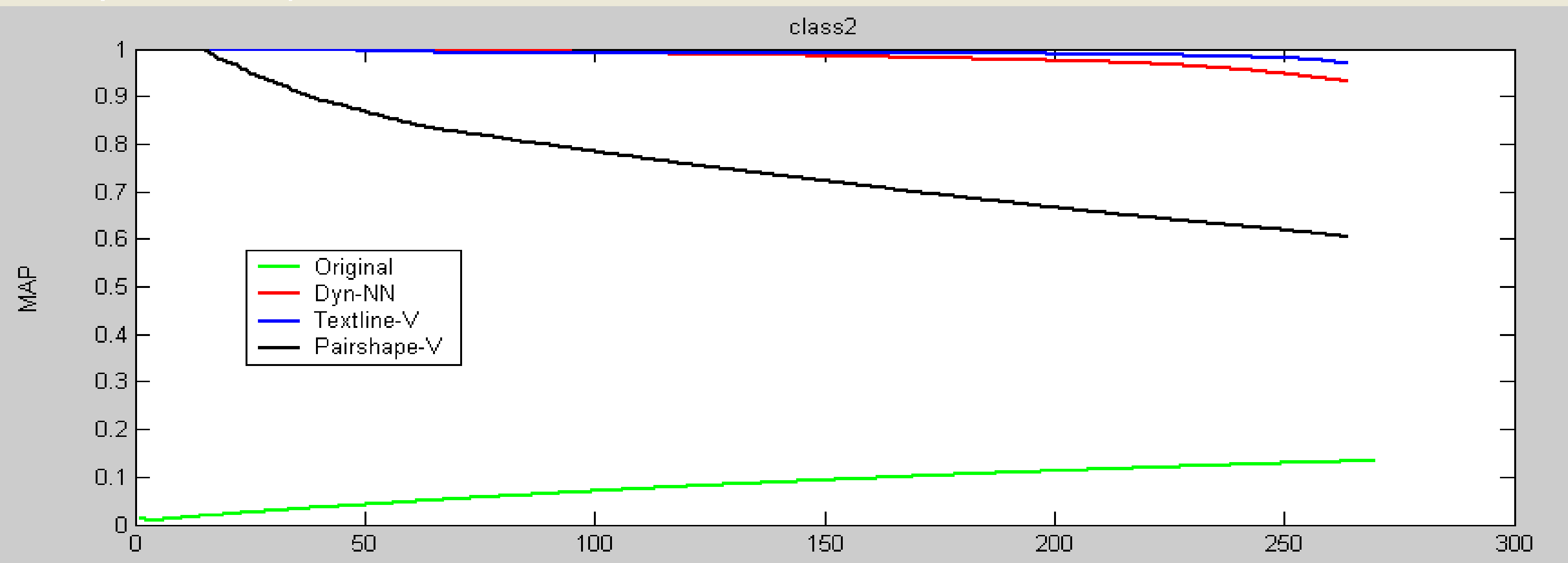
	Original	Dyn-NN	Text-V	Pair_V
1c	0.450	0.011	0.038	0.043
2c	0.062	0.010	0.324	0.087
3c	0.028	0.0002	0.504	0.013
1r2c	0.148	0.063	0.245	0.105
1r1r2c	0.159	0.010	0.103	0.045
1r2c2c	0.121	0.067	0.186	0.139
2c_asym	0.137	0.025	0.360	0.039
2c2c_asym	0.025	0.0002	0.097	0.010
class1	0.009	0.002	0.133	0.003
class2	0.398	0.011	0.004	0.075
class3	0.160	0.026	0.146	0.090
class5	0.302	0.056	0.103	0.085

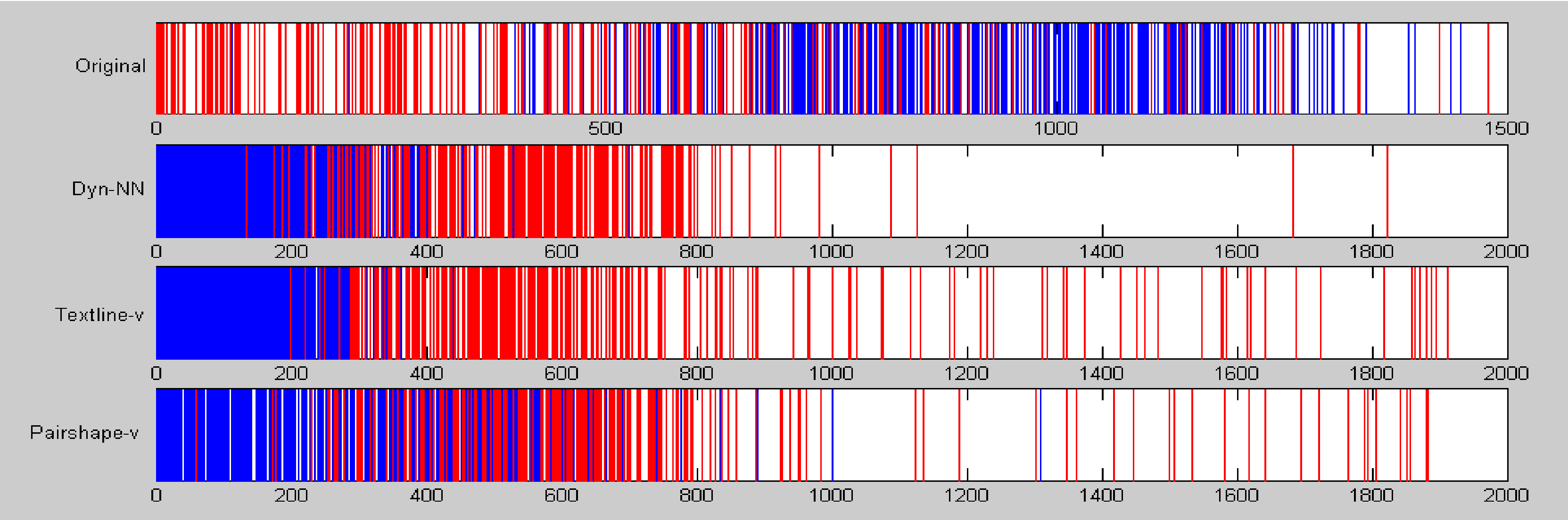

1c
MAP
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120
Original
Dyn-NN
Textline-V
Pairshape-V

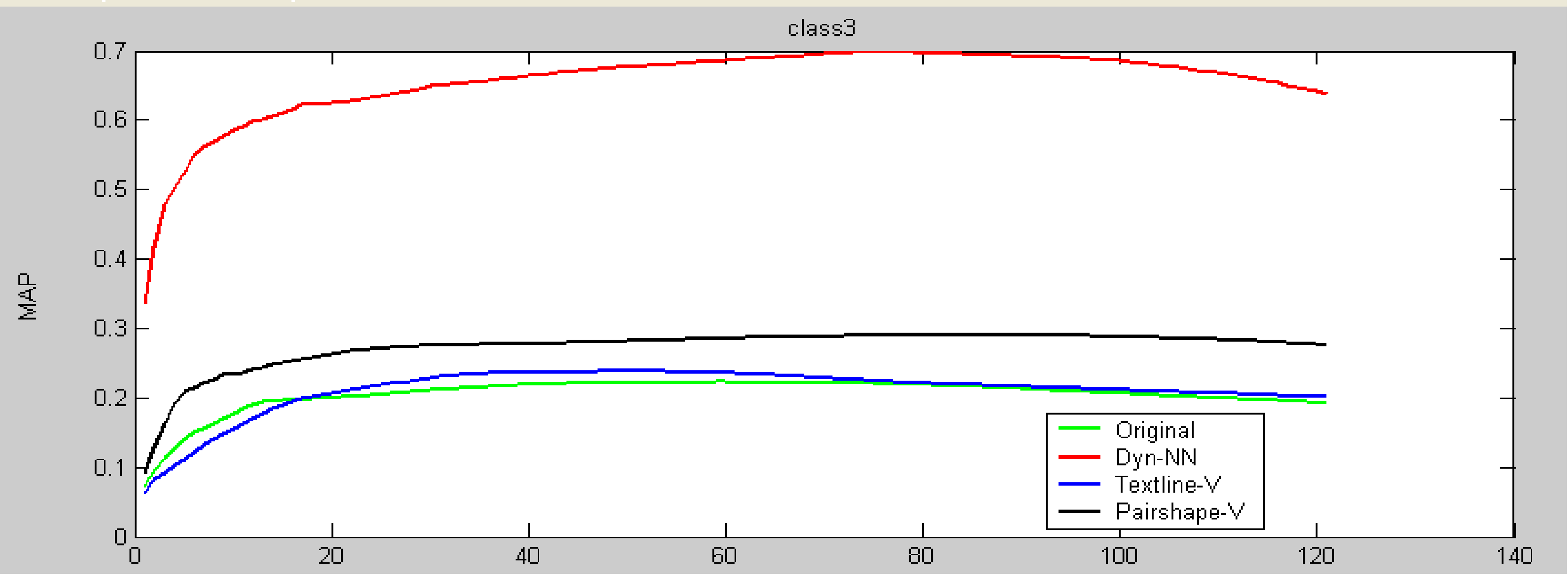

Original
Dyn-NN
Textline-v
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000

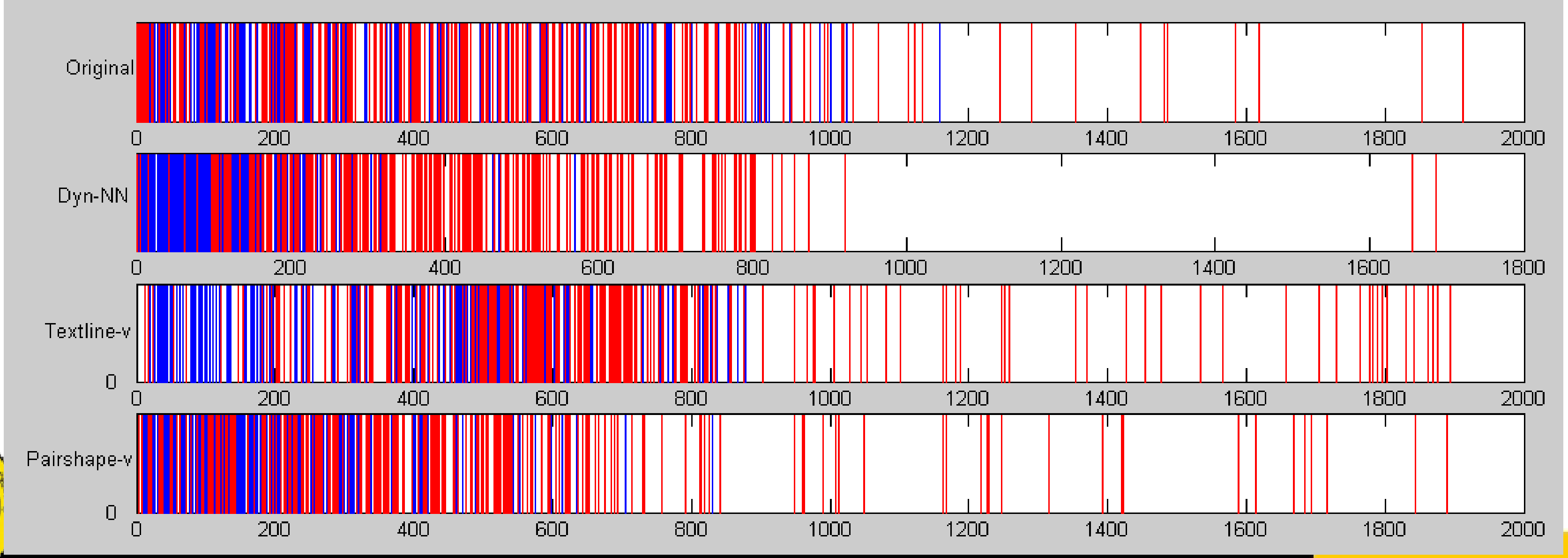

class1
MAP
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
Original
Dyn-NN
Textline-V
Pairshape-V

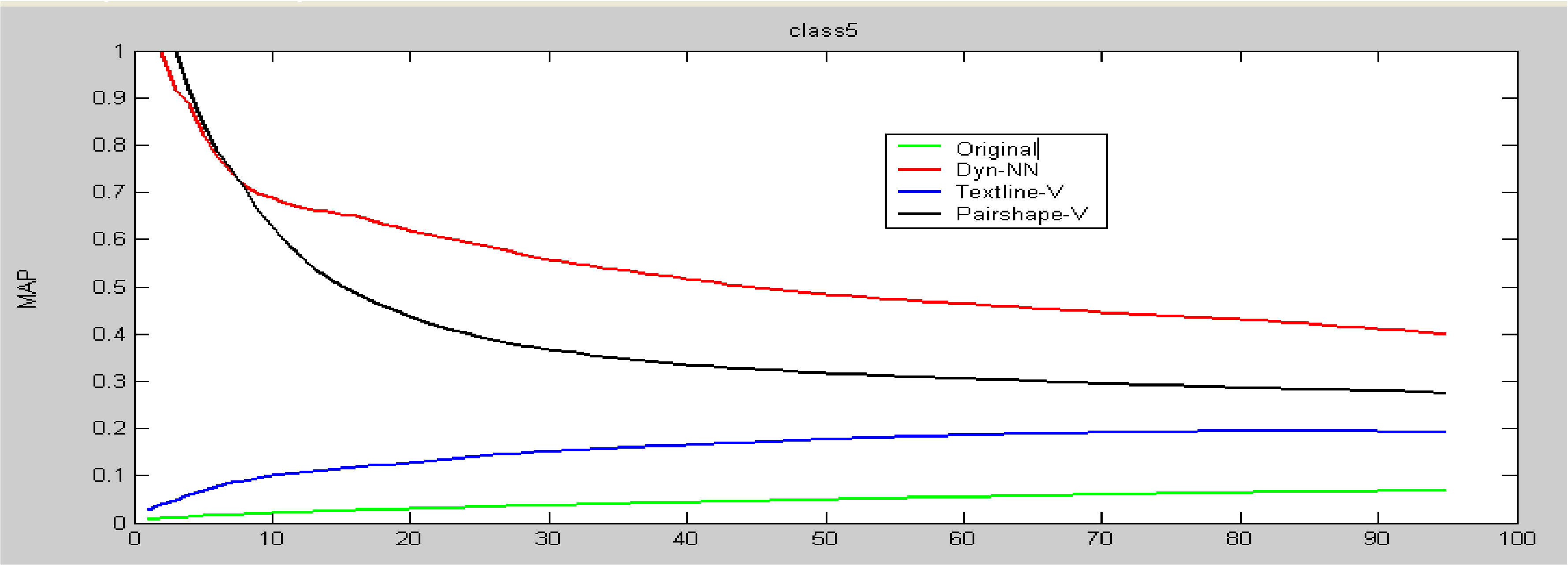

Original
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Dyn-NN
0
200
400
600
800
1000
1200
1400
1600
1800
Textline-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000

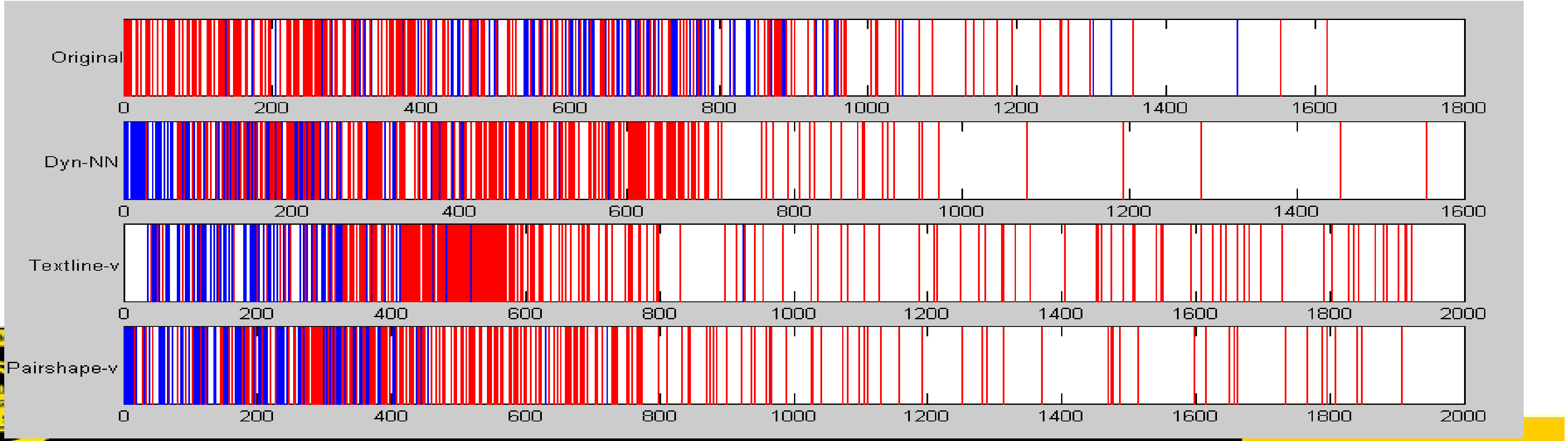

class2
MAP
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
Original
Dyn-NN
Textline-V
Pairshape-V


Original
0
500
1000
1500
Dyn-NN
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Textline-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000


class3
MAP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
20
40
60
80
100
120
140
Original
Dyn-NN
Textline-V
Pairshape-V


Original
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Dyn-NN
0
200
400
600
800
1000
1200
1400
1600
1800
Textline-v
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Pairshape-v
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000


class5
MAP
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
Original
Dyn-NN
Textline-V
Pairshape-V


Original
0
200
400
600
800
1000
1200
1400
1600
1800
Dyn-NN
0
200
400
600
800
1000
1200
1400
1600
Textline-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000

Summary and Future Work

- Conclusions:
  - Time efficiency
  - Space efficiency: only need to store classifier parameters and class signatures
  - Easy to combine new layout classes
  - Generalizability : is able to tell, to some degree, whether a new pair of instances unseen in the training set are of similar layout
- Future Work:
  - Find out the optimal number of chops for a given number of classes
  - Guarantee non-overlapping of classes
  - Try different classifiers, like NB, SVM

Software

- Currently Implemented as DocLib
- Line Detection Modules Improved

Technical Presentations

- Page Segmentation (and rule line separation)
- Page Layout Similarity
- Document ID/Script ID

This afternoon

- Logo Detection and Recognition
- Signature Detection
- Font OCR

Script and ImageID

- ScriptID
  - Given a set of handwritten document images, identify the scripts.
  - Dataset: UMD handwritten dataset + Arabic dataset
- ImageID
  - Given an arbitrary image, identify that it is
    - document image
    - image with text
    - Image w/o text
  - Dataset: ~3700 images from Internet.

ScriptID

- Motivation
- Challenges
- Observation
- Descriptor
- Implementation
- Results

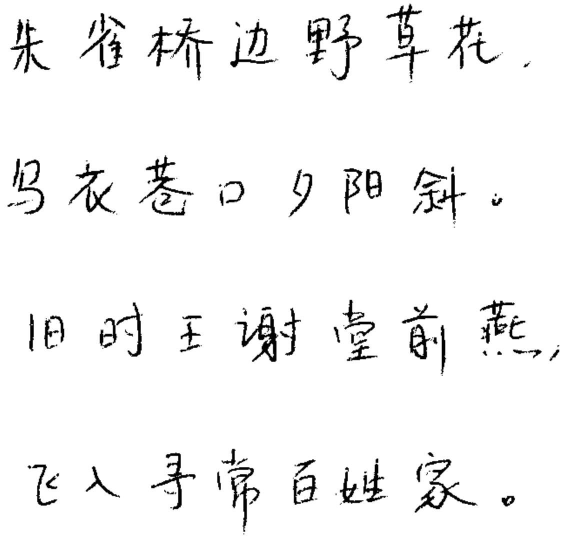
The Motivation

- Speedup the recognition process
  - Turn on the OCR engine, when necessary;
- Improve the accuracy
  - Select different OCR engines for different scripts;
- Understand the human perception
  - Can we recognize different scripts before recognizing individual characters?

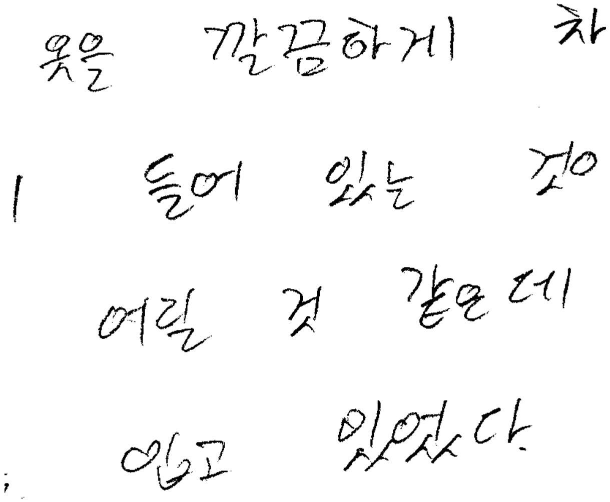
The Challenge

- Handwritten documents
  - Template matching cannot be used in general.
- The method needs to be fast
  - Naïve trial-and-error methodology doesn’t work
- The method needs to be invariant to
  - Scale
  - Rotation

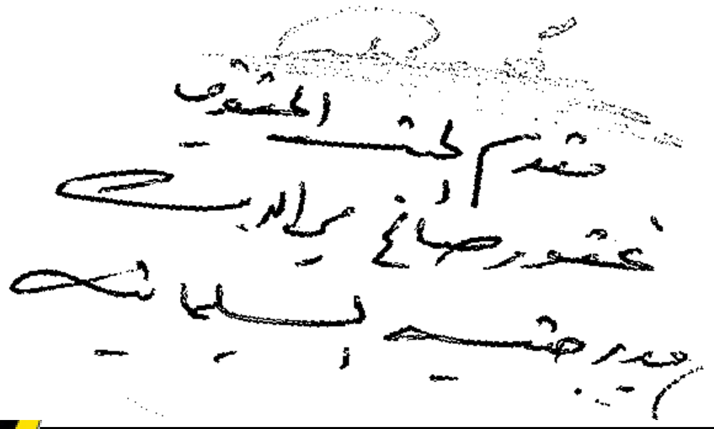
The Observation



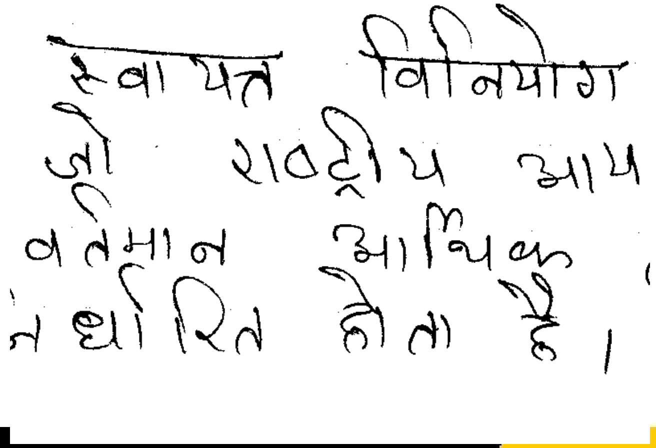
朱雀桥边野草花.
乌衣巷口夕阳斜。
旧时王谢堂前燕,
飞入寻常百姓家。



옷을 깔끔하게 차
들어 있는 것이
어린 것 같은데
입고 있었다.



المشق
[illegible]
[illegible]
[illegible]



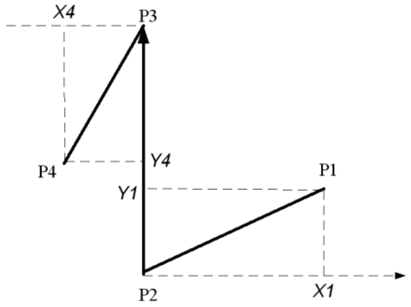
सेवा यत विनियोग
जो राष्ट्रीय आय
वर्तमान आर्थिक
निर्धारित होता है।

The Observation (con’t)

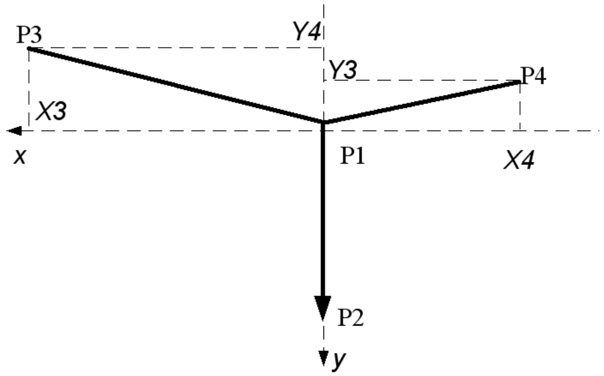
- The relationship of connected edges could be used for description;
- The dominant descriptors for different scripts could be different;
- The statistics of the descriptors could be used for discriminating different scripts.

The descriptor

- Fit edges to small lines
- Adjacent lines: encode the relative coordinates w.r.t pivot point.
  - C / Z shape



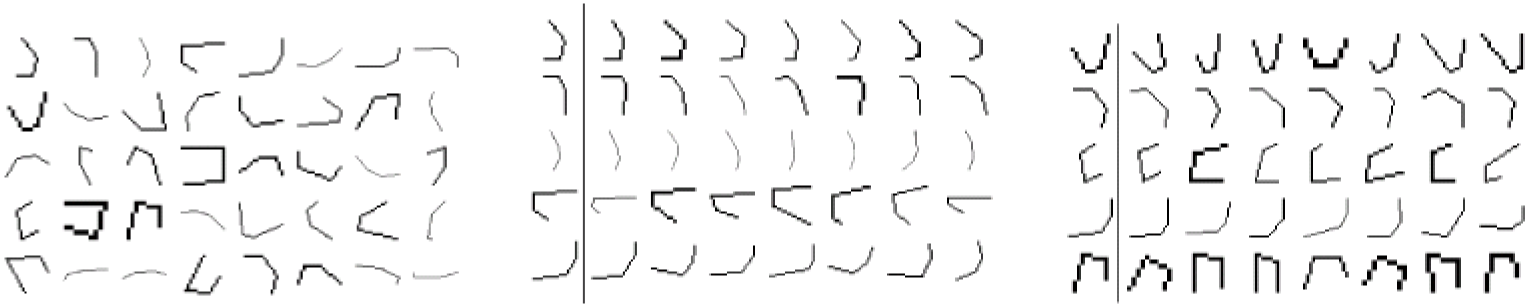
  - Y shape



Yu et al, Object Detection Using Shape Codebook, BMVC 2007

The codebook for the descriptor

- The advantage of the codebook
  - Generic
  - Quantization -> fast
- generate the codebook
  - A large dataset
  - Extract descriptor
  - Cluster the descriptor

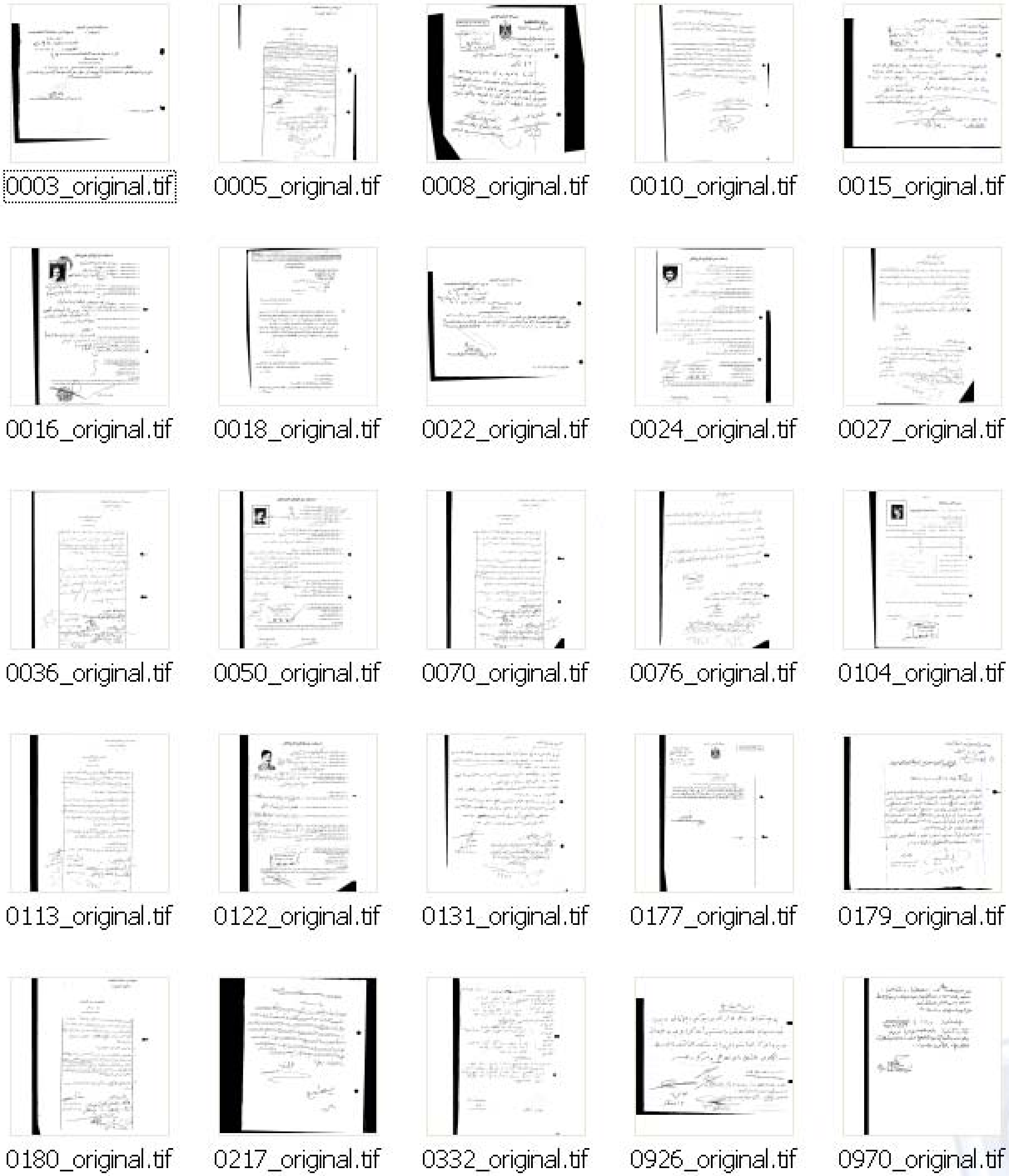


Classifier: Support Vector Machine

- Suppose we have *N* classes
- For each class, we train 1 SVM using images from this class vs other classes.
- Result: *N* SVM classifiers (linear classifier in high dimensional space)

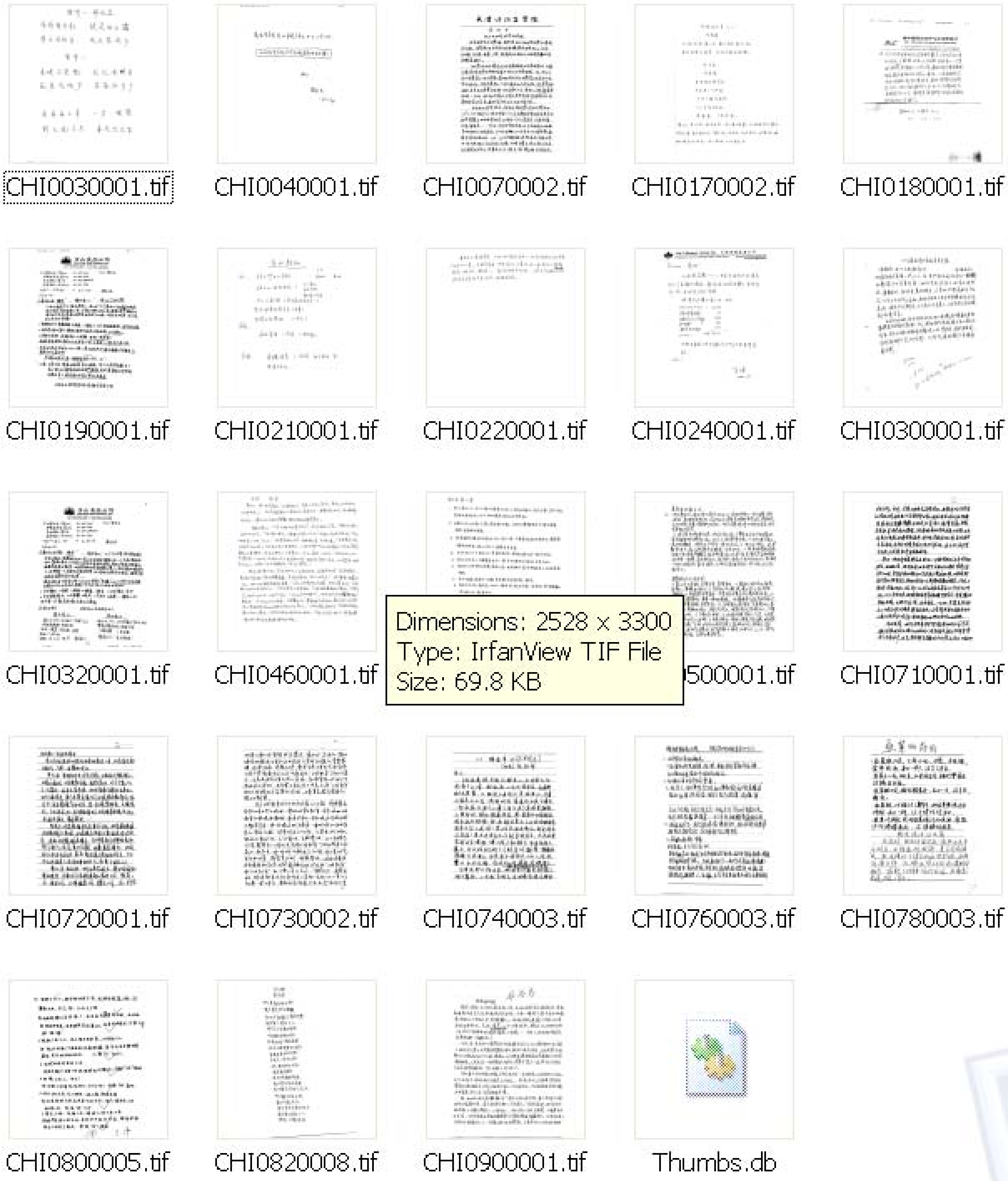
Dataset for Classification

- Arabic



Dataset for Classification

- Chinese



CHI0030001.tif CHI0040001.tif CHI0070002.tif CHI0170002.tif CHI0180001.tif

CHI0190001.tif CHI0210001.tif CHI0220001.tif CHI0240001.tif CHI0300001.tif

CHI0320001.tif CHI0460001.tif 500001.tif CHI0710001.tif

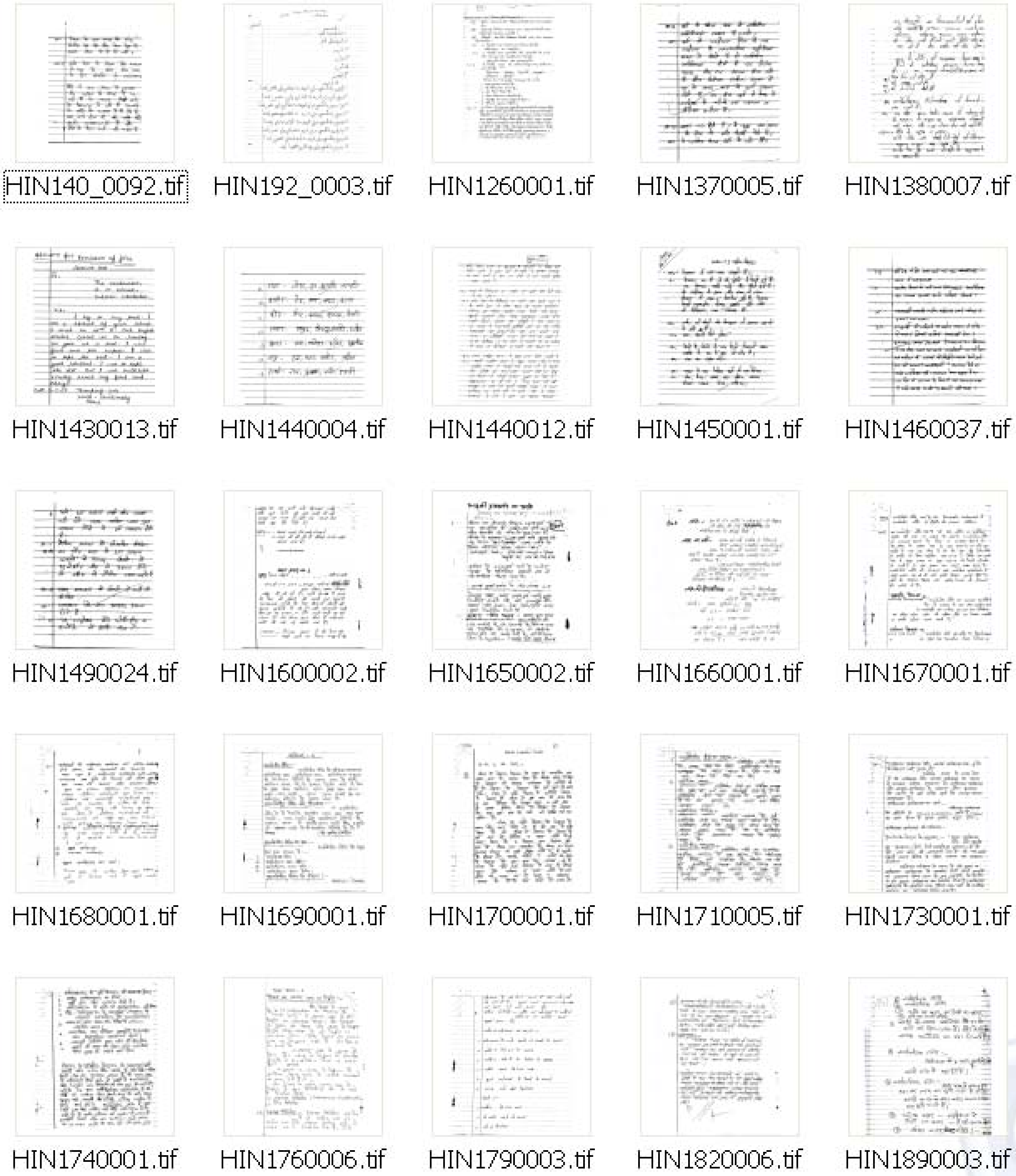
CHI0720001.tif CHI0730002.tif CHI0740003.tif CHI0760003.tif CHI0780003.tif

CHI0800005.tif CHI0820008.tif CHI0900001.tif Thumbs.db

Dimensions: 2528 x 3300
Type: IrfanView TIF File
Size: 69.8 KB

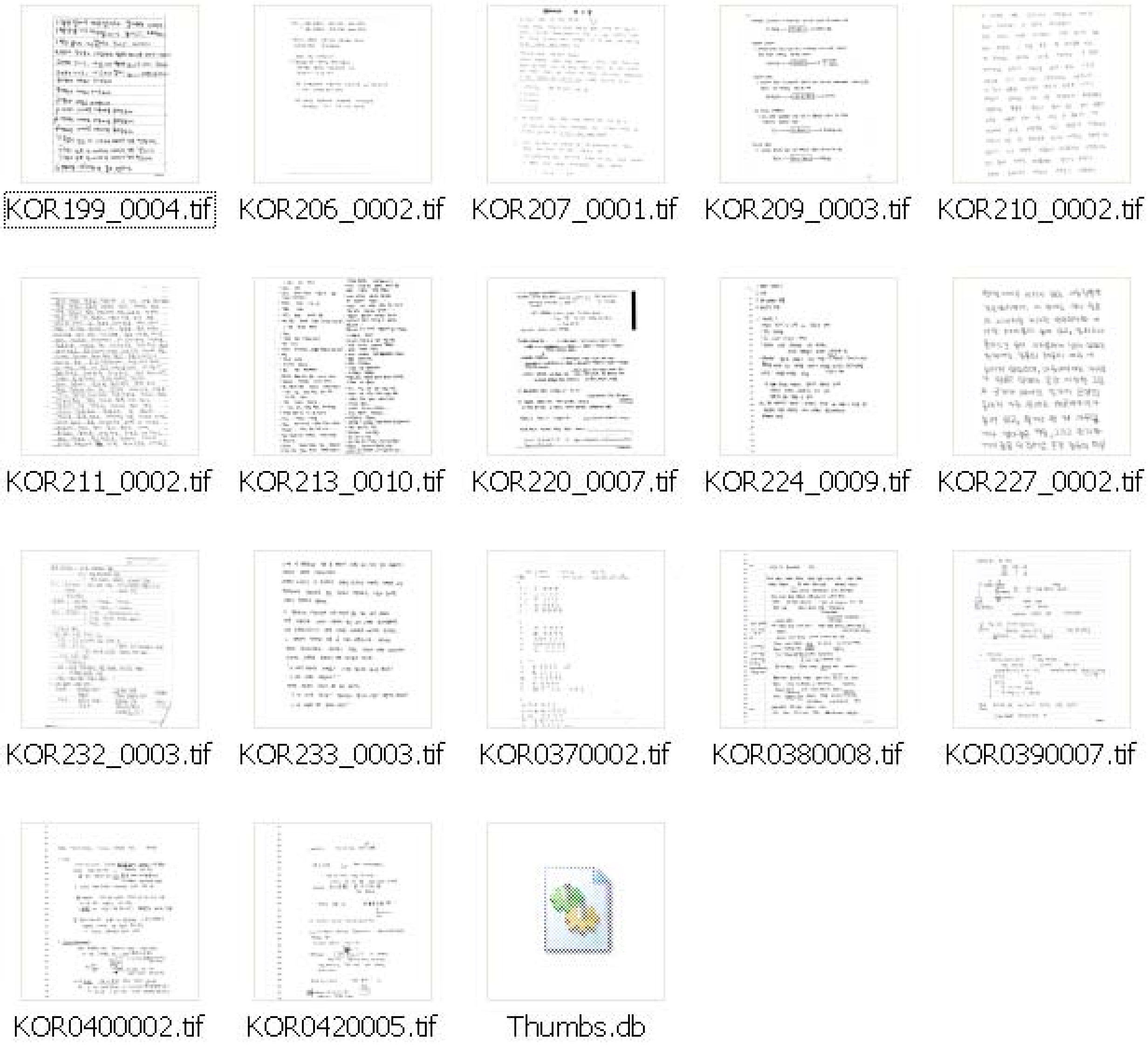
Dataset for Classification

- Hindi



Dataset for Classification

- Korean



The implementation

- Given a document image
  - Preprocessing
    - Binarize if necessary
    - Skeletonize
    - Clean the image using mathematical morphology.
  - Extract descriptors
    - Extract line segments
    - Compute shape descriptors
    - Quantize the shape descriptors and compute their histogram.
  - Train and classify

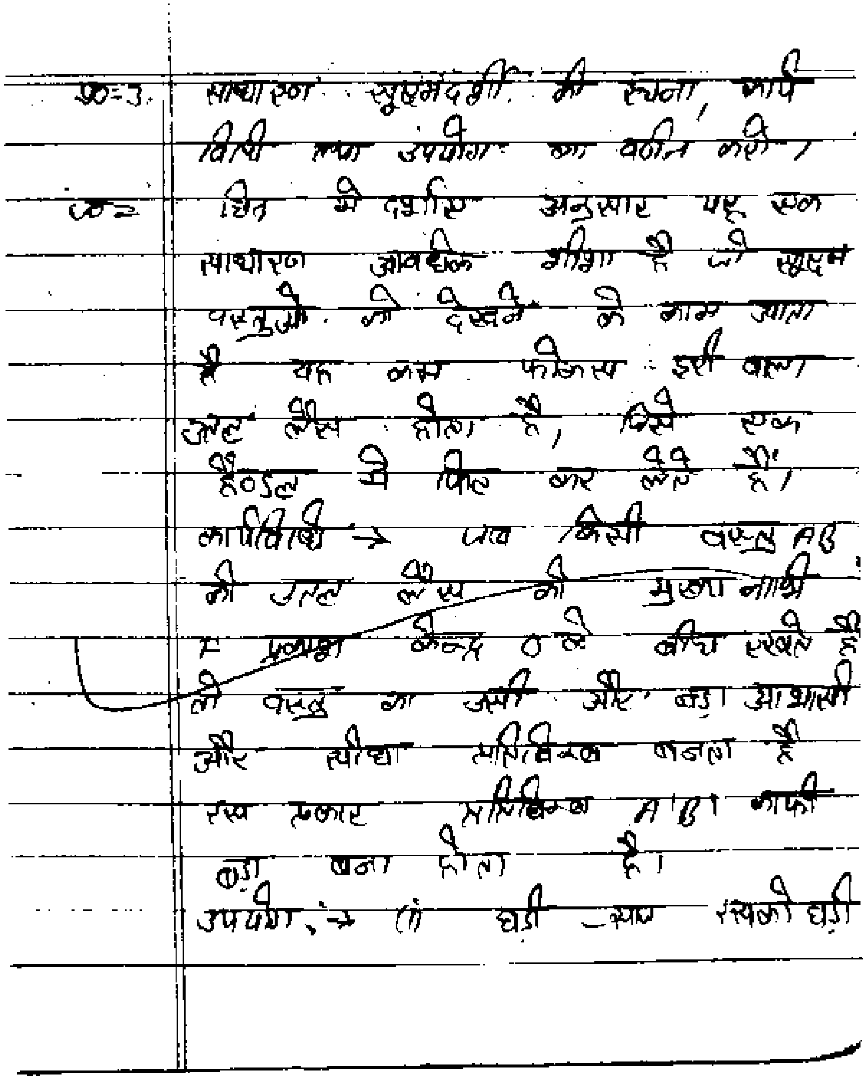
Result

- Confusion matrix (experimental result, july 2007)

	Arabic	Chinese	Hindi	Korean
Arabic	11 (74%)	1	2	1
Chinese	0	10 (77%)	0	3
Hindi	1	1	10 (83%)	0
Korean	1	3	0	9 (70%)

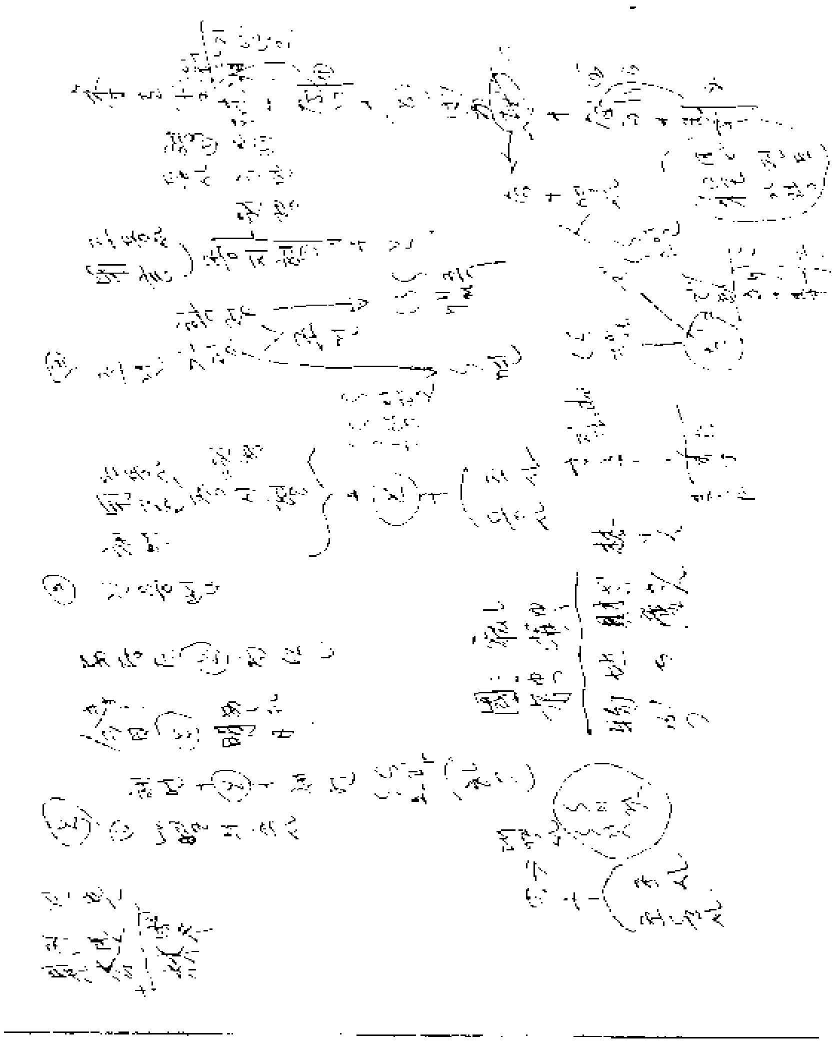
Failed examples

Arabic

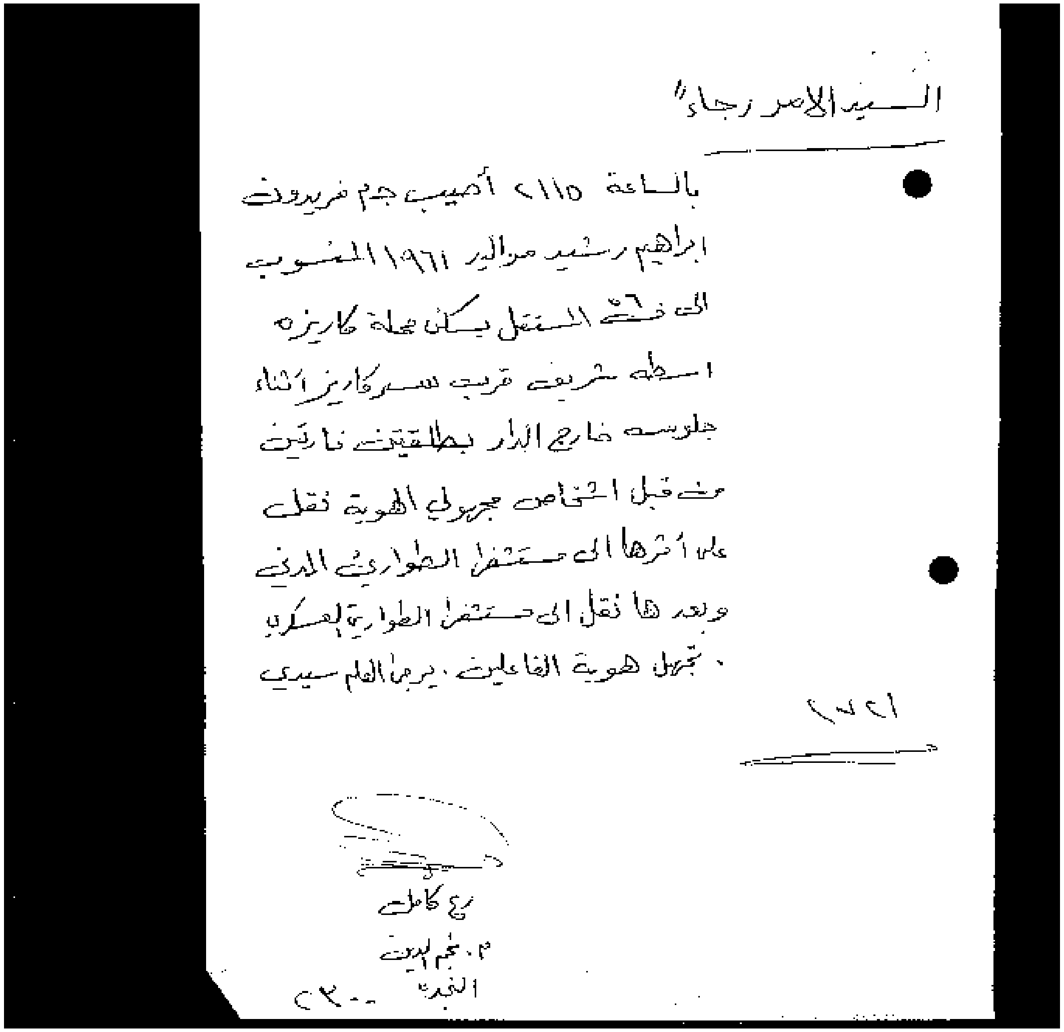


प्र०3. साधारण सूक्ष्मदर्शी की रचना, कार्य
विधि तथा उपयोग का वर्णन करो।
उ०- चित्र में दर्शाए अनुसार यह एक
साधारण आवर्धक शीशा है जो सूक्ष्म
वस्तुओं को देखने के काम आता
है यह कम फोकस दूरी वाला
उत्तल लैंस होता है, जिसे एक
हैंडल से फिट कर लेते हैं।
कार्यविधि → जब किसी वस्तु AB
को उत्तल लैंस की मुख्य नाभि
F व प्रकाश केन्द्र O के बीच रखते हैं
तो वस्तु का उसी ओर बड़ा आभासी
और सीधा प्रतिबिम्ब बनता है
इस प्रकार प्रतिबिम्ब A'B' काफी
बड़ा बना होता है।
उपयोग → (i) घड़ी साज रखकर घड़ी

Chinese



Failure example (Korean)



السيد الآمر رجاءً

بالساعة ٢١١٥ أصيب جرم فريدون
ابراهيم رشيد مواليد ١٩٦١ المنسوب
الى فوج٦ المستقل بسكن محلة كاريزه
اسطه شريف قرب سوركاريز اثناء
جلوسه خارج الدار بطلقتين ناريتين
من قبل اشخاص مجهولي الهوية نقل
على اثرها الى مستشفى الطوارئ المدني
وبعدها نقل الى مستشفى الطوارئ العسكري
. تجهل هوية الفاعلين . يرجى العلم سيدي

٢٧٢١

رع كامل
م.نجم الدين
النجدة
٢٠٠

ImageID

- Motivation
- Challenge
- The Approach
- Results

The Motivation

- Adopt different vision modules
  - For different categories we can adopt different strategy in computer vision
- Improve efficiency
  - Use the category as prior.
- Speedup OCR module in real world environment.

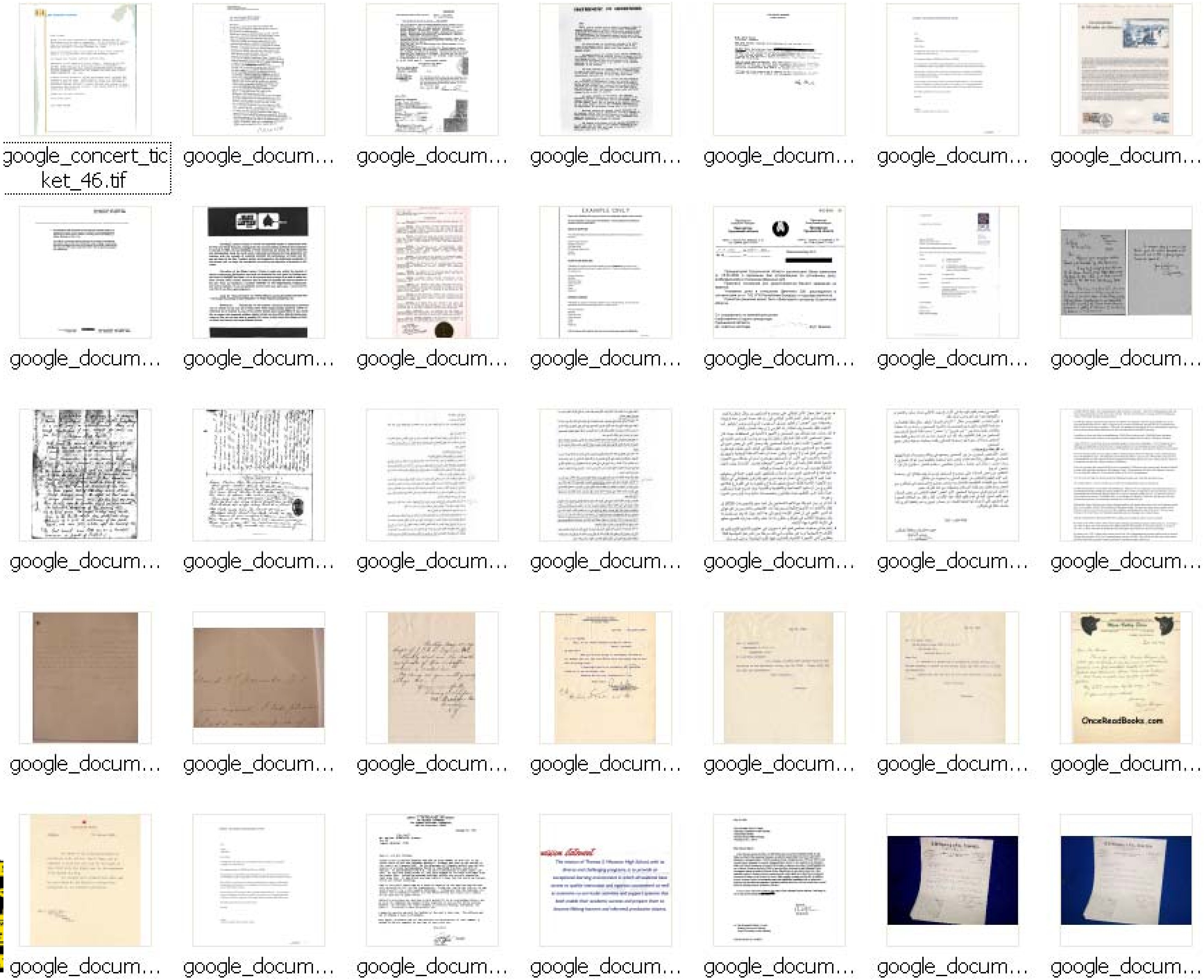
The Challenge

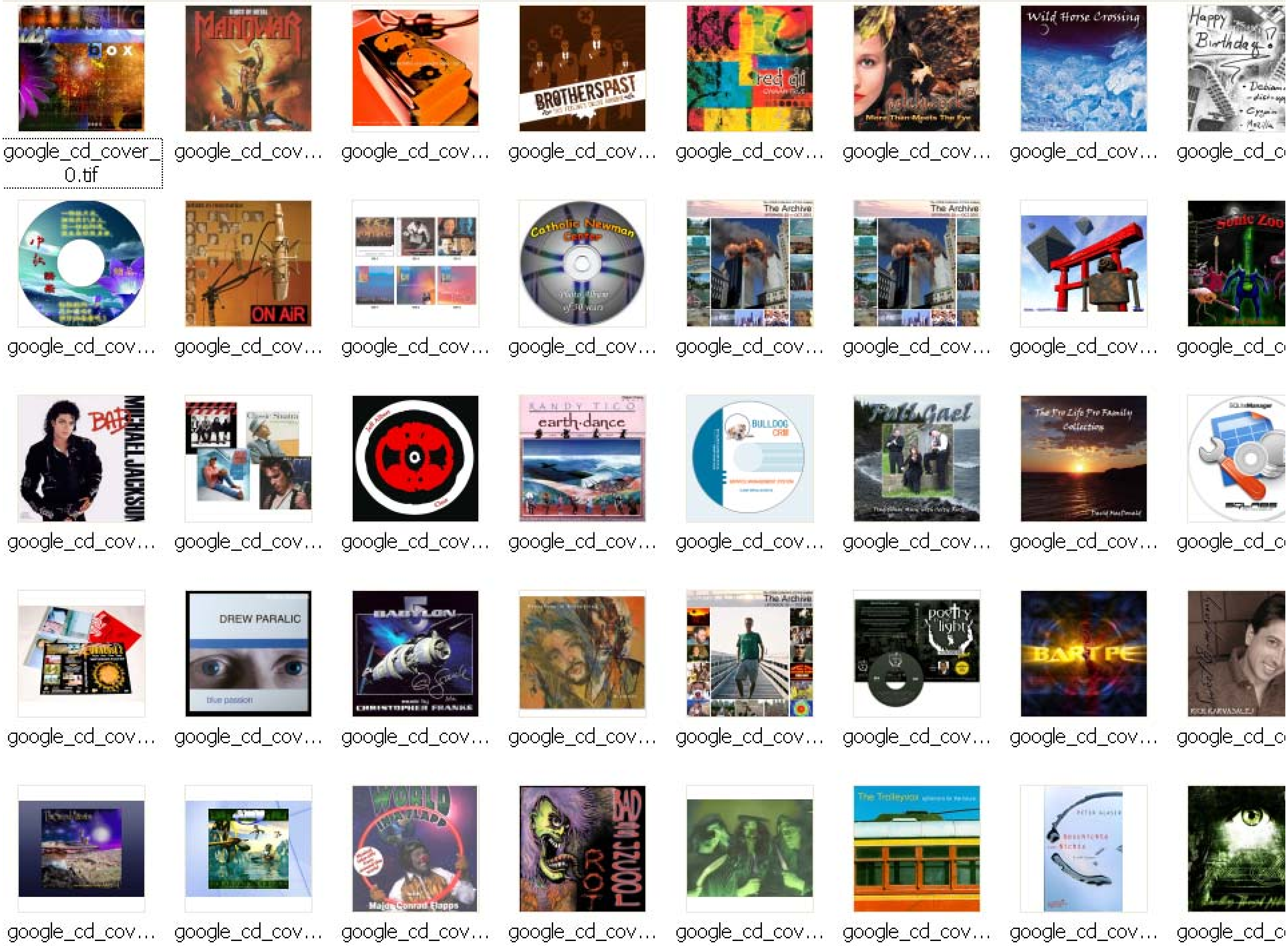
- Images are arbitrary
  - – Appearance model cannot be used for the classification.
  - – We use the same shape descriptor because the code book is generic.
- Ambiguity
  - – “images / text vs images”, e.g., Coke can.
  - – “doc vs images / text”, e.g. “publication cover” usually has figures.

Dataset for ImageID

- Collected form Internet, through search using different keywords
- Manual inspection, removal of duplicate images.

Page Classification Datasets (Google Image)	
Document	797
Image with Text	1695
Non-Document	1275
Total	3767


google_concert_ticket_46.tif
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…



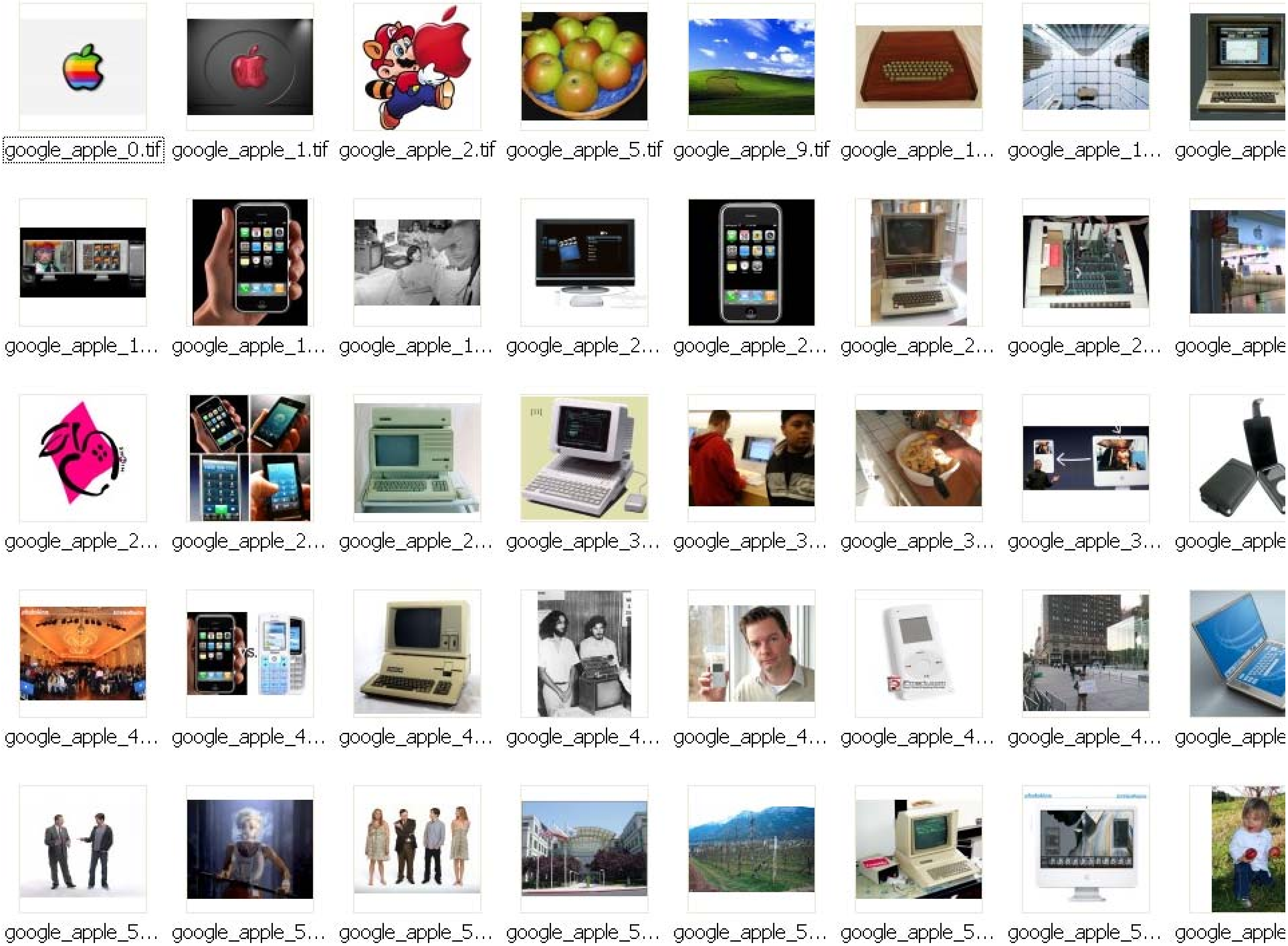
google_cd_cover_0.tif google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_c

google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_c

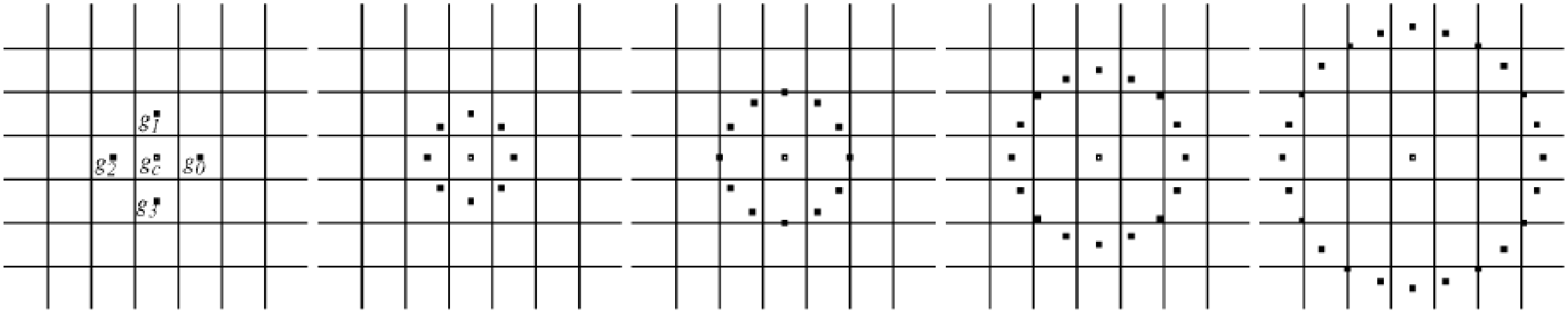
google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_c

google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_c

google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_cov… google_cd_c


google_apple_0.tif
google_apple_1.tif
google_apple_2.tif
google_apple_5.tif
google_apple_9.tif
google_apple_1...
google_apple_1...
google_apple
google_apple_1...
google_apple_1...
google_apple_1...
google_apple_2...
google_apple_2...
google_apple_2...
google_apple_2...
google_apple
google_apple_2...
google_apple_2...
google_apple_2...
google_apple_3...
google_apple_3...
google_apple_3...
google_apple_3...
google_apple
google_apple_4...
google_apple_4...
google_apple_4...
google_apple_4...
google_apple_4...
google_apple_4...
google_apple_4...
google_apple
google_apple_5...
google_apple_5...
google_apple_5...
google_apple_5...
google_apple_5...
google_apple_5...
google_apple_5...
google_apple

The Spatial Descriptor



(*P*=4,*R*=1.0) (*P*=8,*R*=1.0) (*P*=12,*R*=1.5) (*P*=16,*R*=2.0) (*P*=24,*R*=3.0)

- P: number of neighbor pixels
- R: neighbor size

LBP: Local Binary Pattern

- Define
  - Texture: Joint distribution of center pog_ct given neighbor sampling_p (p=0,…, P-1)

$$T = t(g_c, g_0, \ldots, g_{P-1})$$

- Example

g_3	g_2	g_1
g_4	g_c	g_0
g_5	g_6	g_7

The LBP representation

- Given an image.
- Transform the distribution vector into an P-bit pattern code ("Binary pattern")

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\ (g_p - g_c)\ 2^p$$

  - s: scale factor

Other variations of LBP

- Rotation invariant
- Different neighbor points and area
- “uniform” pattern

The performance

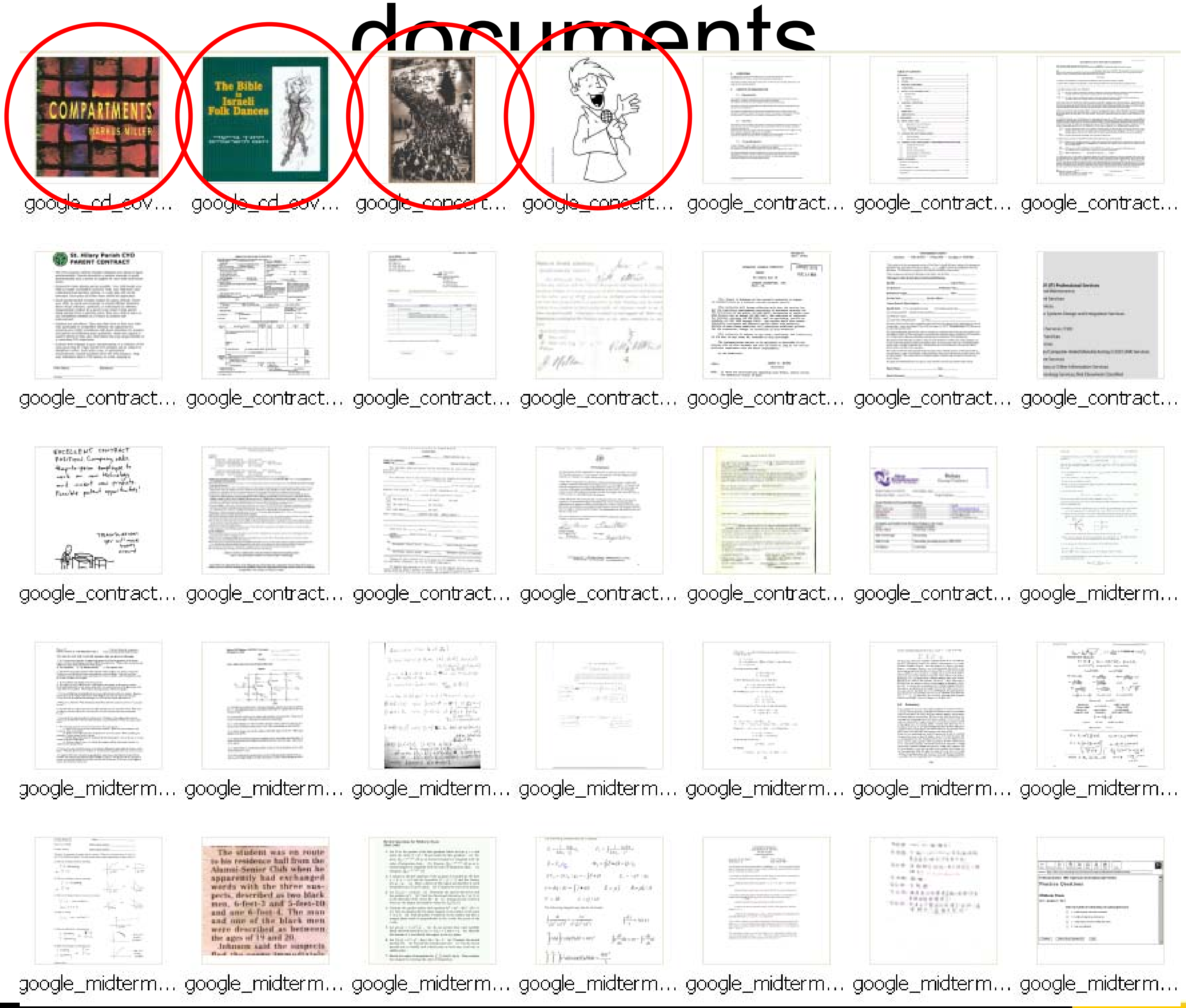
- Confusion matrix

	Doc	Image w/	Non doc
Doc	0.8557	0.1340	0.0103
Image w/	0.1725	0.6011	0.2264
Non doc	0.0444	0.1422	0.8133

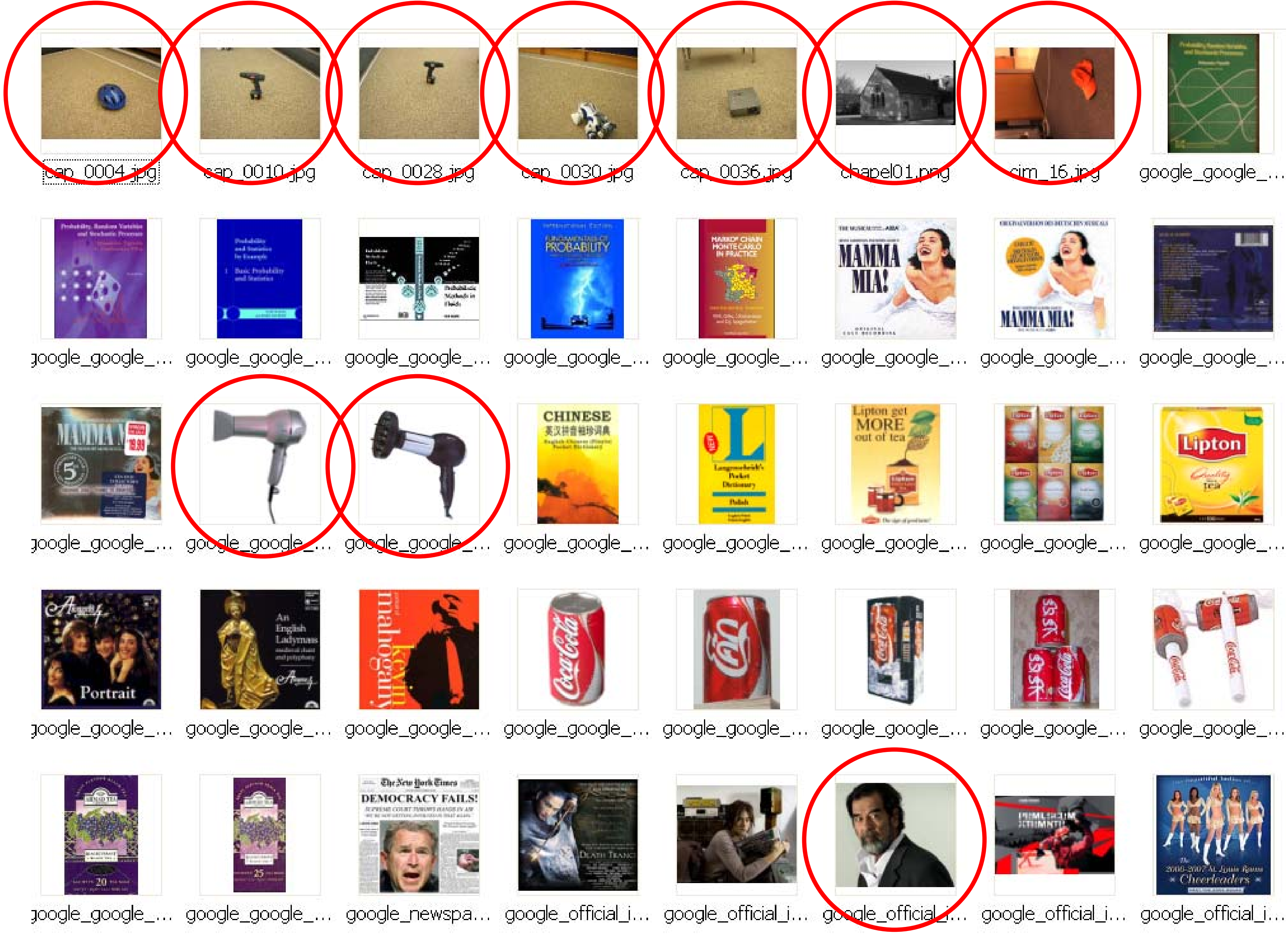
The Module

- Input
  - Training: an text file contains a list of training images.
  - Testing: a filename to an image.
- Output
  - Training: an SVM classifier (model.txt)
  - Testing: XML format (JEDI readable) for corresponding input image.
- Performance
  - 700 seconds for 3000 images
  - Similar speed for every image
  - No exceptions and memory leaks

Results – classified as documents



Images w/ text



cap_0004.jpg cap_0010.jpg cap_0028.jpg cap_0030.jpg cap_0036.jpg chapel01.png cim_16.jpg google_google_...

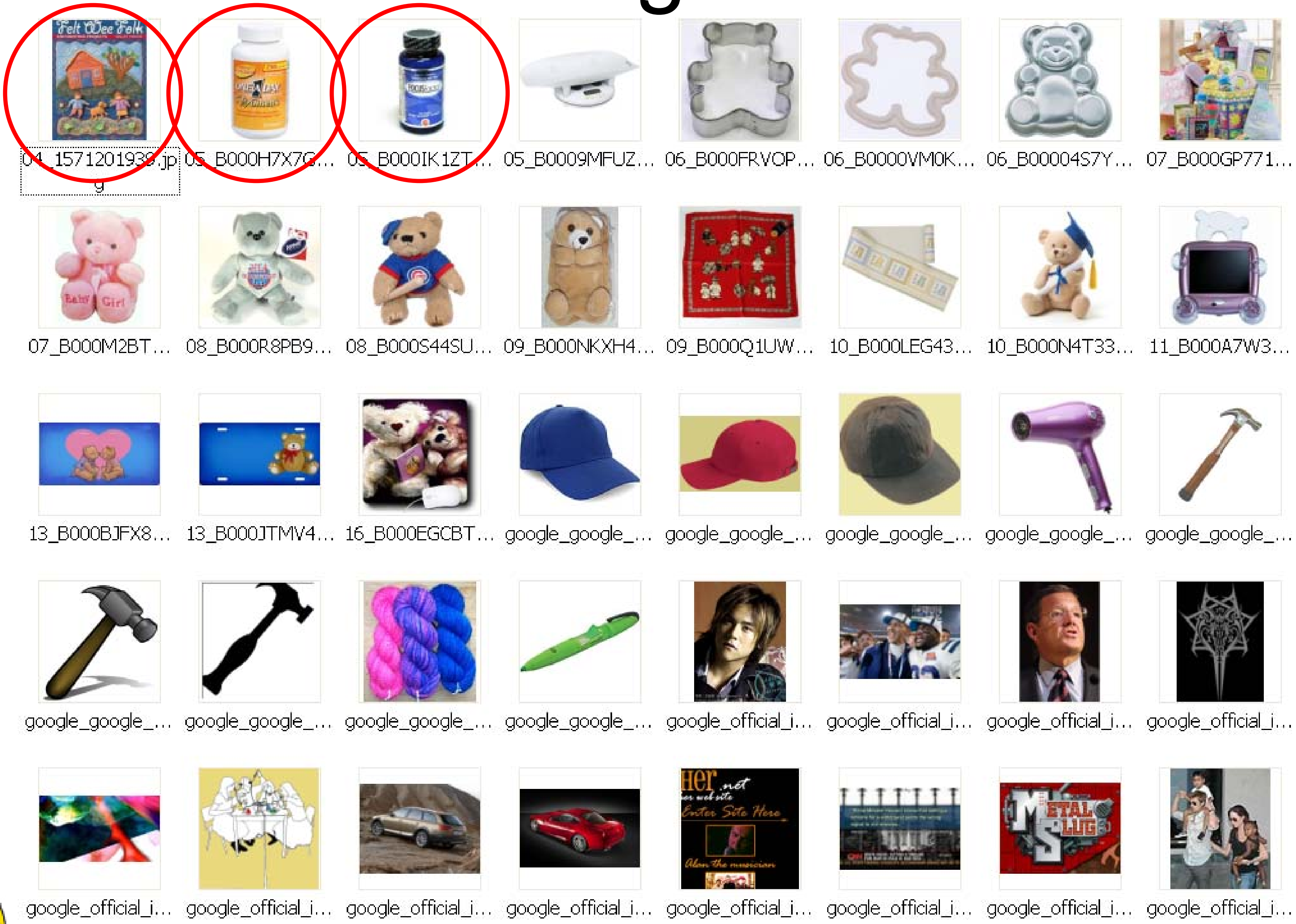
google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_...

google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_...

google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_... google_google_...

google_google_... google_google_... google_newspa... google_official_i... google_official_i... google_official_i... google_official_i... google_official_i...

Images



04_1571201939.jpg

05_B000H7X7G… 05_B000IK1ZT… 05_B0009MFUZ… 06_B000FRVOP… 06_B0000VM0K… 06_B00004S7Y… 07_B000GP771…

07_B000M2BT… 08_B000R8PB9… 08_B000S44SU… 09_B000NKXH4… 09_B000Q1UW… 10_B000LEG43… 10_B000N4T33… 11_B000A7W3…

13_B000BJFX8… 13_B000JTMV4… 16_B000EGCBT… google_google_… google_google_… google_google_… google_google_… google_google_…

google_google_… google_google_… google_google_… google_google_… google_official_i… google_official_i… google_official_i… google_official_i…

google_official_i… google_official_i… google_official_i… google_official_i… google_official_i… google_official_i… google_official_i… google_official_i…

Improvement

- Incorperate the distribution of grayscale:
  - an important clue for classification
- Try larger neighbor area for LBP
- Combine with other descriptors
  - Appearance model

Future work

- ScriptID
  - – Test more scripts. 10-15 would be a reasonable goal
- ImageID
  - – Improve the performance of the classification of the image w/ text vs images .

Technical Presentations

- Page Segmentation (and rule line separation)
- Page Layout Similarity
- Document ID/Script ID

This afternoon

- Logo Detection and Recognition
- Signature Detection
- Font OCR