CLEAT:

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

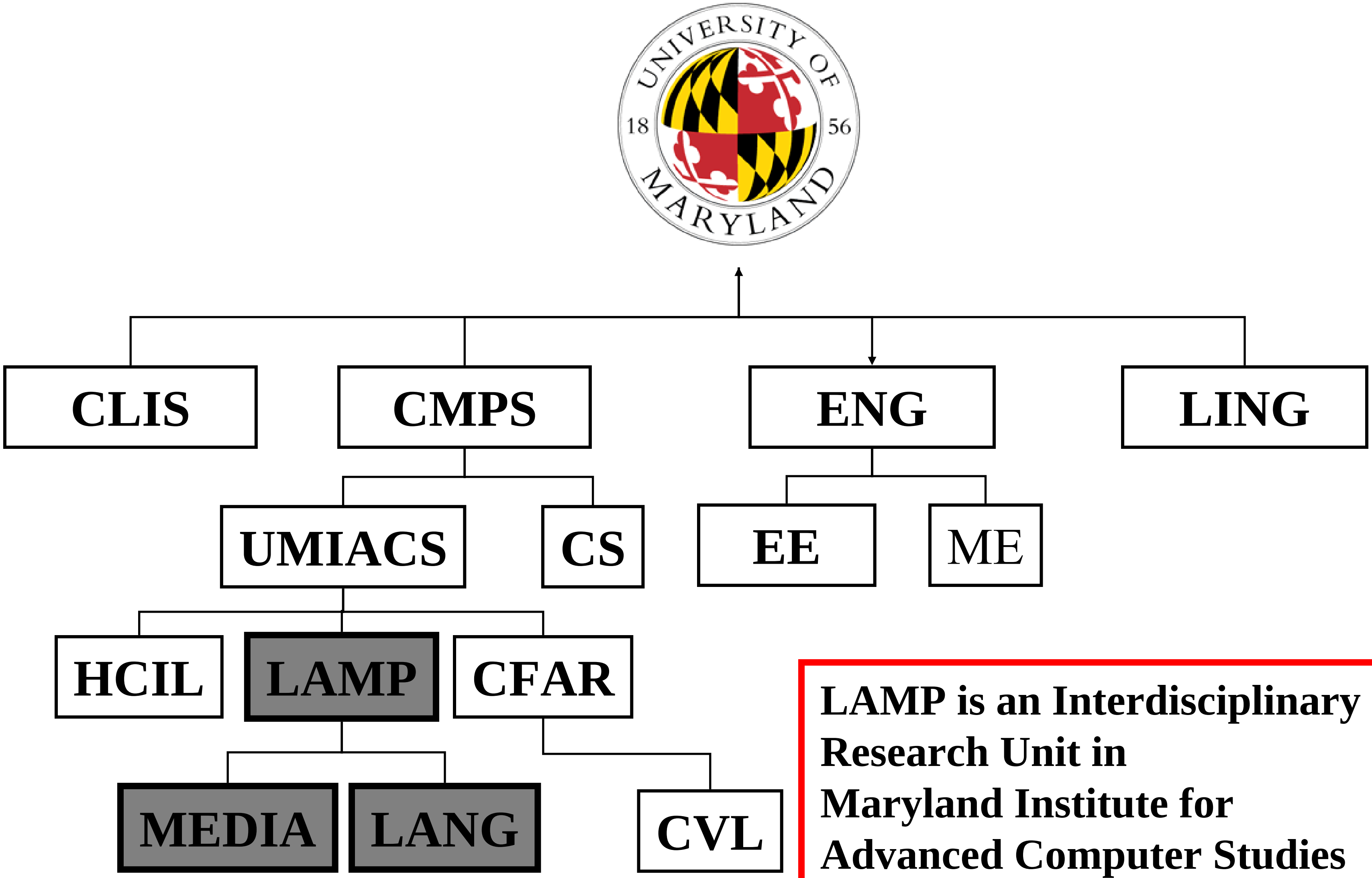
David Doermann
University of Maryland

Who are we?
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



UNIVERSITY OF MARYLAND 1856

CLIS

CMPS

ENG

LING

UMIACS

CS

EE

ME

HCIL

LAMP

CFAR

MEDIA

LANG

CVL

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

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

Research Focal Areas

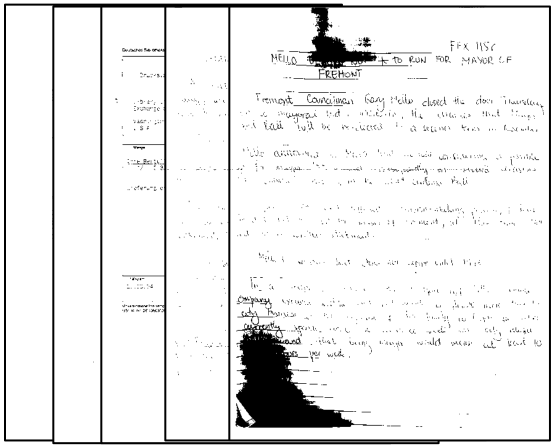
- Document image analysis
  - Providing fundamental tools for the enhancement summarization, navigation, indexing and retrieval in document image databases
- Content based video analysis
  - Providing access to video content through extraction, structure representation, classification, visualization 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

Intelligence Value Estimation

- How can we take **large, noisy, unstructured, heterogeneous** collections of image and video data to:
  - Mine the nuggets?
  - Bubble the important things to the top?
  - Provide tools for Information Discovery?

Challenges We Face….

- An overwhelming number of documents
  - Only a small fraction will ever be seen
- Huge variations in types, qualities and “value”
- Documents value diminishes with time
- We need to bring relevant documents to the top of the stack



FFX 1157

MELLO [illegible] TO RUN FOR MAYOR OF FREMONT

Fremont Councilman Gary Mello closed the door Thursday [illegible] mayoral bid [illegible] [illegible] will be [illegible] a second term [illegible]

Mello [illegible] [illegible] considering [illegible] possible [illegible] for mayor [illegible] [illegible] challenge Mello

[illegible] statement.

[illegible]

[illegible] company [illegible] city [illegible] currently [illegible] [illegible] that being mayor would mean at least 10 [illegible] per week.

Approach

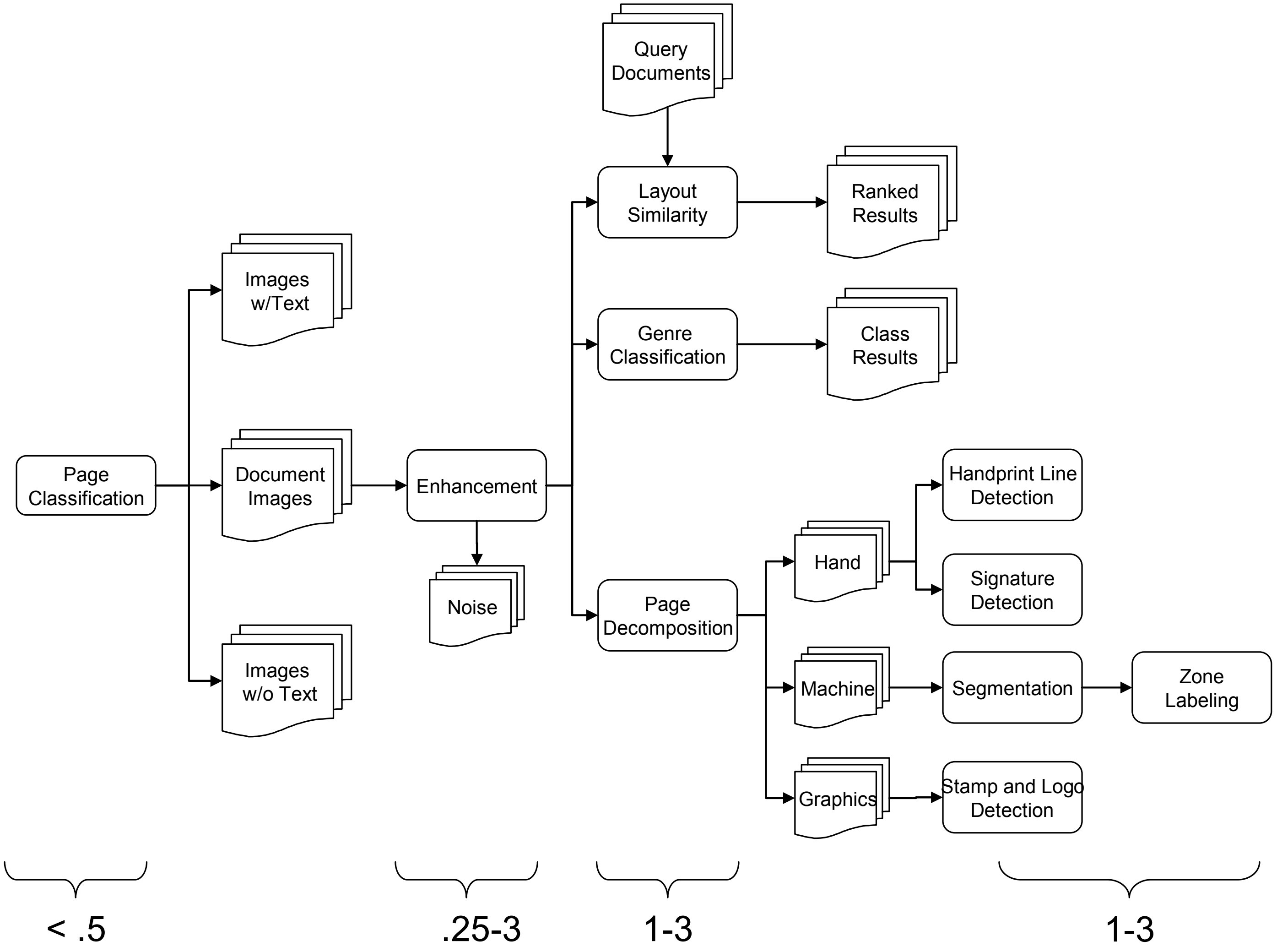
- Build robustness to noise into algorithms
  - Train noise as its own class
  - Integration of recognition and segmentation

- Provide mid level tools to organize collections
  - Genre Classification
  - Logo, Stamp and Signature Detection/Recognition

- Focus on Ranking rather then “conversion”
  - Page Layout Similarity

- Provide tools necessary for efficient research and evaluation
  - Datasets
  - GEDI – Groundtruth and Evaluation

Project Overview



Query Documents

Layout Similarity

Ranked Results

Images w/Text

Genre Classification

Class Results

Page Classification

Document Images

Enhancement

Handprint Line Detection

Hand

Signature Detection

Noise

Page Decomposition

Images w/o Text

Machine

Segmentation

Zone Labeling

Graphics

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

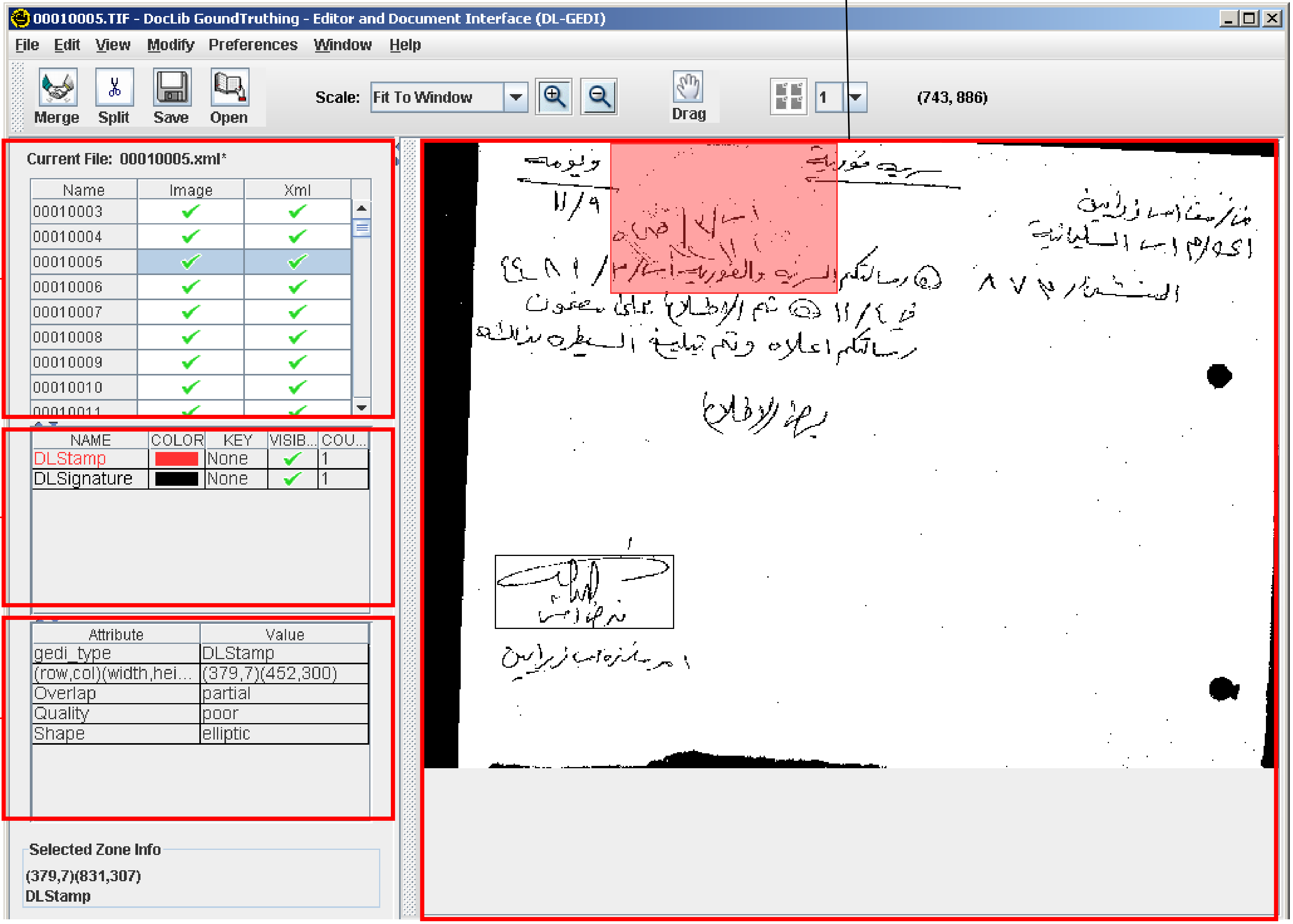
GEDI – Java Interface

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

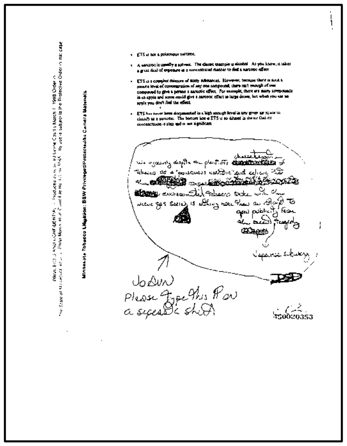
Selected Zone Info

(379,7)(831,307)
DLStamp

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

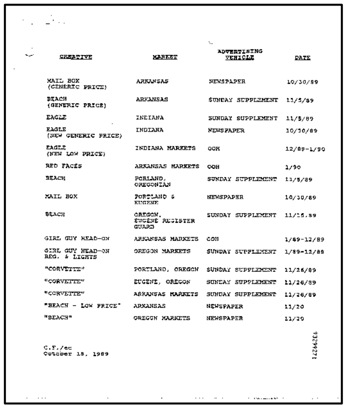


Minnesota Tobacco Litigation: BSW Privileged/Protected/In Camera Materials

- ETS is not a poisonous narcotic.
- A narcotic is usually a solvent. The classic example is alcohol. As you know, it takes a great deal of exposure in a concentrated manner to feel a narcotic effect.
- ETS is a complex mixture of many substances. However, because there is such a minute level of concentration of any one compound, there isn't enough of one compound to give a person a narcotic effect. For example, there are many compounds in an apple and some could give a narcotic effect in large doses, but when you eat an apple you don't feel the effect.
- ETS has never been documented in a high enough level in any given air space to classify as a narcotic. The bottom line is ETS is so diluted in the air that its concentration is slim and is not significant.

JoAnn
Please type this on a separate sheet

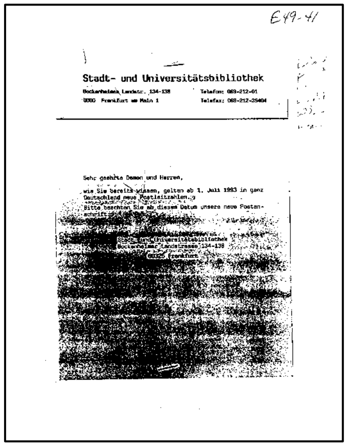
450020353



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	1/89-12/89
"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	EUGENE, OREGON	SUNDAY SUPPLEMENT	11/26/89
"CORVETTE"	ARKANSAS MARKETS	SUNDAY SUPPLEMENT	11/26/89
"BEACH - LOW PRICE"	ARKANSAS	NEWSPAPER	11/20
"BEACH"	OREGON MARKETS	NEWSPAPER	11/20

C.F./ec
October 18, 1989

93296271



E49-41

Stadt- und Universitätsbibliothek
Bockenheimer Landstr. 134-138 Telefon: 069-212-01
6000 Frankfurt am Main 1 Telefax: 069-212-39404

Sehr geehrte Damen und Herren,

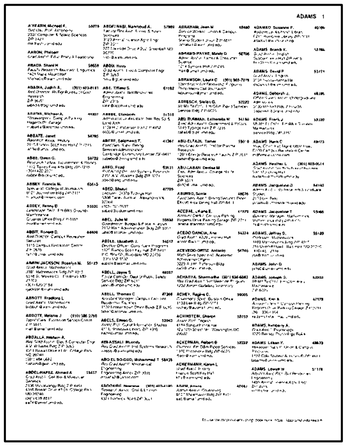
wie Sie bereits wissen, gelten ab 1. Juli 1993 in ganz Deutschland neue Postleitzahlen.

Bitte beachten Sie ab diesem Datum unsere neue Postanschrift

Stadt- und Universitätsbibliothek
Bockenheimer Landstrasse 134-138
60325 Frankfurt

A a

aardvark



ADAMS 1

The Washington Post

After Zarqawi, No Clear Path In Weary Iraq

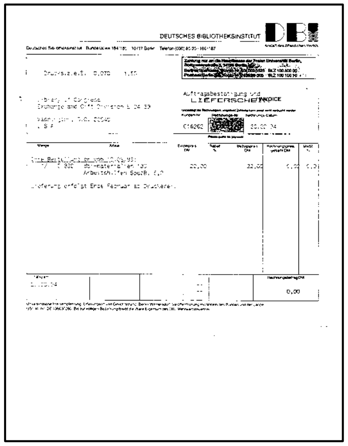
Difficult Questions Surround Legacy of Insurgent Leader

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

MARCH 26, 2007 | WWW.FORBES.COM

Forbes

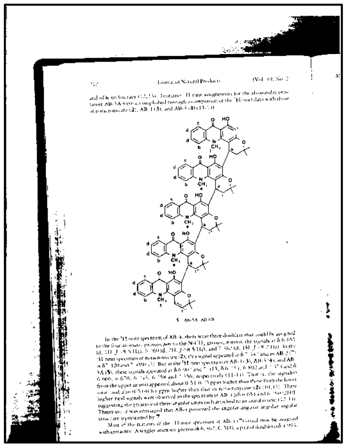
BILLIONAIRES

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

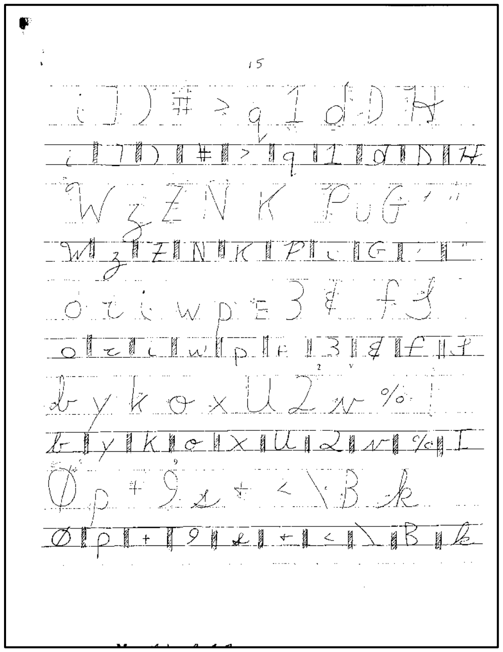
Collateral damage

Iraq braced as Saddam rejects Bush ultimatum

Fed holds rates but signals uncertain future

US probes Ahold collusion claims

Language And Media Processing
LAMP



15

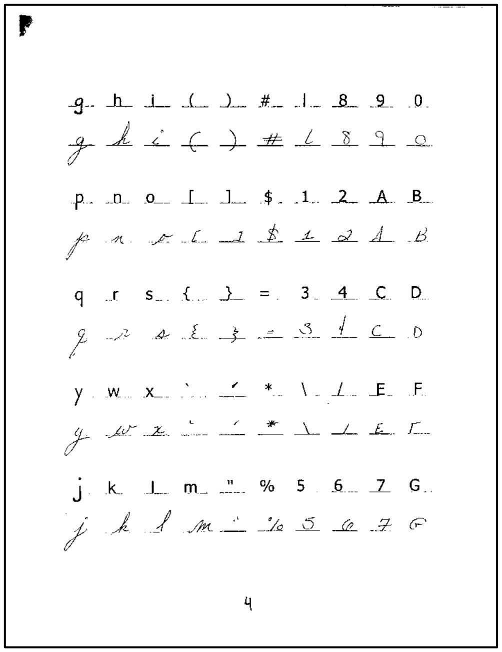
i]) # > q 1 d D H

W z Z N K P u G ' "

o z i w p E 3 f 3

b y k o x U 2 w % !

Ø p + 9 z + < \ B k



g h i ( ) # 1 8 9 0

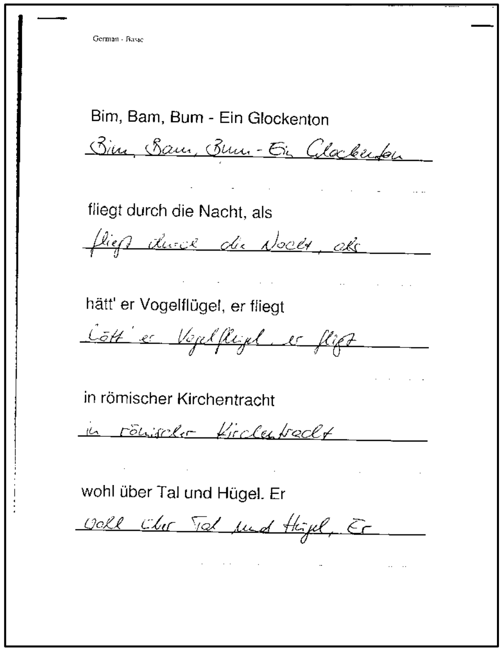
p n o [ ] $ 1 2 A B

q r s { } = 3 4 C D

y w x ` ´ * \ / E F

j k l m " % 5 6 7 G

4



German - Basic

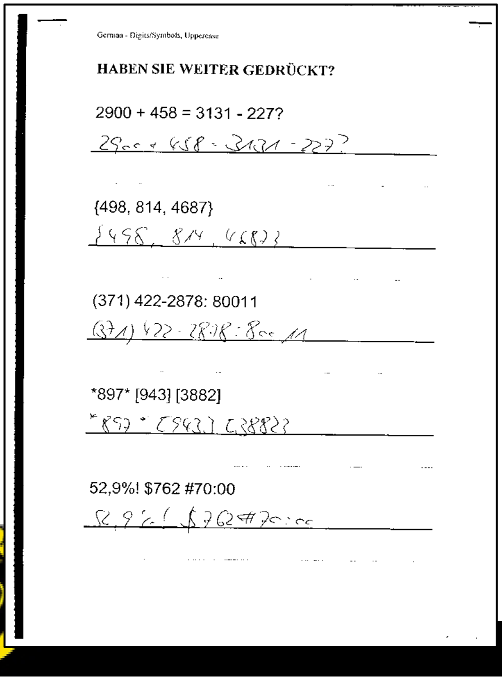
Bim, Bam, Bum - Ein Glockenton

fliegt durch die Nacht, als

hätt' er Vogelflügel, er fliegt

in römischer Kirchentracht

wohl über Tal und Hügel. Er



German - Digits/Symbols, Uppercase

HABEN SIE WEITER GEDRÜCKT?

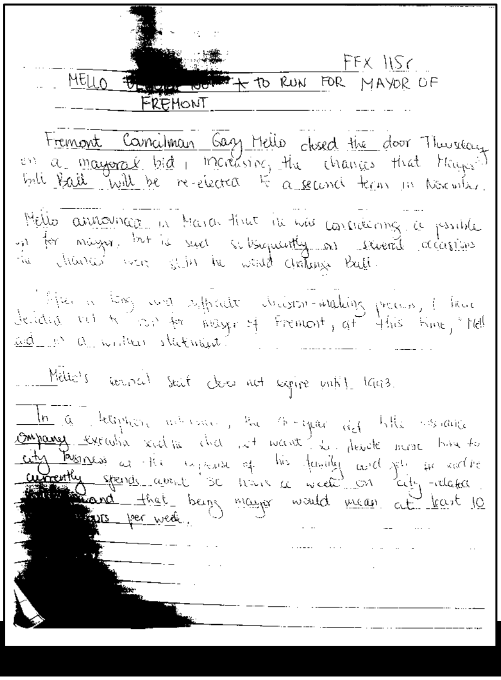
2900 + 458 = 3131 - 227?

{498, 814, 4687}

(371) 422-2878: 80011

897 [943] [3882]

52,9%! $762 #70:00



FFX 115r

MELLO [illegible] TO RUN FOR MAYOR OF FREMONT

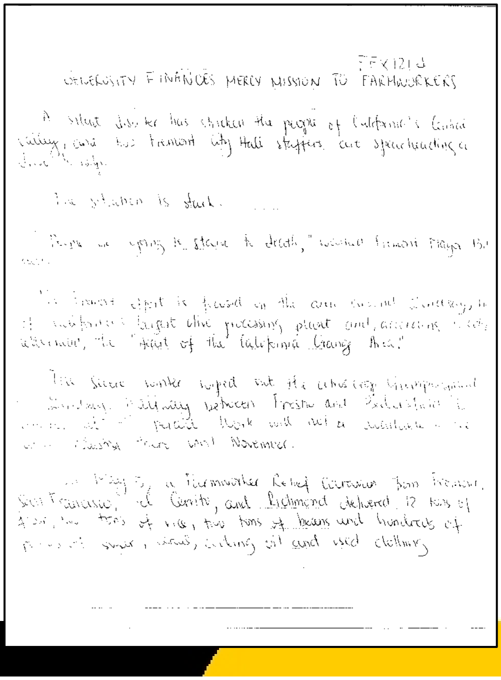
Fremont Councilman Gary Mello closed the door Thursday on a mayoral bid, increasing the chances that Mayor Bill Ball will be re-elected to a second term in November.

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," Mello said in a written statement.

Mello's council seat does not expire until 1993.

In a telephone interview, the [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 hours per week.



FFX121J

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 [illegible], home 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] between Fresno and Bakersfield [illegible] work will not be available [illegible] 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 [illegible] 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

25,000 pages ground truthed

	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

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

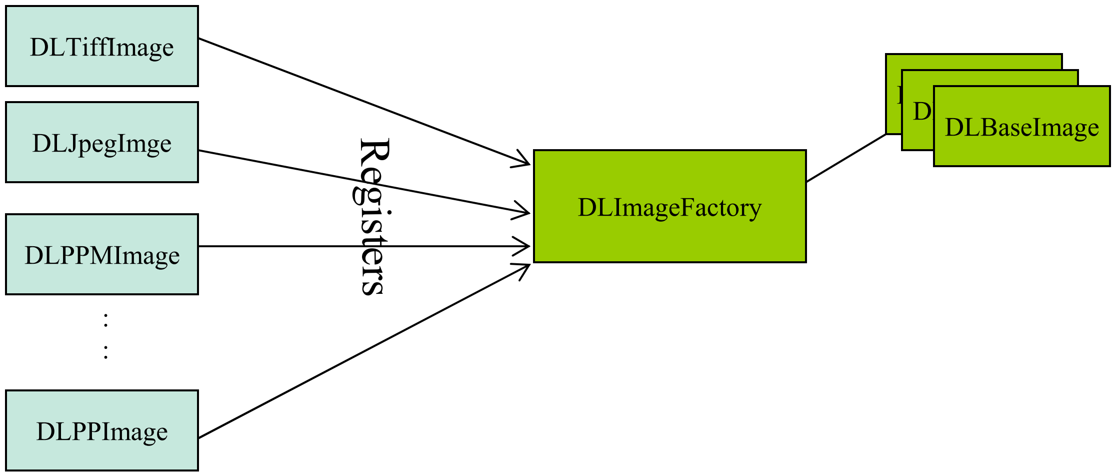
- Core DocLib components matured and stable (in use by a variety of government installations)\
- Addons being integrated/implemented, primarily by developers
- Freely available to government researchers
- Core supported on Solaris, Linux and Windows

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



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

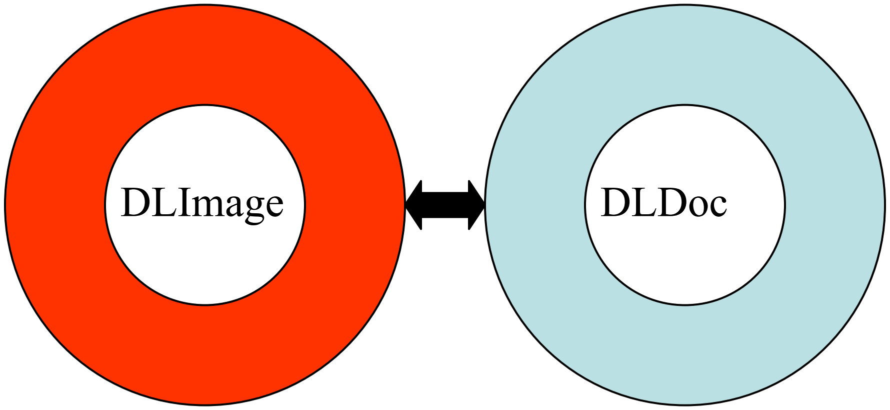
DocLib´s architecture rests on two pillars:

DLImage:

- ➢ **Image Processing**

DLDocument:

- ➢ **Document Processing**



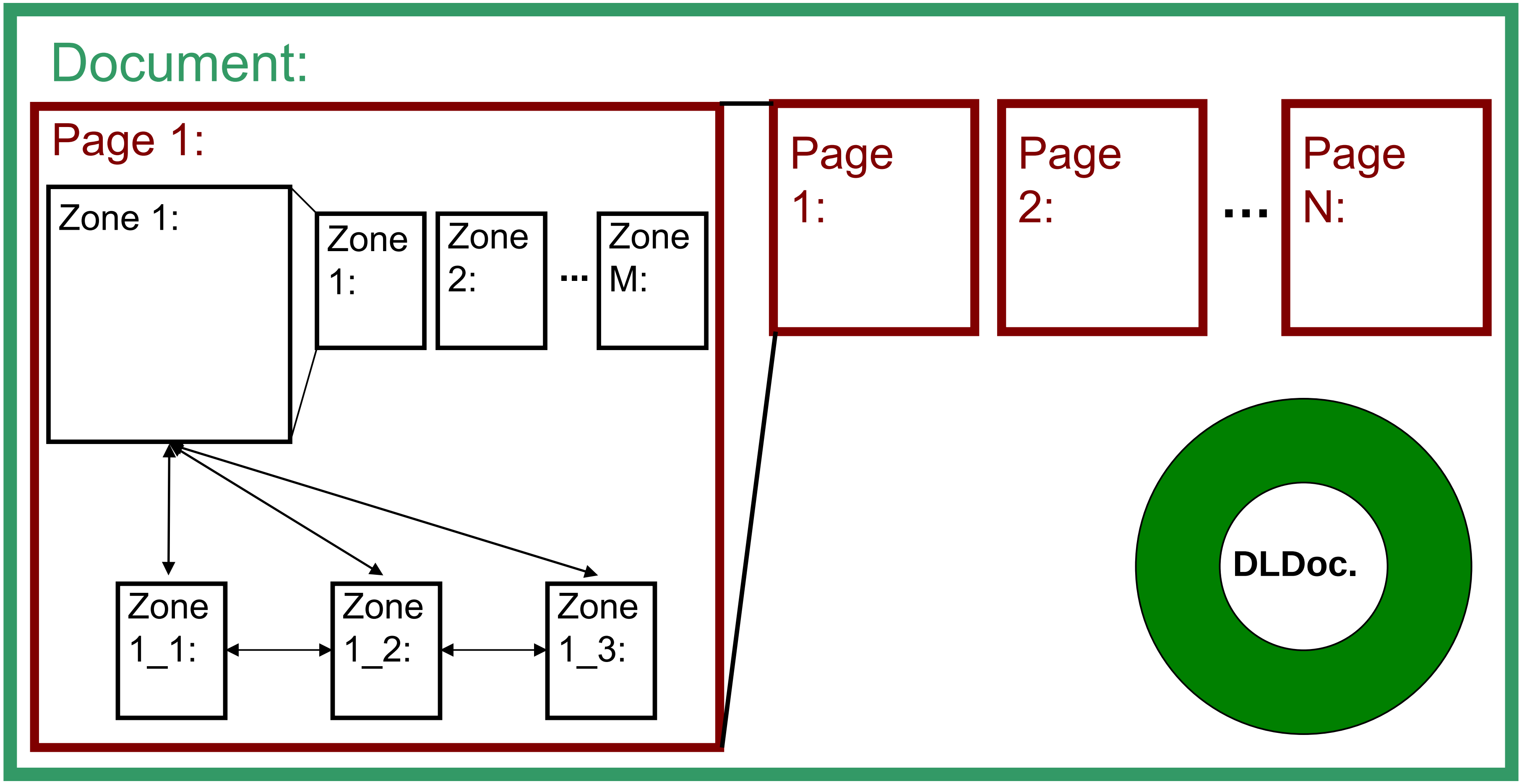
e.g.

- ➢ **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

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>
```

Technical Presentations

- Page Segmentation (and rule line separation)

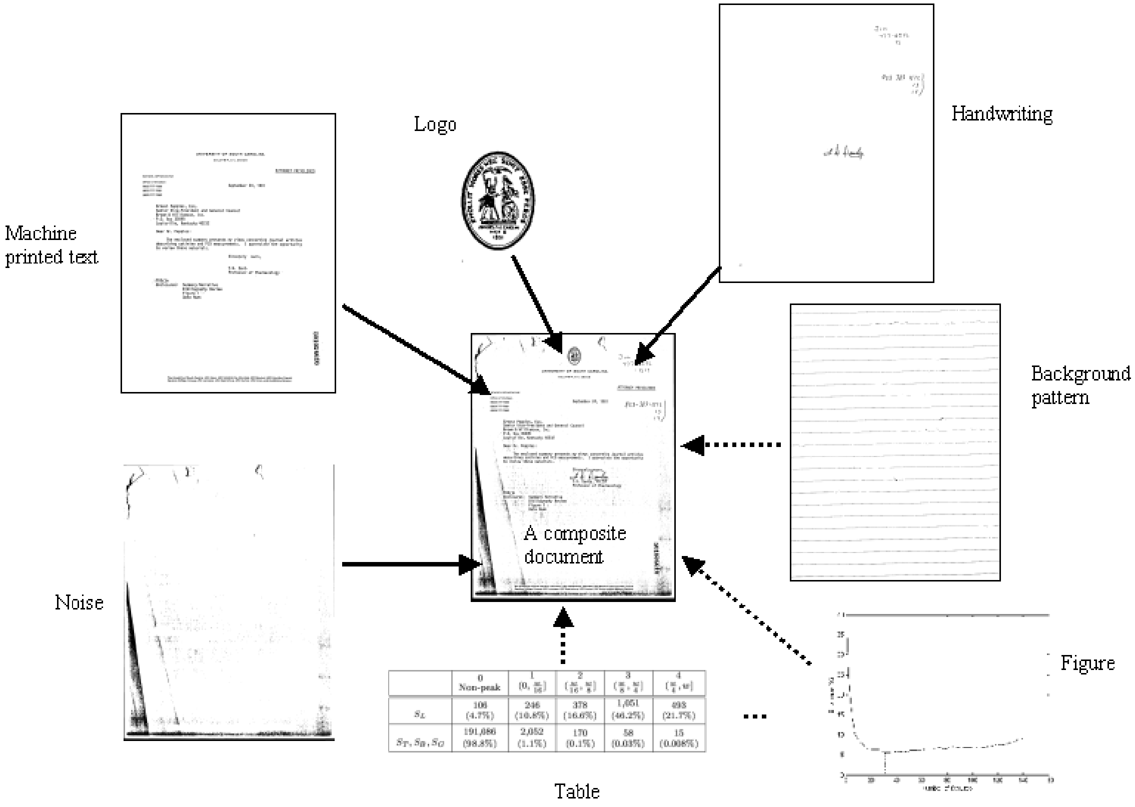
- Logo Detection and Recognition
- Signature Detection
- Stamp Detection

- Document ID/Script ID
- Page Layout Similarity

- Video Research
  - Tracking and Analysis of People
  - Video Content Classification

Page Layer Segmentation

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

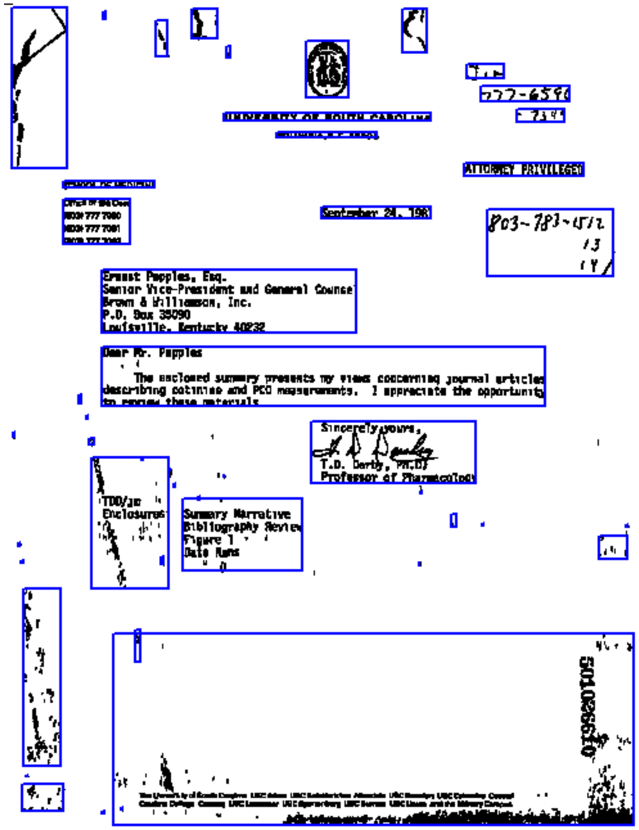


	0 Non-peak	1 $(0, \frac{w}{16}]$	2 $(\frac{w}{16}, \frac{w}{8}]$	3 $(\frac{w}{8}, \frac{w}{4}]$	4 $(\frac{w}{4}, w]$
S_L	106 (4.7%)	246 (10.8%)	378 (16.6%)	1,051 (46.2%)	493 (21.7%)
S_T, S_B, S_G	191,086 (98.8%)	2,052 (1.1%)	170 (0.1%)	58 (0.03%)	15 (0.008%)

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.

Page Segmentation for Noisy Documents



UNIVERSITY OF SOUTH CAROLINA

777-6590

7341

ATTORNEY PRIVILEGED

(803) 777 7080
(803) 777 7081

September 24, 198

803-783-4512
13
14

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

Dear Mr. Pepples

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

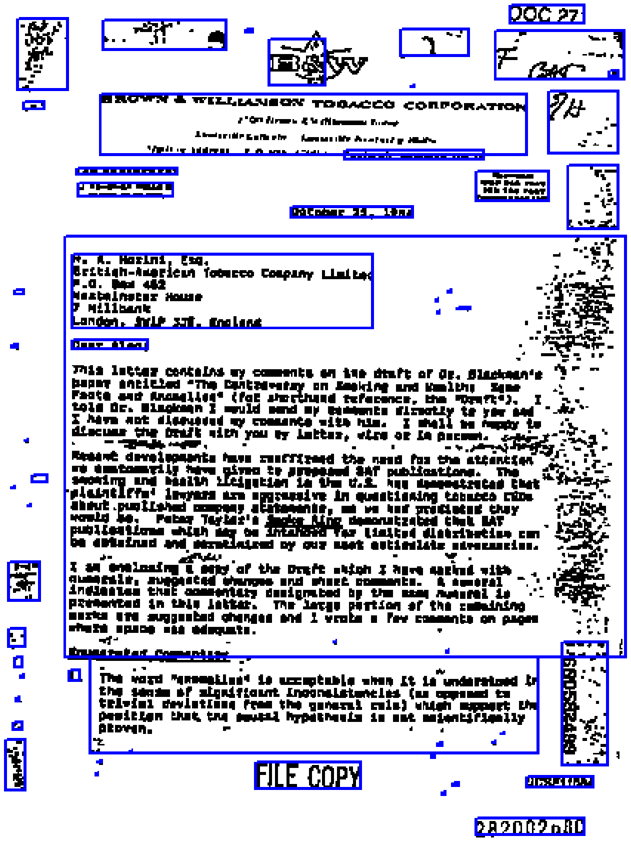
Sincerely yours,

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

TDD/jm
Enclosures

Summary Narrative
Bibliography Review
Figure 1
Data Runs

501096810



DOC 27

BROWN & WILLIAMSON TOBACCO CORPORATION

October [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 cos. 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 Commentary

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

282002680

* 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

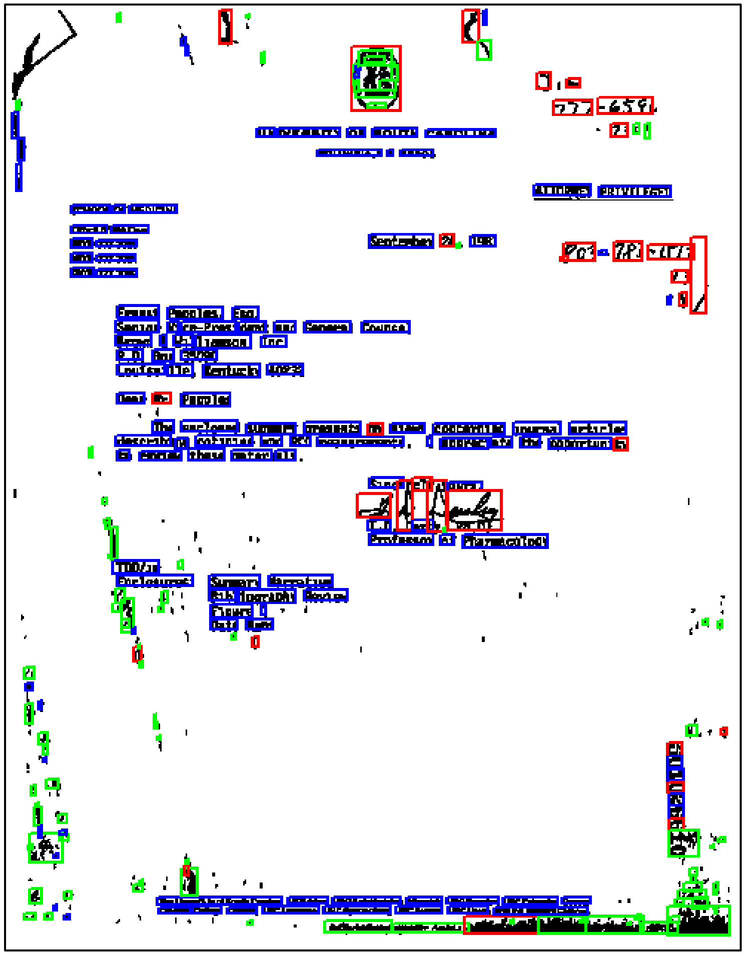
	Usage description	Dimensio	Selected
Structural	Region size, connected components	18	9
Gabor filter	Stroke orientation	16	4
Run-length histogram	Stroke length	20	5
Crossing counts histogram	Stroke complexity	10	6
Co-occurrence	Texture	16	2
2×2 gram	Texture	60	5
Total		140	31

Classification Results with Fisher Classifier

Printed text

Handwriting

Noise



ATTORNEY PRIVILEGED

September 29, 1981

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

Dear Mr. Pepples

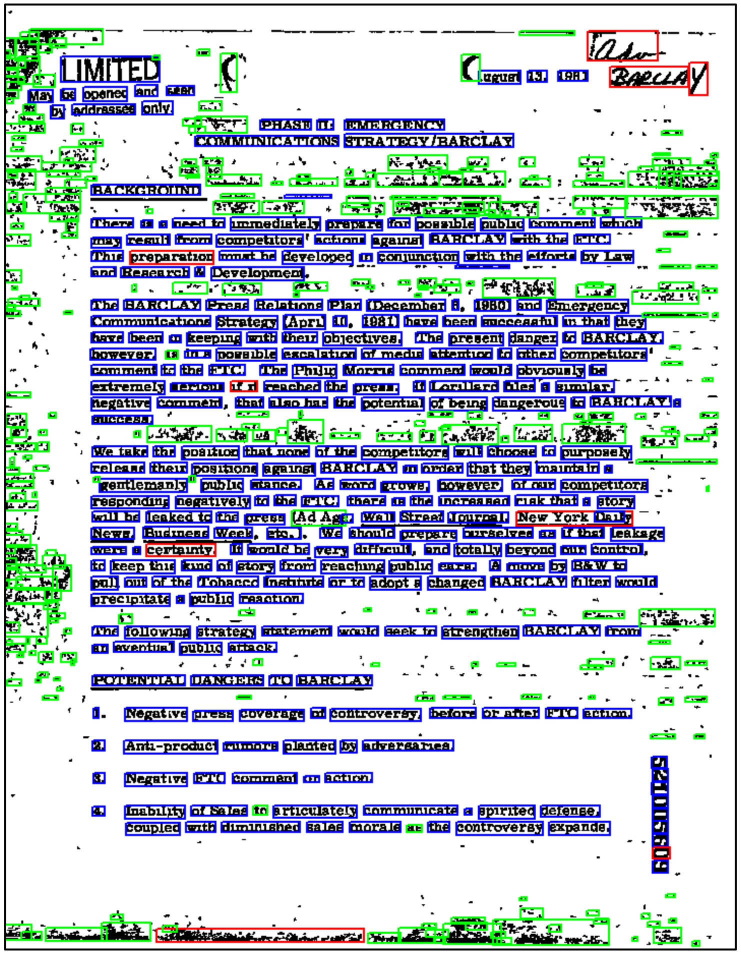
The enclosed summary presents [illegible] items concerning journal articles describing [illegible] and [illegible] management. [illegible] the opportunity to [illegible] these materials.

Sincerely yours,

Professor of Pharmacology

TDD/jr

Enclosures: Summary Literature
Bibliography Review
Figure 1
Data Base



LIMITED

August 13, 1981

Ada
BARCLAY

May be opened and seen
by addressees only.

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

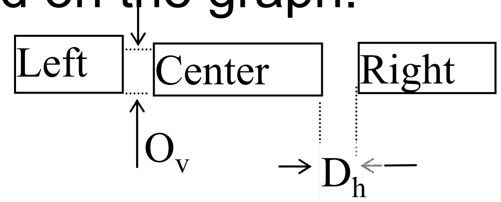
1. Negative press coverage of controversy, before or after FTC action.
2. Anti-product rumors planted by adversaries.
3. Negative FTC comment or action.
4. Inability of Sales to articulately communicate a spirited defense, coupled with diminished sales morale as the controversy expands.

Using Context

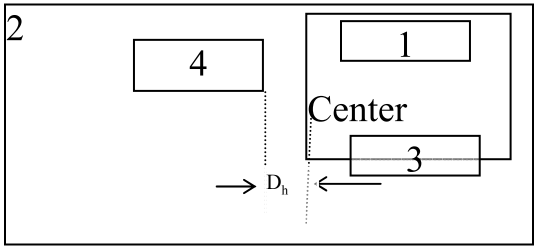
- The results are reasonable with a few mis-classification due to the overlapping of different classes in the feature space.
- Context can be used to refine classification results further
  - Words of printed text tend to lie on the same line.
  - Noise block are likely to overlap each other.
- This 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

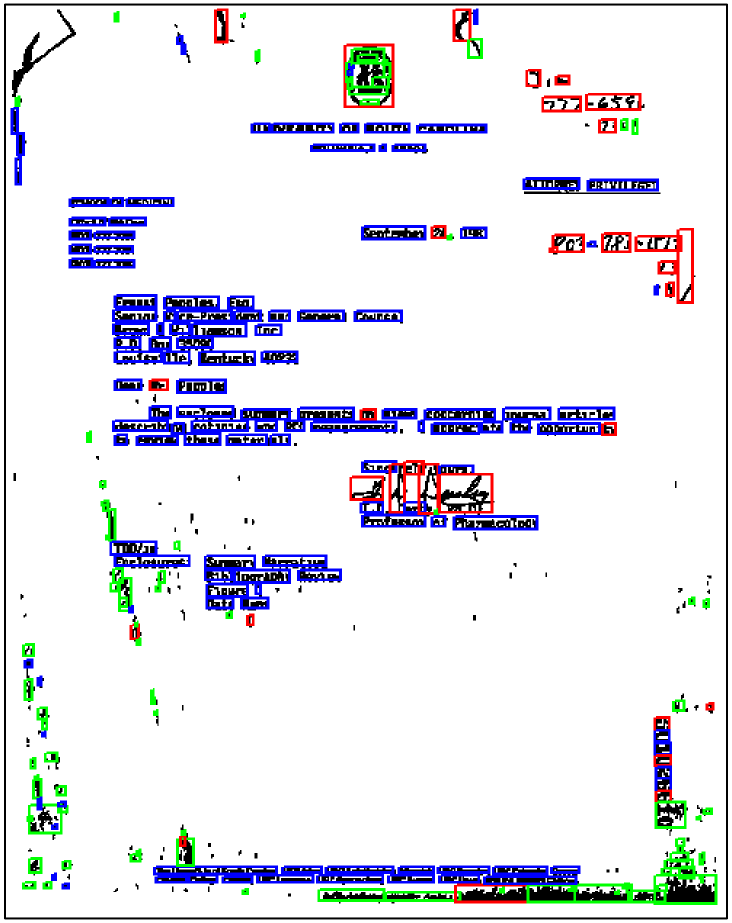


MRF Postprocessing Example

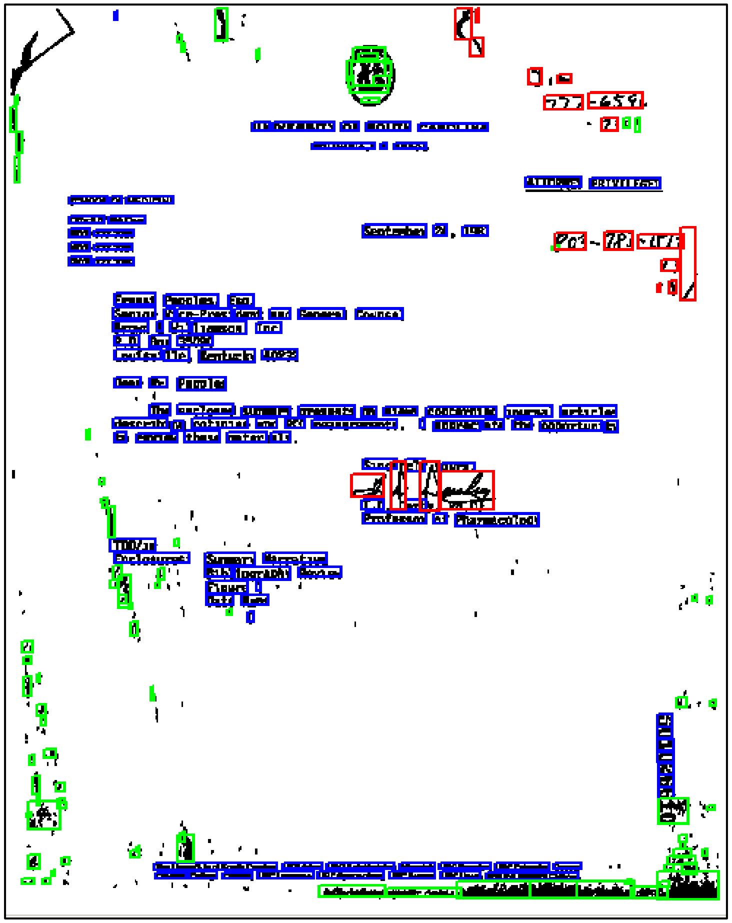
Printed text

Handwriting

Noise



Before MRF-based postprocessing



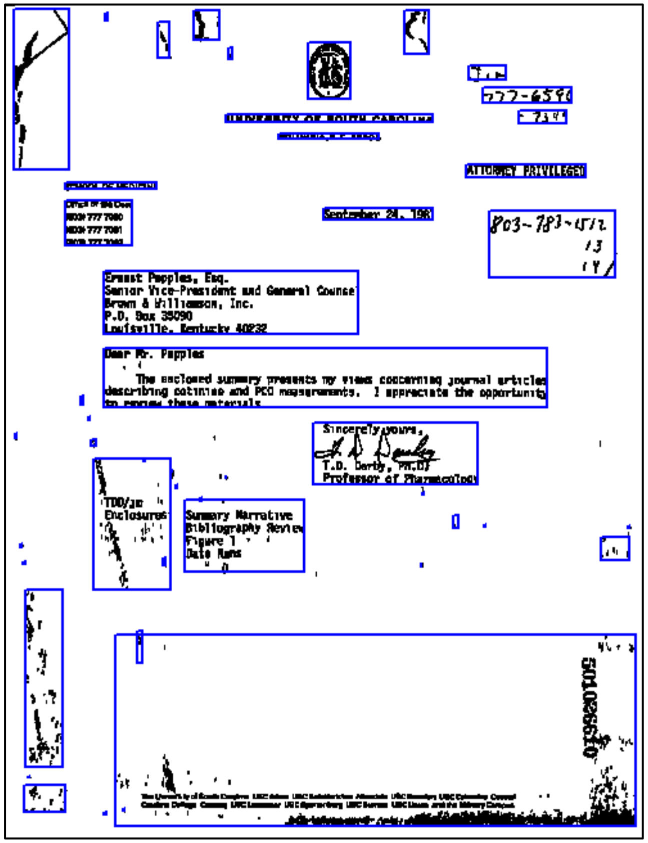
After MRF-based postprocessing

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

ATTORNEY PRIVILEGED

September 24, 198

803-783-1512

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

Dear Mr. Pepples

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

Sincerely yours,

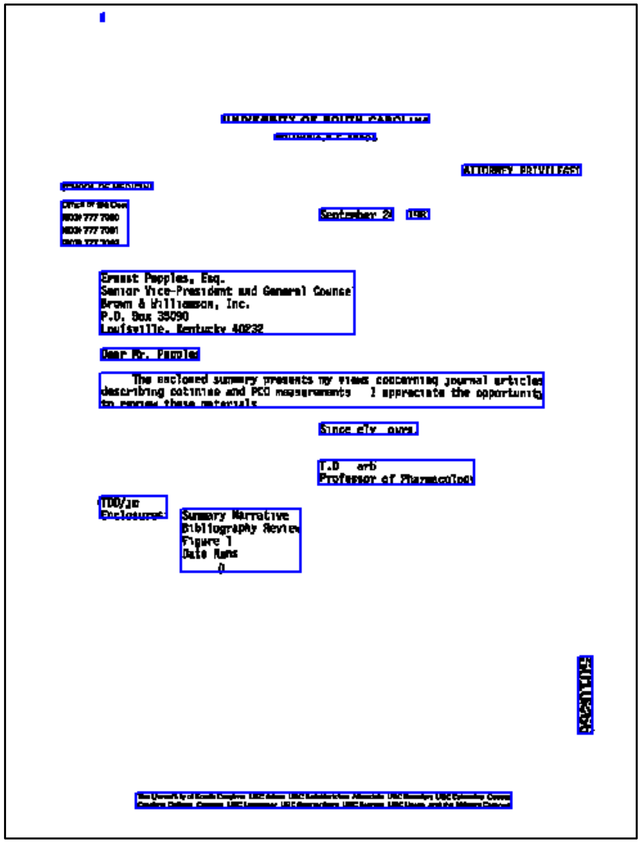
T.D. Darby, Ph.D.
Professor of Pharmacology

TDD/jm
Enclosures:

Summary Narrative
Bibliography Review
Figure 1
Date Runs

501029610

Before enhancement



ATTORNEY PRIVILEGED

September 24 198

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

Dear Mr. Pepples

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

Since ely ours

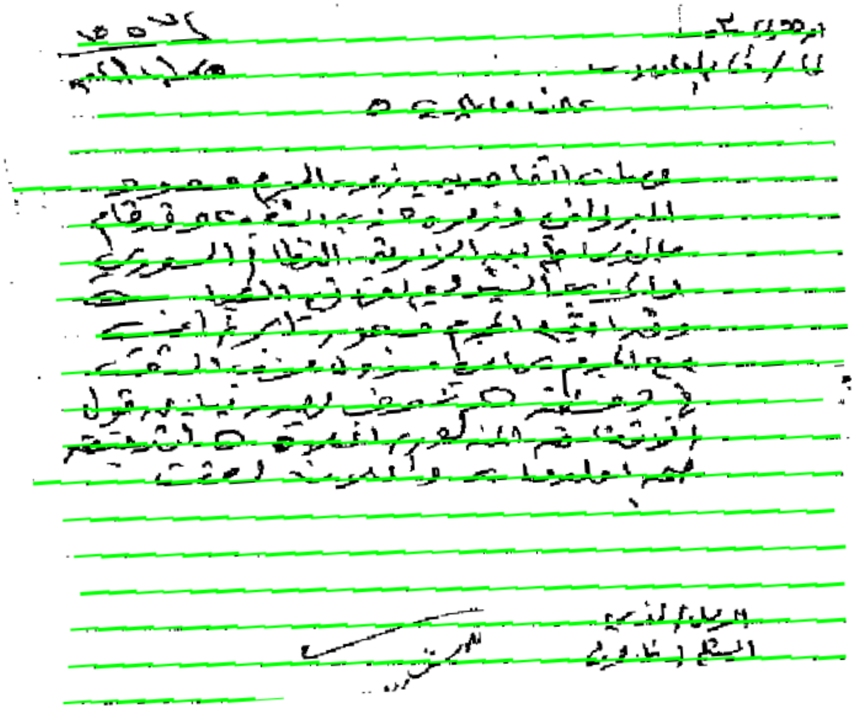
T.D arb
Professor of Pharmacology

TDD/jm
Enclosures:

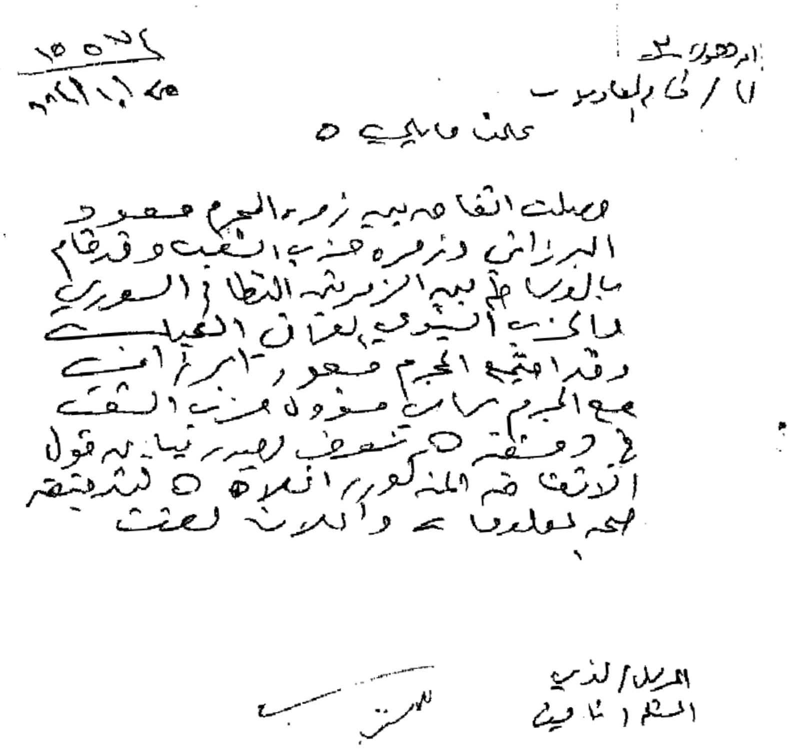
Summary Narrative
Bibliography Review
Figure 1
Date Runs

After enhancement

Rule Line Detection Example



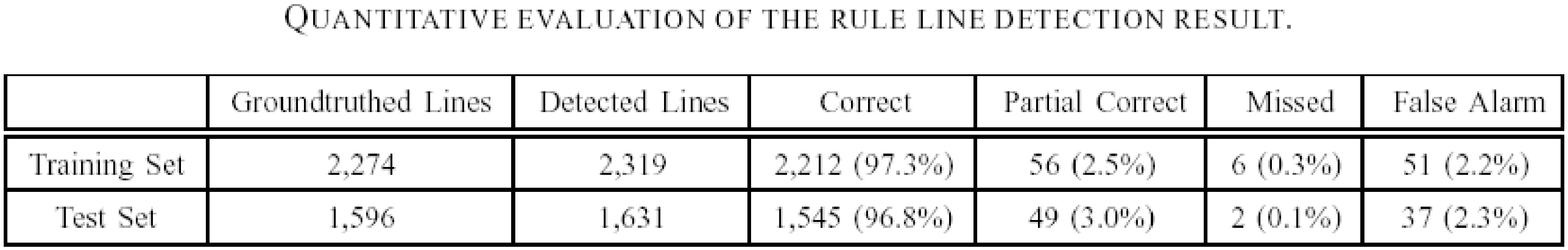
Model-based line detection result



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%)

Technical Presentations

- Page Segmentation (and rule line separation)

- Signature Detection
- Logo Detection and Recognition
- Stamp Detection
- Document ID/Script ID

Metadata Extraction

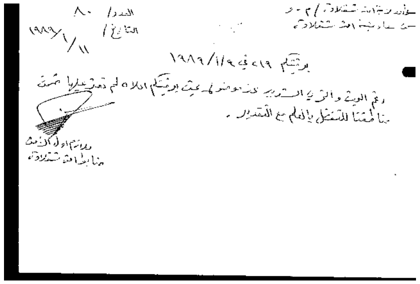
- Page Layout Similarity

- Video Research
  - Tracking and Analysis of People
  - Video Content Classification

Problem Statement

Given a large heterogeneous document image database, we are facing a few very challenging problems

- How can we retrieve documents authored or approved by a specific individual in unconstrained settings?



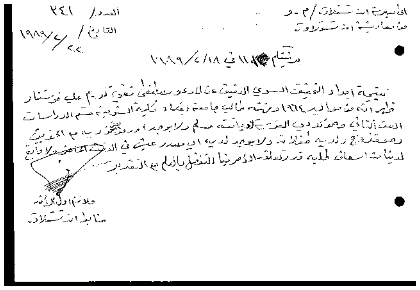
[illegible] امن شقلاوة / م - د
[illegible] امن شقلاوة

العدد / ٨٠
التاريخ / ١٩٨٩/١/١١

برقيتكم ١٩ع في ١٩٨٩/١/٩

رغم البحث والتحري المستمر عن موضوع بحث برقيتكم اعلاه لم [illegible]
[illegible] للتفضل بالعلم مع التقدير .

[illegible]
مدير امن شقلاوة



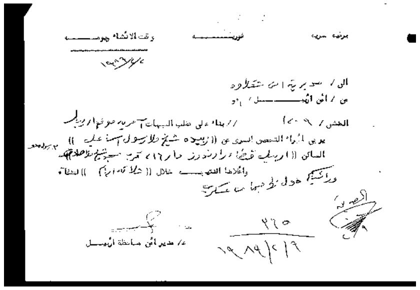
مديرية امن شقلاوة / م - د
[illegible] امن شقلاوة

العدد / ٢٤١
التاريخ / ١٩٨٩/٢/٢٢

برقيتكم ١١ في ١٩٨٩/٢/١٨

[illegible]
[illegible] التفضل بالعلم مع التقدير

[illegible]
مدير امن شقلاوة



مرتبة سرية   فوري   وقت الانشاء يوم
١٩٨٩/٢/٧

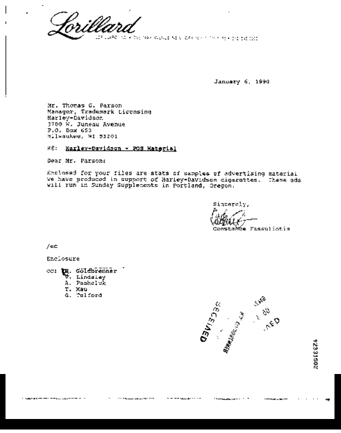
الى / مديرية امن شقلاوة
من / امن اربيـــل / [illegible]

العدد / ١٢٠٩   // بناء على طلب الجهات [illegible] اربيل
يرجى اجراء التحقيق السري عن (([illegible]))
الساكن ((اربيل [illegible])) [illegible]
واعلامنا النتيجة خلال (([illegible])) لطفاً
[illegible]

م/ مدير امن محافظة اربيل

٢٦٥
١٩٨٩/٢/٩

- How can we retrieve documents originating from an organization?



Lorillard

January 6, 1990

Mr. Thomas G. Parson
Manager, Trademark Licensing
Harley-Davidson
3700 W. Juneau Avenue
P.O. Box 653
Milwaukee, WI 53201

RE: **Harley-Davidson - POS Material**

Dear Mr. Parson:

Enclosed for your files are stats of samples of advertising material we have produced in support of Harley-Davidson cigarettes. These ads will run in Sunday Supplements in Portland, Oregon.

Sincerely,

Constance Fasuliotis

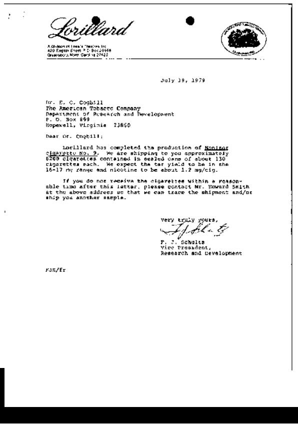
/ec

Enclosure

cc: R. Goldbrenner
V. Lindsley
A. Pashcluk
T. Mau
G. Telford

RECEIVED

92331502



Lorillard
A Division of Loew's Theatres Inc
420 English Street P.O. Box 20958
Greensboro, North Carolina 27420

July 19, 1979

Dr. E. C. Cogbill
The American Tobacco Company
Department of Research and Development
P. O. Box 899
Hopewell, Virginia 23860

Dear Dr. Cogbill:

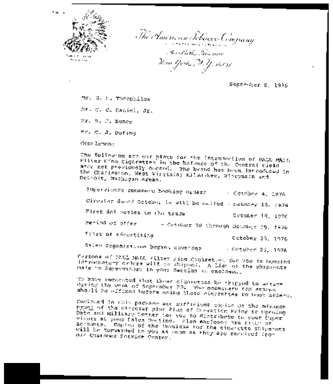
Lorillard has completed the production of Monitor cigarette No. 9. We are shipping to you approximately 6200 cigarettes contained in sealed cans of about 130 cigarettes each. We expect the tar yield to be in the 16-17 mg range and nicotine to be about 1.2 mg/cig.

If you do not receive the cigarettes within a reasonable time after this letter, please contact Mr. Howard Smith at the above address so that we can trace the shipment and/or ship you another sample.

Very truly yours,

F. J. Schultz
Vice President,
Research and Development

FJS/fr



The American Tobacco Company
245 Park Avenue
New York, N.Y. 10017

September 8, 1976

Mr. S. L. Theophilos

Mr. C. C. Daniel, Jr.

Mr. R. J. Bence

Mr. C. J. Dofing

Gentlemen:

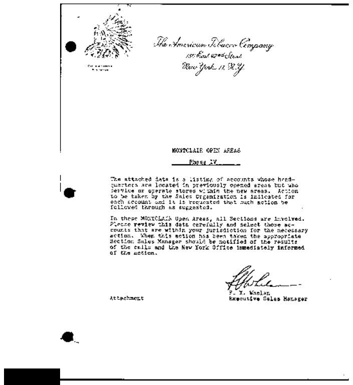
The following are our plans for the introduction of PALL MALL Filter King Cigarettes in the balance of the Central Field area not previously opened. The brand has been introduced in the Charleston, West Virginia; Milwaukee, Wisconsin and Detroit, Michigan areas.

Supervisors commence booking orders - October 4, 1976
Circular dated October 1a will be mailed - October 15, 1976
First deliveries to the trade October 18, 1976
Period of offer - October 18 through October 29, 1976
Start of advertising October 25, 1976
Sales Organization begins coverage - October 25, 1976

Cartons of PALL MALL Filter King Cigarettes for use in booking introductory orders will be shipped. A list of the shipments made to Supervisors in your Section is enclosed.

We have requested that these cigarettes be shipped to arrive during the week of September 20. The necessary tax stamps should be affixed before using these cigarettes to book orders.

Enclosed in this package are sufficient copies of the advance proof of the circular plus Plan of Operation Prior to Opening Date and Military Letter for you to distribute to your Supervisors at your Sales Meeting. Also enclosed are lists of accounts. Copies of the invoices for the cigarette shipments will be forwarded to you as soon as they are received from our Customer Service Center.



The American Tobacco Company
150 East 42nd Street
New York 17, N.Y.

MONTCLAIR OPEN AREAS

Phase IV

The attached data is a listing of accounts whose headquarters are located in previously opened areas but who service or operate stores within the new areas. Action to be taken by the Sales Organization is indicated for each account and it is requested that such action be followed through as suggested.

In these MONTCLAIR Open Areas, all Sections are involved. Please review this data carefully and select those accounts that are within your jurisdiction for the necessary action. When this action has been taken the appropriate Section Sales Manager should be notified of the results of the calls and the New York Office immediately informed of the action.

Attachment

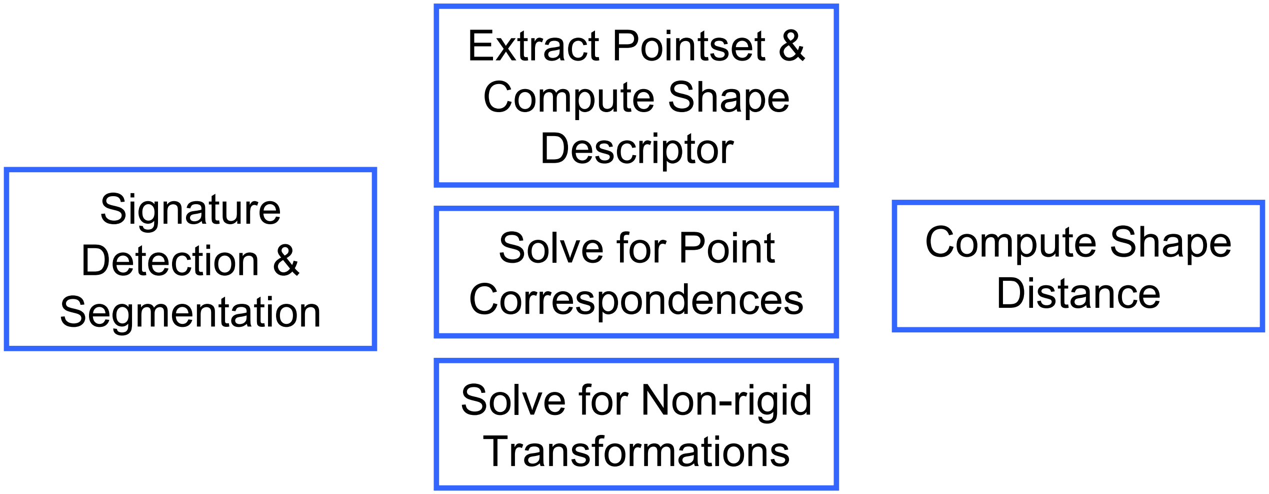
F. X. Whelan
Executive Sales Manager

Motivation

- Signatures and logos provide exciting new dimensions for document image mining
- Solution to these problems are also important in document analysis systems in a range of application domains
  - Signature verification and identification
  - Business process automation

Our Tasks

- Two problems are of fundamental interest to general content-based image retrieval
  - Detection and segmentation
  - Matching
    - Representation
    - Similarity measures
    - Matching algorithms



Signature Detection & Segmentation

Extract Pointset & Compute Shape Descriptor

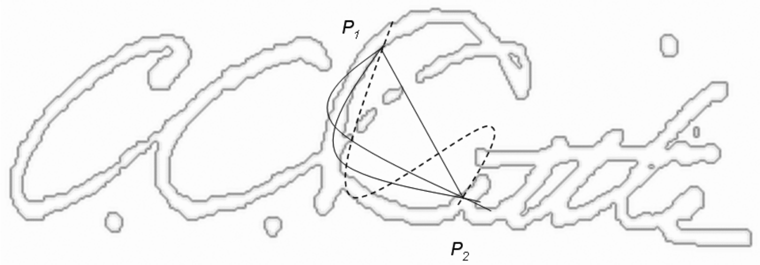
Solve for Point Correspondences

Solve for Non-rigid Transformations

Compute Shape Distance

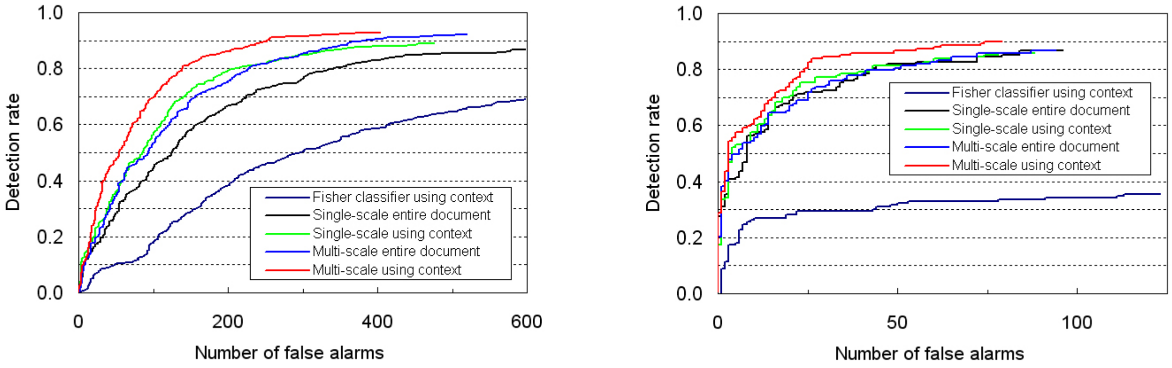
Overview of our approach

- We treat a signature as a global symbol. Rather than focusing on local features that typically have large variations, our approach aims to capture the structural saliency of a signature by searching over multiple scales
- We consider identifying salient structure and grouping its parts in two separate steps
- Two keys questions we addressed are:
  - How to effectively model off-line signature production under reasonable assumptions without its temporal information
  - What to effectively measure the structural saliency of signatures under such production model



Evaluation

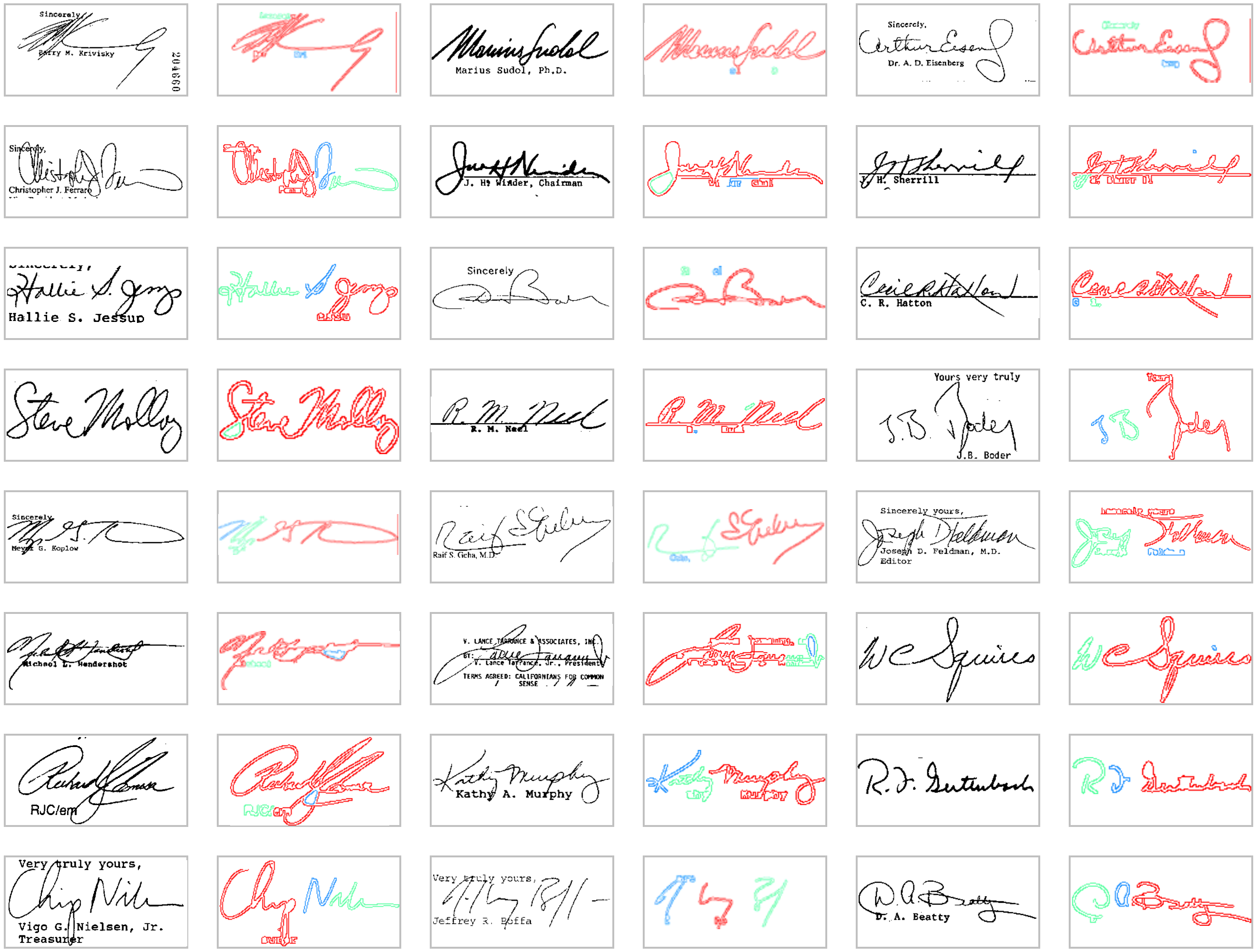
- We used two large collections of real world documents—Tobacco-800 and University of Maryland Arabic datasets.
- Using document context, our multi-scale signature detector achieves 92.8% and 86.6% detection rates for the Tobacco-800 and Maryland Arabic datasets, at 0.3 false-positives per image.



(a) **(b)**

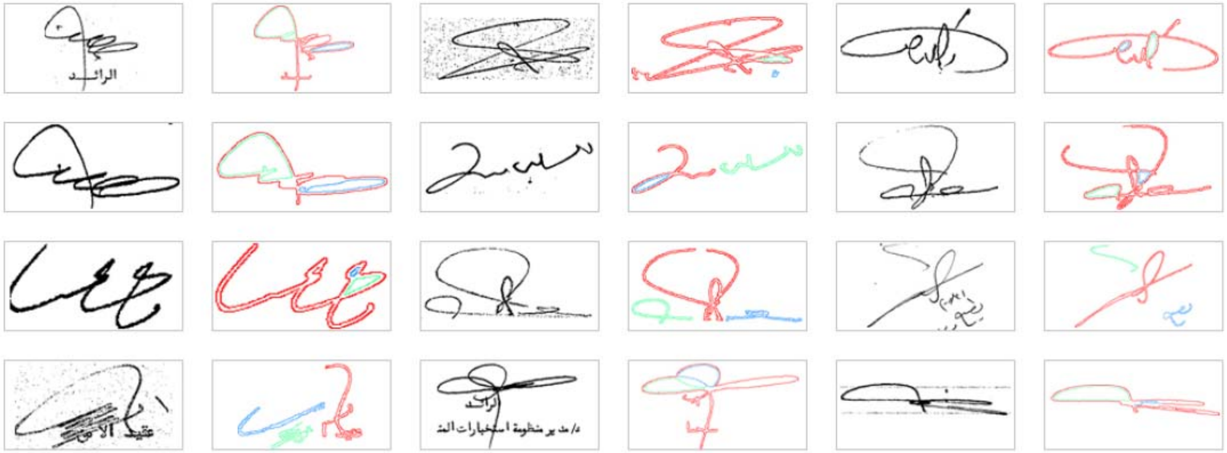
ROC curves for (a) Tobacco-800 dataset and (b) Maryland Arabic dataset.

Evaluation



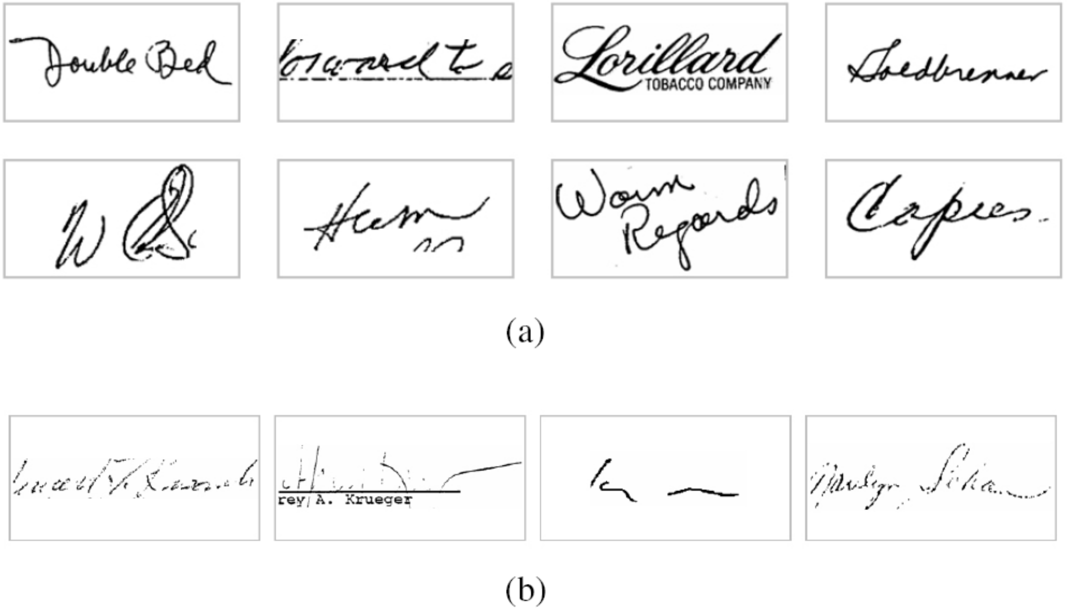
Examples of detected signatures from Tobacco-800 and their saliency maps.

Evaluation



Examples of detected signatures from Maryland Arabic dataset and their saliency maps.

Evaluation

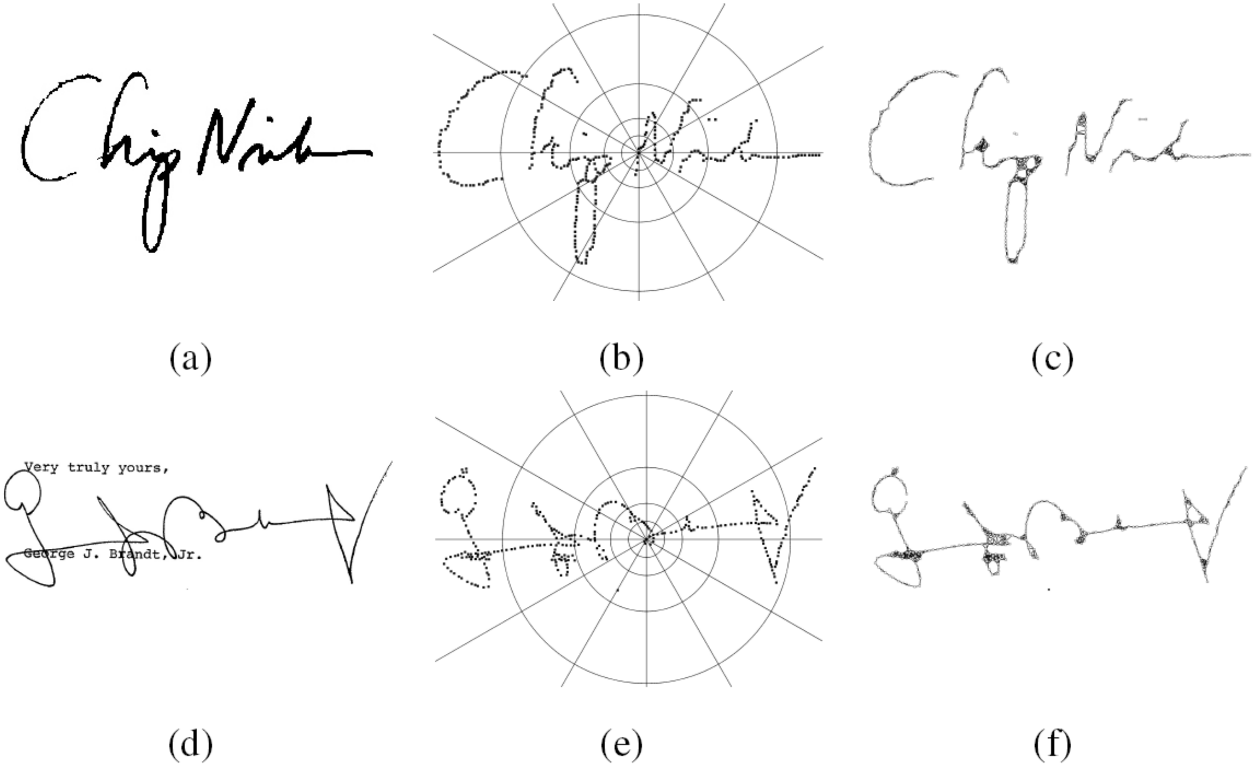


(a)

(b)

Examples of (a) falsely alarms (b) missed signatures

Shape representation

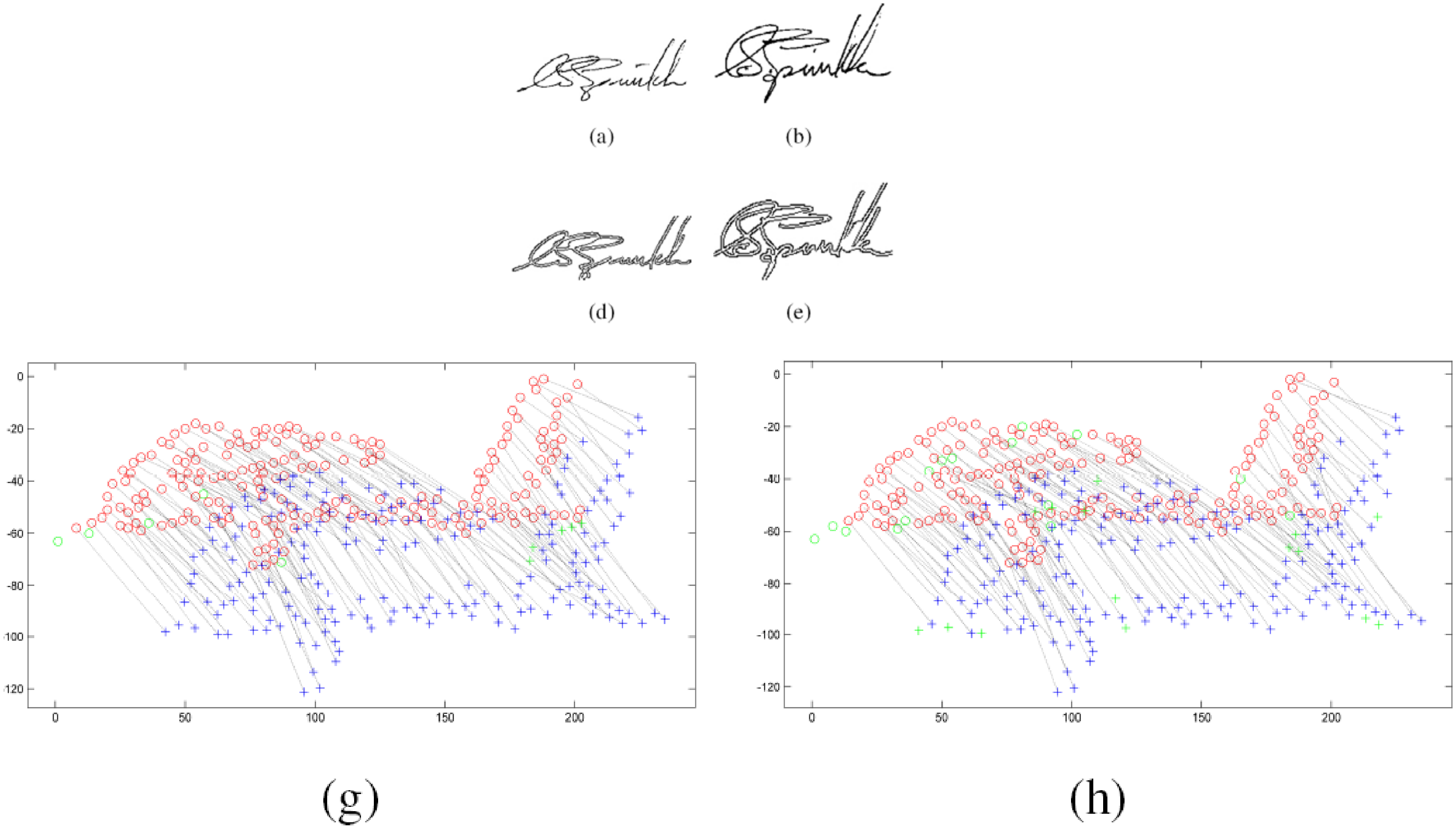


(a) (b) (c)

(d) (e) (f)

Shape contexts [Belongie *et al*., 2002] and local-neighborhood-graph [Zheng and Doermann, 2006] constructed from detected and segmented signatures.

Shape matching


(a) (b)

(d) (e)

(g) (h)

Illustration of signature matching using shape contexts and local-neighborhood-graph

Shape matching

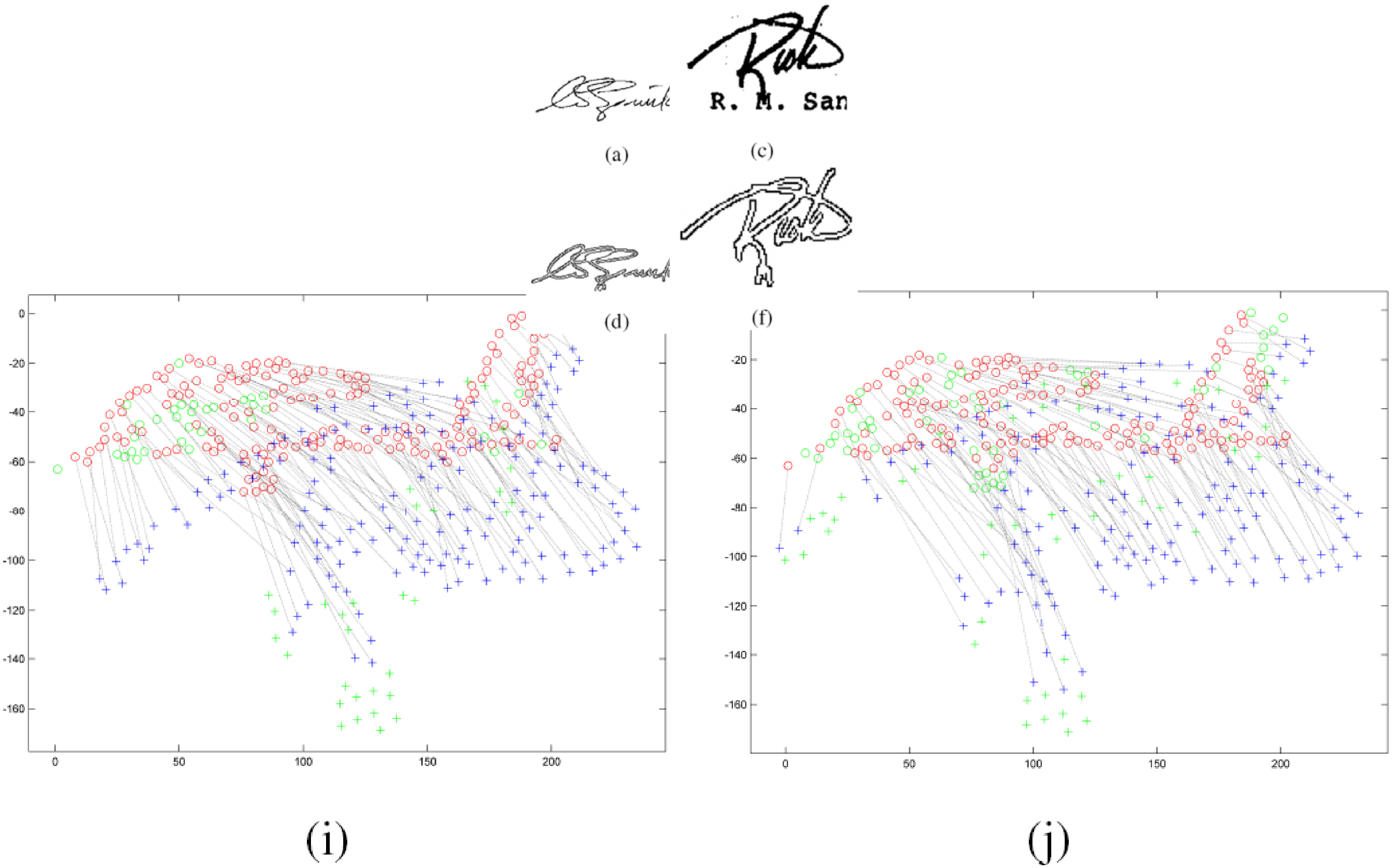
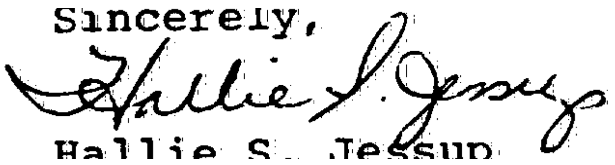


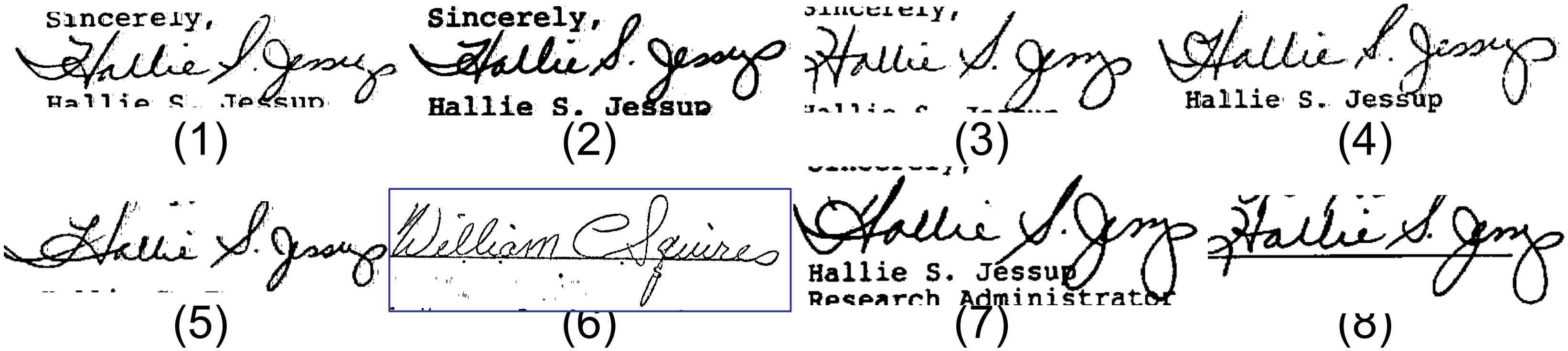
Illustration of signature matching using shape contexts and local-neighborhood-graph

Shape matching evaluation

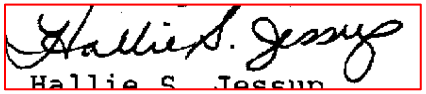
A query with eight relevant signature instances



Top eight retrieved in the ranked list



Relevant instance outside the top eight in the ranked list



(10)

A signature query example. Among the total of eight relevant signature instances, seven appear in the top eight of the 460-element ranked list, giving an average precision of 94.2%, and an R-Precision of 87.5%. The irrelevant signature that is ranked among the top eight is highlighted with a blue bounding box.

Signature matching results

Table 1: Signature retrieval result using different similarity measures.

Similarity measures	Mean average precision	Mean R-precision
D_{sc}	66.9%	62.8%
D_{af}	61.3%	57.0%
D_{be}	59.8%	55.6%
D_{re}	52.5%	48.3%
$D_{sc} + D_{be}$	78.7%	74.3%
$D_{sc} + D_{af} + D_{sc} + D_{re}$	84.5%	80.8%

Table 2: Signature retrieval result using multiple instances of signatures from the same person in each query.

Number of instances	Mean average precision	Mean R-precision
One	84.5%	80.8%
Two	88.6%	85.2%
Three	91.3%	88.1%

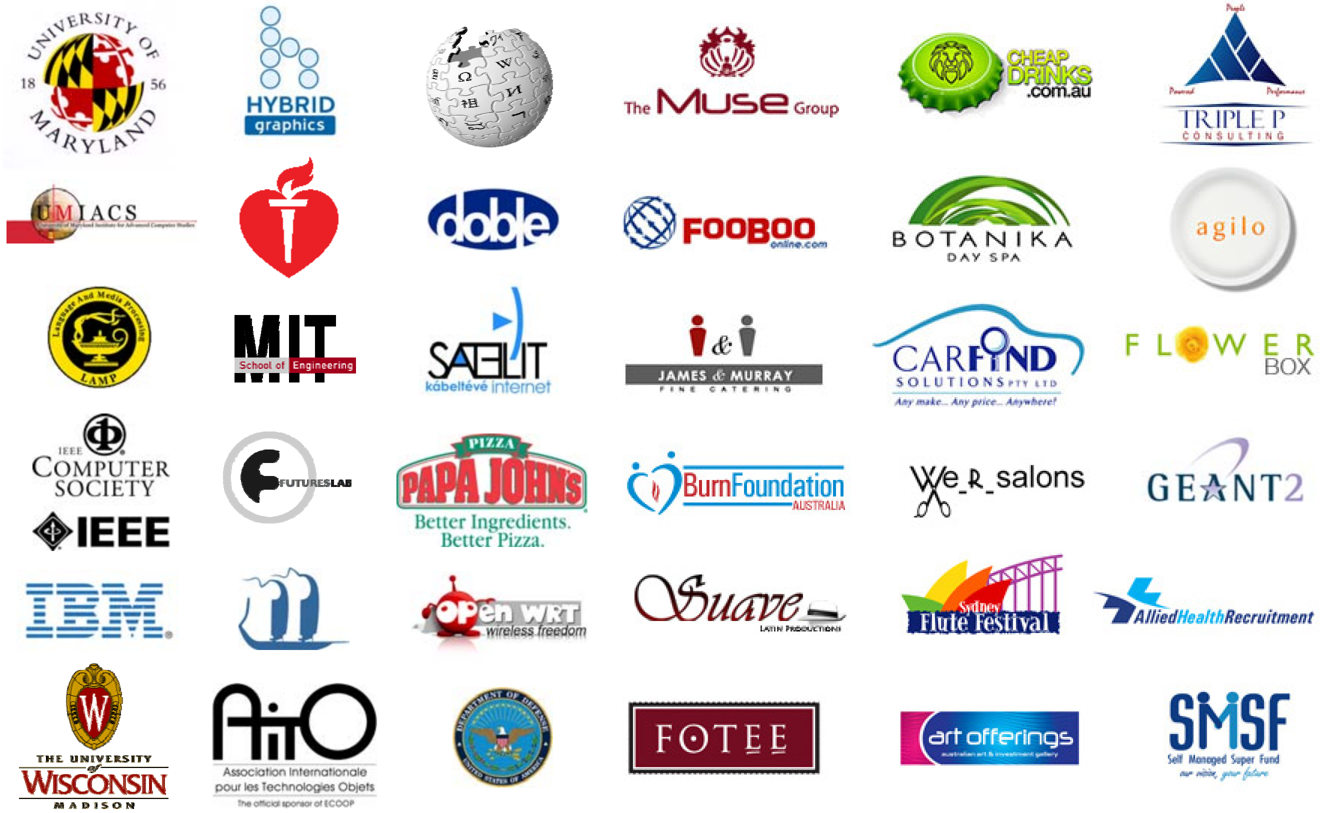
Logo Detection and Recognition

  - enables identification of the source of documents from a given organization
  - Most studies assume good logo detection and segmentation is available

- Challenges
  - Detection is required for any prior to extraction
  - Extraction is required for any shape based matching/recognition process

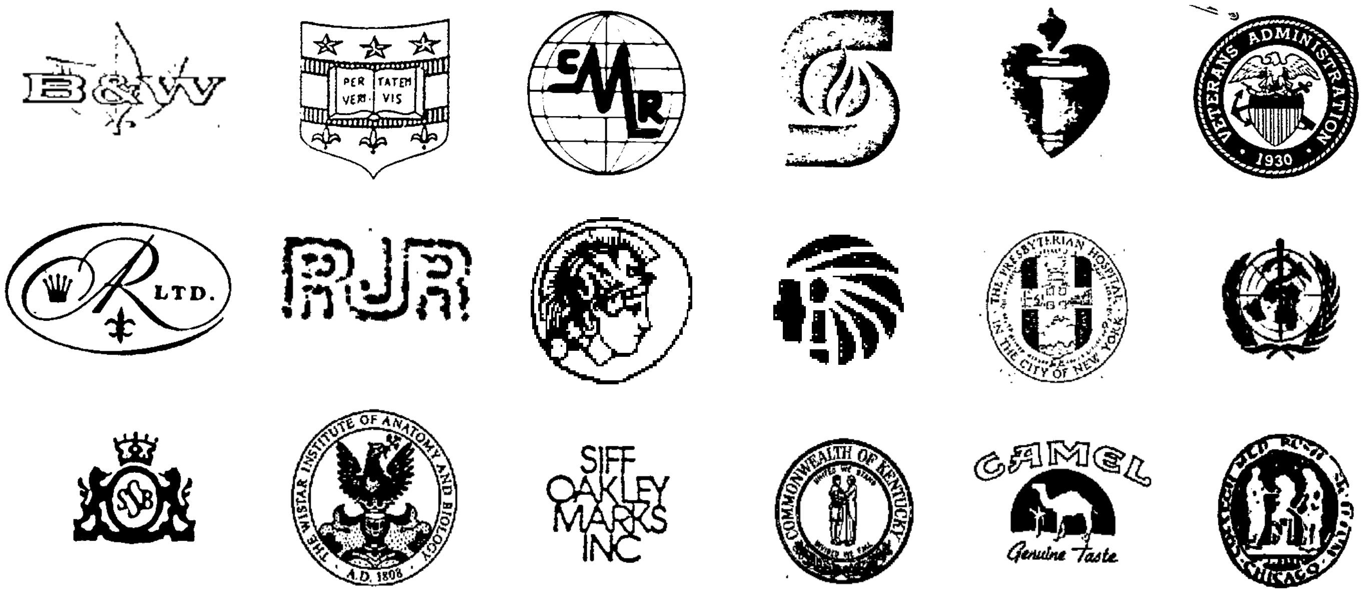
Challenges

- Extremely large intra-class variations among logos
- Continuum between graphics, logos and text



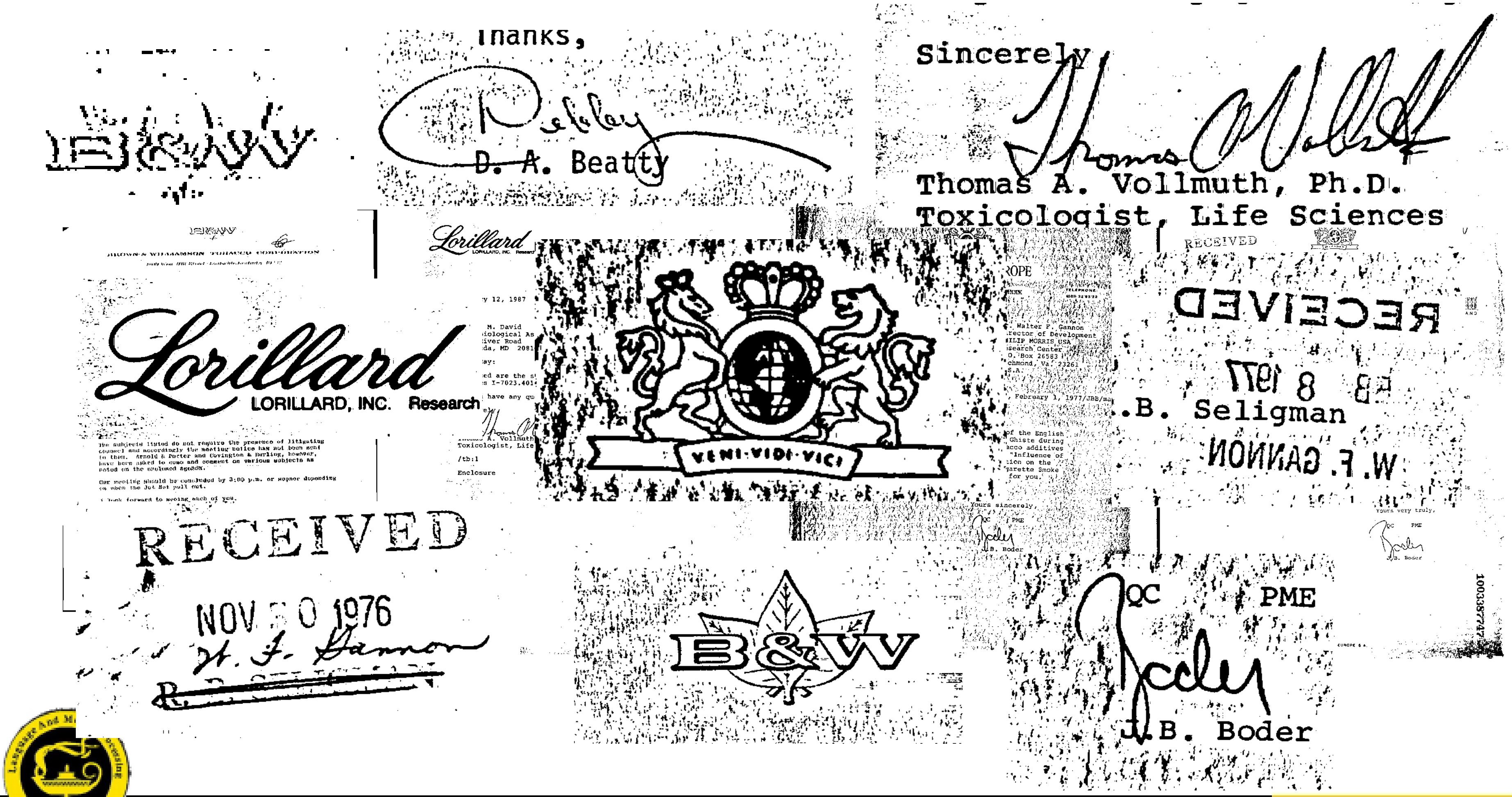
Challenges

- Diverse document layouts, scanning qualities, and image degradations on real document datasets



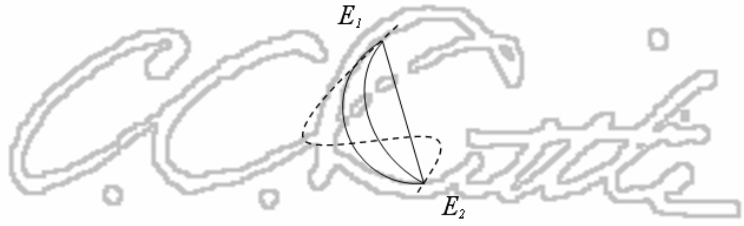
Claim #1

- Documents exists where *spatial* segmentation of Logos, Signatures and Stamps is not an option!


Thanks,
D. A. Beatty
Sincerely,
Thomas A. Vollmuth, Ph.D.
Toxicologist, Life Sciences
B&W
Lorillard
LORILLARD, INC. Research
VENI·VIDI·VICI
RECEIVED
.B. Seligman
RECEIVED
NOV 30 1976
H. F. Gannon
B&W
QC PME
J.B. Boder

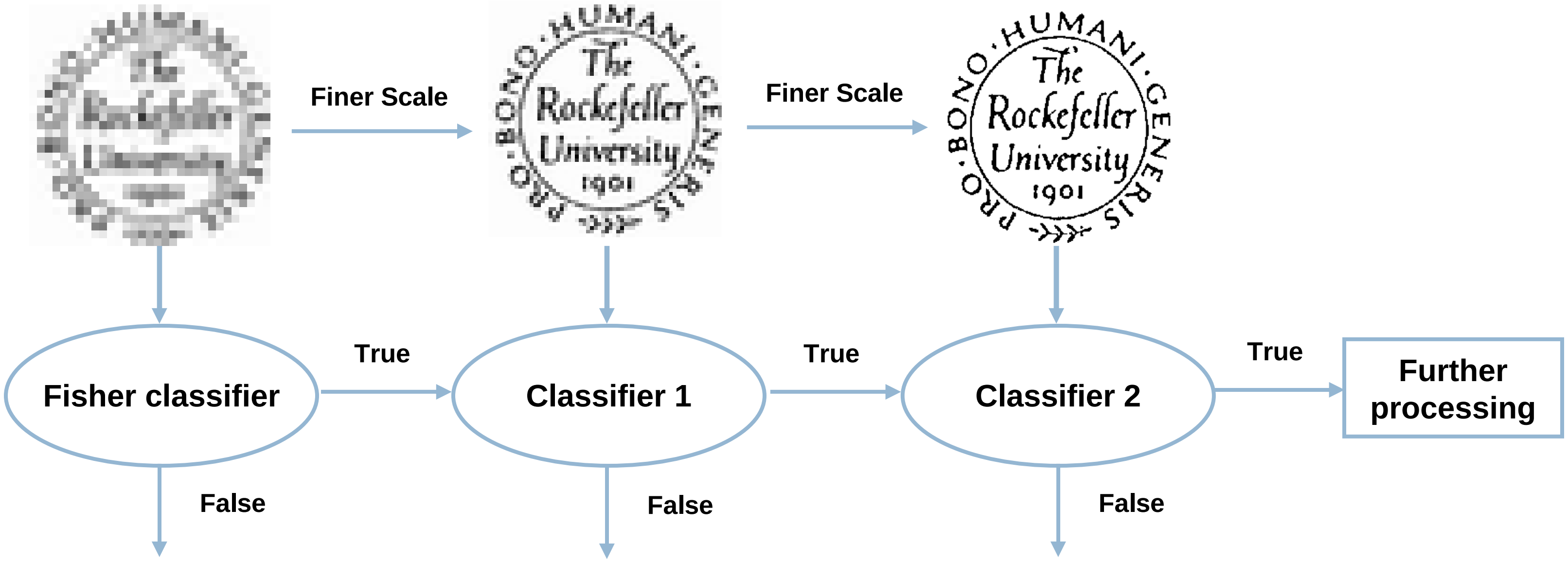
Claim #2

- Considering the more general problem of ***Detection*** *(as opposed to segmentation->classification)* allows us to integrate identification and extraction, and possibly recognition
- The concept has successfully been applied to:
  - *Guangyu Zhu, Yefeng Zheng, David Doermann and Stefan Jaeger. Multi-scale Structural Saliency for* ***Signature Detection****. (CVPR 2007).*
  - *Guangyu Zhu, Stefan Jaeger and David Doermann. A Robust* ***Stamp Detection*** *Framework on Degraded Documents. SPIE 2006.*



Multiscale Detection

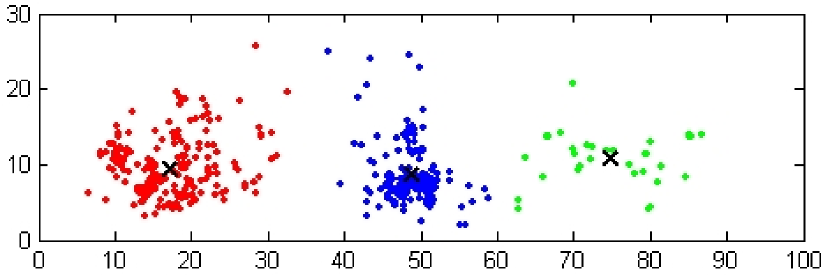
- Each logo candidate region is further classified at successively finer image scales by a cascade of simple classifiers



- The overall classifier is a strong learner, even if each individual classifier is in fact a weak learner

Feature selection and extraction

- How can we explore document context for logo detection?



Clustering result of logo positions using *k*-means (*k* = 3)

$$D_c(P) = \min_{i\in\{1,2,\cdots,k\}} (|p_x - c_x^i| + \lambda|p_y - c_y^i|)$$

- We define context distance as

Context Distance	**Area**	**Symmetry**
Spatial Density	**Aspect Ratio**	**Text Uniformity**

Evaluation

- We use tobacco-800, a large public dataset that consists of 1290 real-world documents (full dataset 49 million pages)
- Use accuracy and precision as evaluation metrics

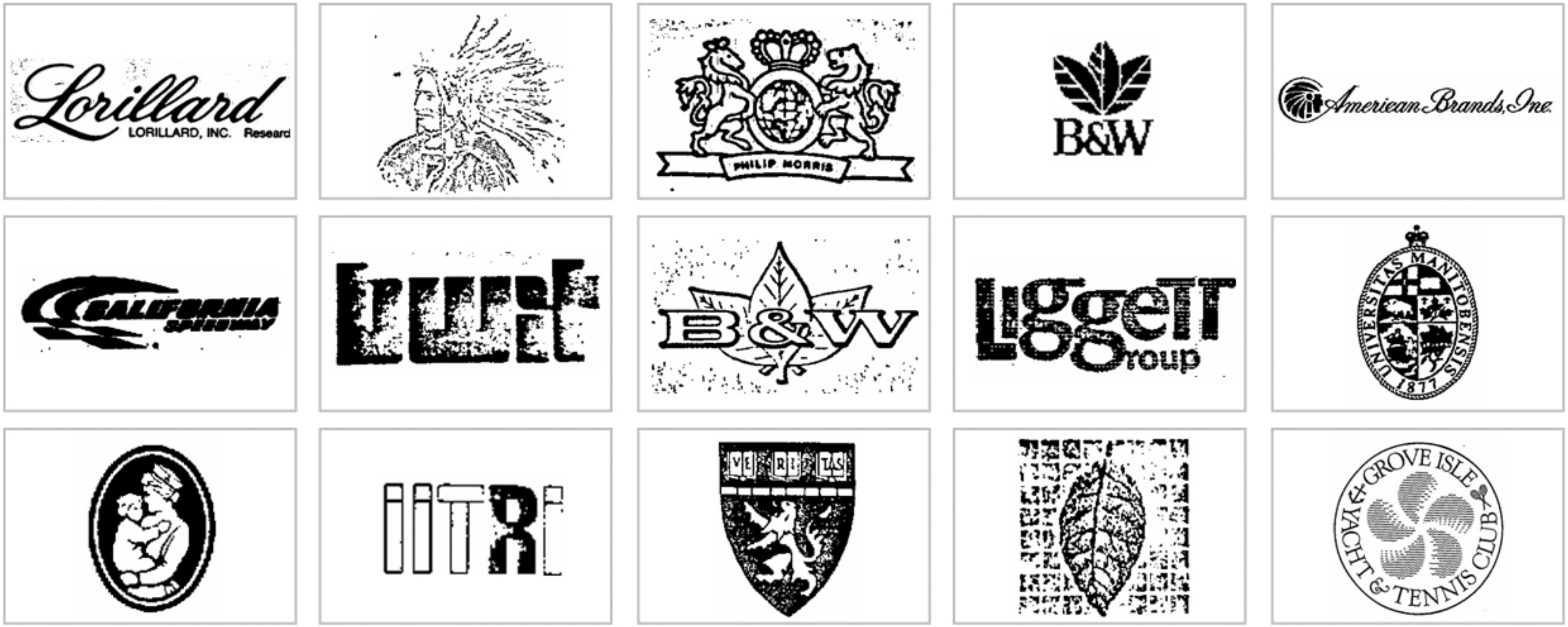
$$\text{Accuracy} = \frac{\#\text{ of correctly detected logos}}{\#\text{ of logos in groundtruth}} \qquad \text{Precision} = \frac{\#\text{ of correctly detected logos}}{\#\text{ of detected logos}}$$

- Detection is at least > 75% and < 125% pixel are overlap (determined from shape matching approach – Zhang. PAMI 2006)

Summary of logo detection performance on the Tobacco-800 dataset

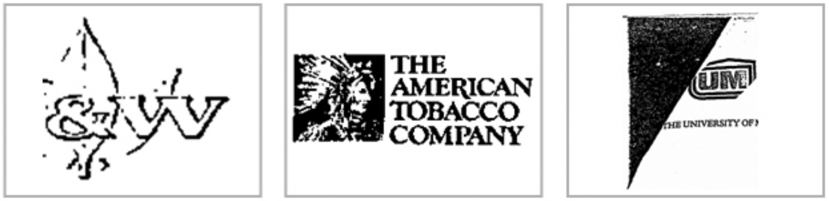
	Accuracy	Precision
Improved spatial density [9]	39.3%	32.1%
Fisher classifier only, *i.e.*, |S| = 1	59.2%	41.7%
Multi-scale approach with |S| = 2	57.0%	68.1%
Multi-scale approach with |S| = 3	84.2%	73.5%

Evaluation

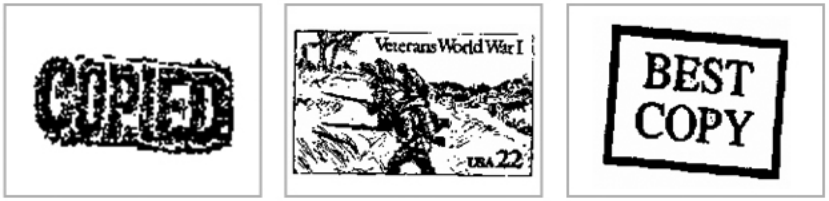


Examples of correctly detected logos from Tobacco-800

Evaluation

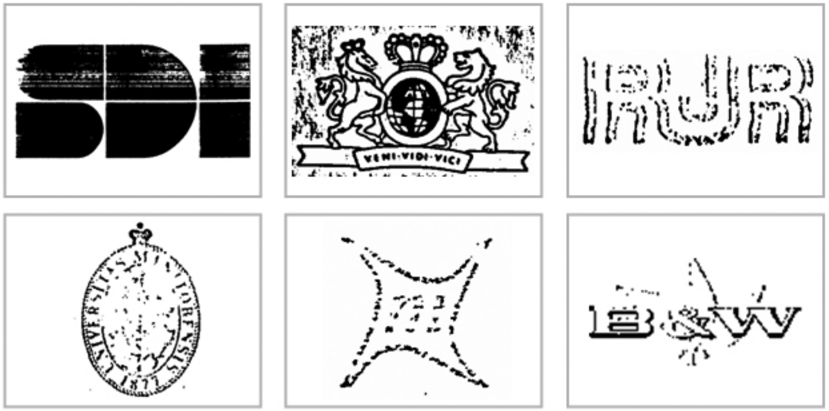


(a) Over/under-segmented logos



(b) Non logos

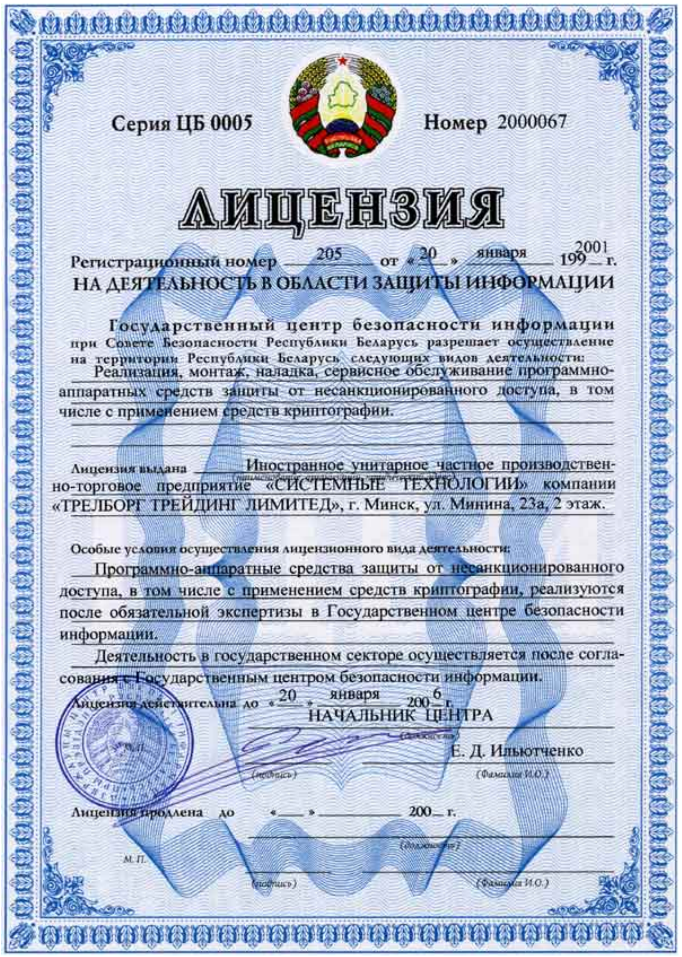
Examples of incorrectly detected logos



Examples of missed logos

Challenges in stamp detection

- Unique characteristics of stamps
  - Unstable and unpredictable patterns in documents
  - Outliers and occlusions are typical
  - Much lower spatial density compared to logo
  - Stamp instances appear as weaker regions within a full spectrum of background – text, figures, tables, watermark
  - Not generally valid to assume its location within the source



Серия ЦБ 0005 Номер 2000067

ЛИЦЕНЗИЯ

Регистрационный номер 205 от «20» января 199_ 2001 г.

НА ДЕЯТЕЛЬНОСТЬ В ОБЛАСТИ ЗАЩИТЫ ИНФОРМАЦИИ

Государственный центр безопасности информации при Совете Безопасности Республики Беларусь разрешает осуществление на территории Республики Беларусь следующих видов деятельности:

Реализация, монтаж, наладка, сервисное обслуживание программно-аппаратных средств защиты от несанкционированного доступа, в том числе с применением средств криптографии.

Лицензия выдана Иностранное унитарное частное производственно-торговое предприятие «СИСТЕМНЫЕ ТЕХНОЛОГИИ» компании «ТРЕЛБОРГ ТРЕЙДИНГ ЛИМИТЕД», г. Минск, ул. Минина, 23а, 2 этаж.

Особые условия осуществления лицензионного вида деятельности:

Программно-аппаратные средства защиты от несанкционированного доступа, в том числе с применением средств криптографии, реализуются после обязательной экспертизы в Государственном центре безопасности информации.

Деятельность в государственном секторе осуществляется после согласования с Государственным центром безопасности информации.

Лицензия действительна до «20» января 2006 г.

НАЧАЛЬНИК ЦЕНТРА
(должность)

М.П. (подпись) Е. Д. Ильютченко
(Фамилия И.О.)

Лицензия продлена до «__» ______ 200_ г.

(должность)

М. П. (подпись) (Фамилия И.О.)

Our stamp detection approach

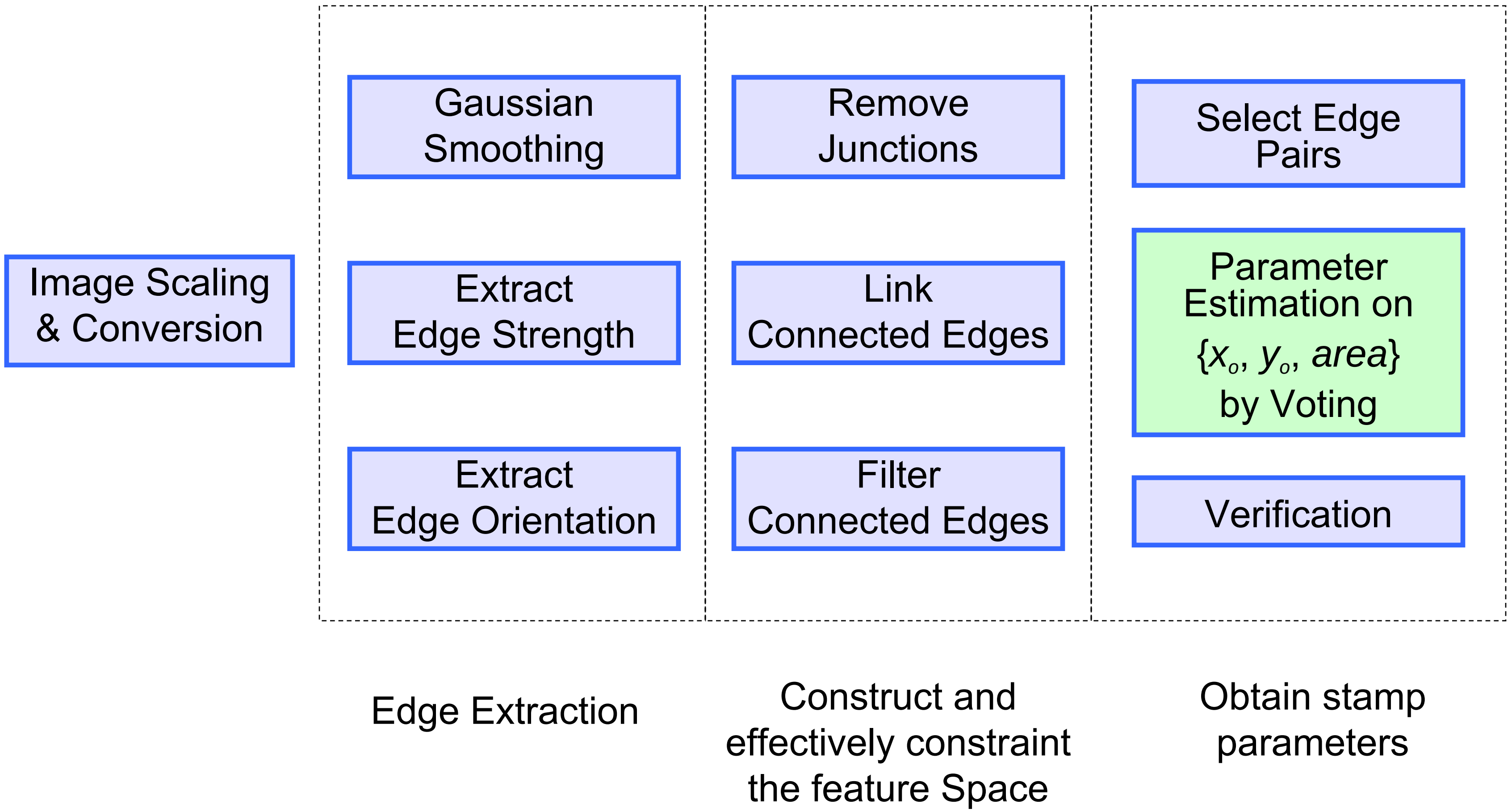


Image Scaling & Conversion

Gaussian Smoothing

Extract Edge Strength

Extract Edge Orientation

Edge Extraction

Remove Junctions

Link Connected Edges

Filter Connected Edges

Construct and effectively constraint the feature Space

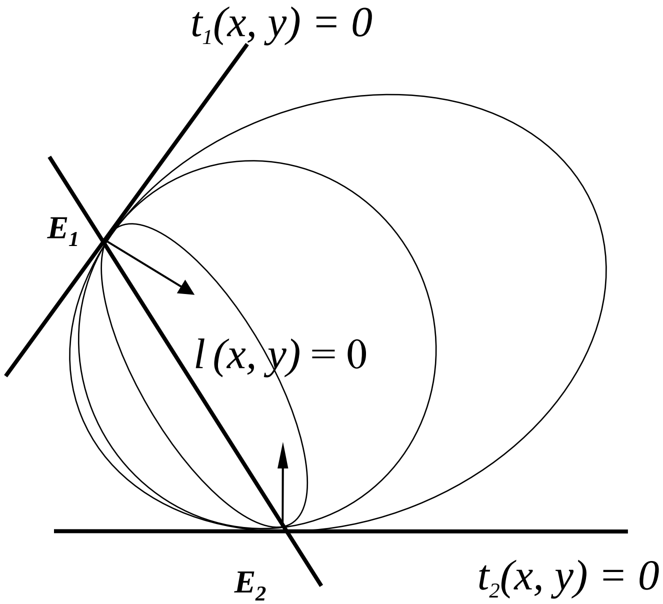
Select Edge Pairs

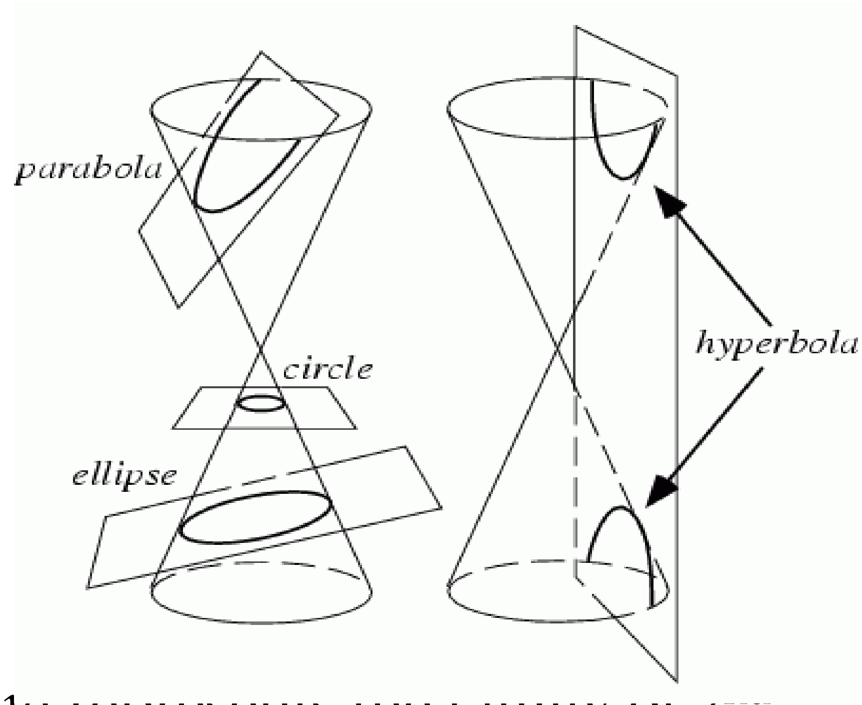
Parameter Estimation on $\{x_o, y_o, area\}$ by Voting

Verification

Obtain stamp parameters

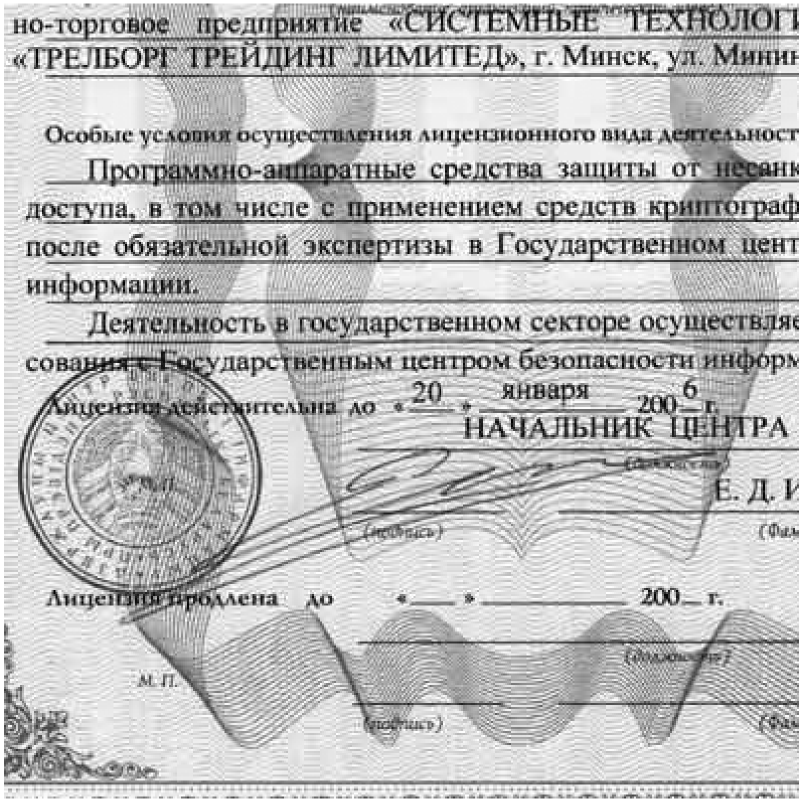
Ellipse detection method using pairs of edges





The quadratic function $f(x, y)$ represents the family of 2^{nd}-order curves that pass points E_1 and E_2 and tangent to lines $t_1(x, y)$ and $t_2(x, y)$.

Demo



но-торговое предприятие «СИСТЕМНЫЕ ТЕХНОЛОГИ
«ТРЕЛБОРГ ТРЕЙДИНГ ЛИМИТЕД», г. Минск, ул. Минин

Особые условия осуществления лицензионного вида деятельност

Программно-аппаратные средства защиты от несанк
доступа, в том числе с применением средств криптограф
после обязательной экспертизы в Государственном цент
информации.

Деятельность в государственном секторе осуществляе
сования с Государственным центром безопасности информ

Лицензия действительна до «20» января 2006 г.

НАЧАЛЬНИК ЦЕНТРА

(должность)

(подпись)

Е. Д. И

(Фам

Лицензия продлена до «___» ___________ 200_ г.

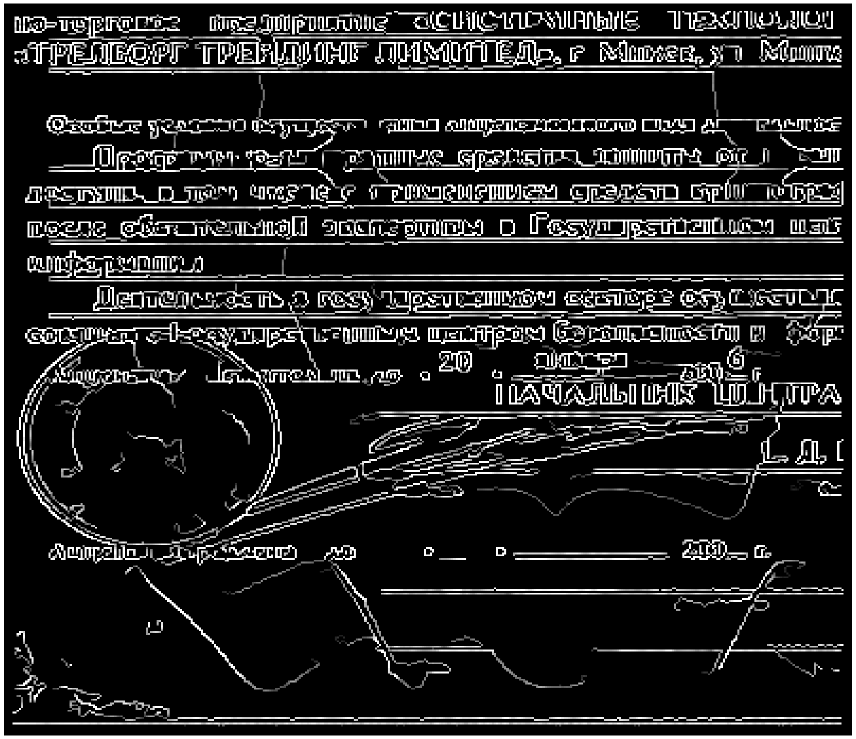
(должность)

М. П.

(подпись)

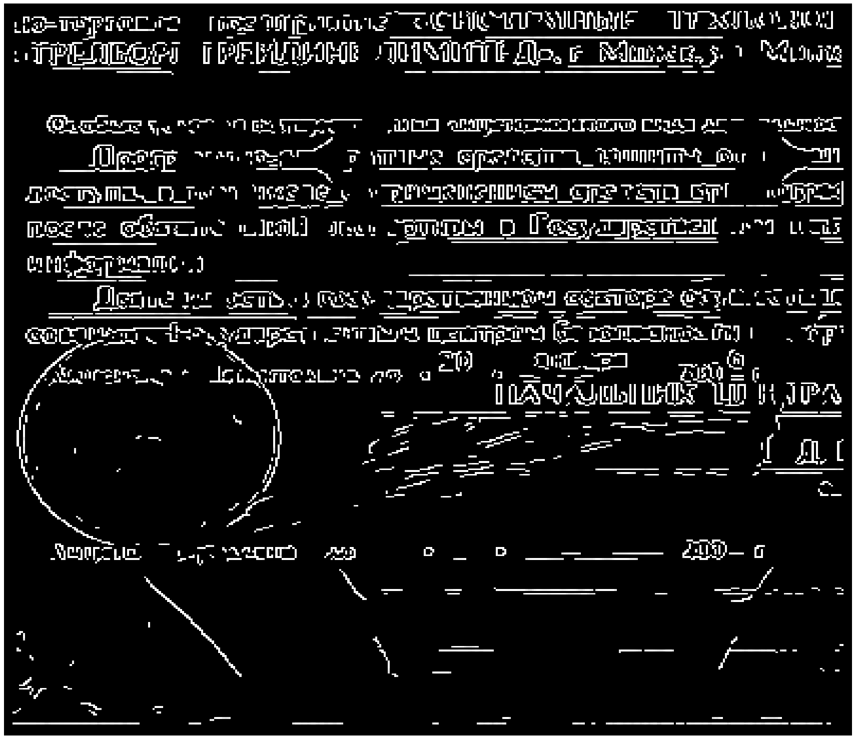
(Фам

Region of a sample image

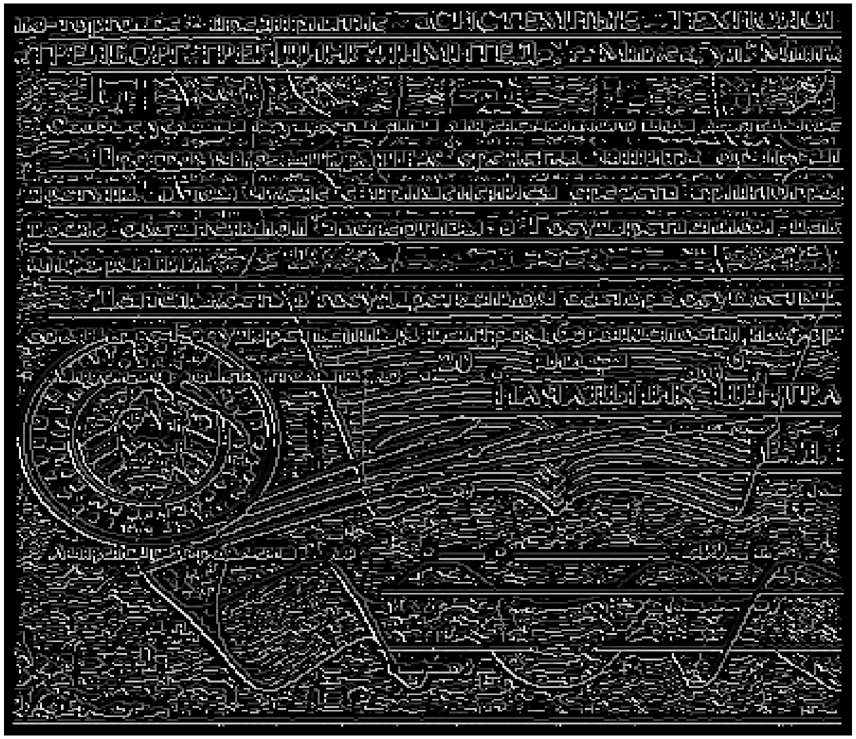


Strength of edge gradient

Demo

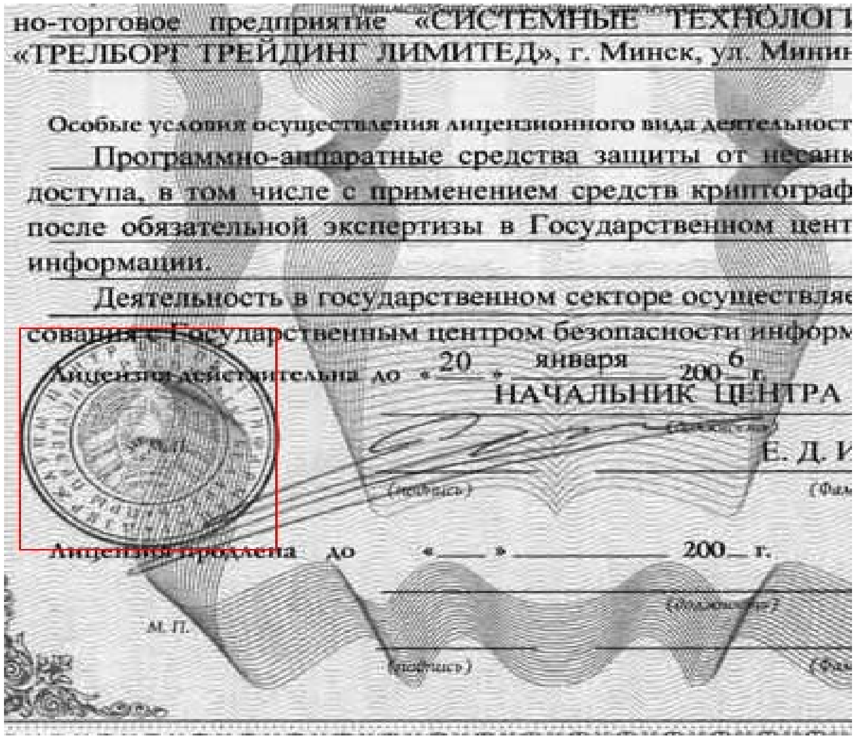


Strong edges



Orientation of edge gradient

Demo



но-торговое предприятие «СИСТЕМНЫЕ ТЕХНОЛОГІ
«ТРЕЛБОРГ ТРЕЙДИНГ ЛИМИТЕД», г. Минск, ул. Минин

Особые условия осуществления лицензионного вида деятельност

Программно-аппаратные средства защиты от несанк
доступа, в том числе с применением средств криптограф
после обязательной экспертизы в Государственном цент
информации.

Деятельность в государственном секторе осуществляе
сования с Государственным центром безопасности информ

Лицензия действительна до «20» января 2006 г.

НАЧАЛЬНИК ЦЕНТРА

(должность)

Е. Д. И

(подпись) (Фам

Лицензия продлена до «___» __________ 200_ г.

(должность)

М. П.

(подпись) (Фам

Top 10 candidates in the 3-D parameter space in ellipse center and area, i.e. (x_o, y_o, $area$)

(68, 238, 11313), score = [5485509]

(56, 202, 6464), score = [501958]

(52, 226, 8080), score = [431456]

(72, 206, 8080), score = [352608]

(84, 266, 6464), score = [278291]

(84, 210, 6464), score = [260775]

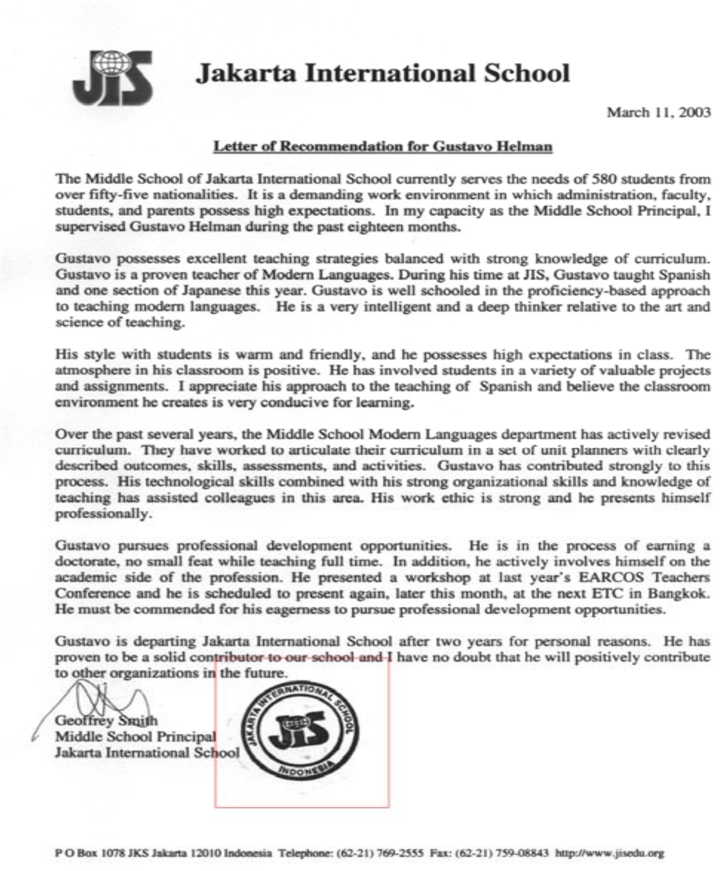
(44, 222, 8080), score = [247448]

(28, 270, 3232), score = [241991]

(40, 202, 4848), score = [224263]

(76, 230, 9696), score = [215384]

Demo



Jakarta International School

March 11, 2003

Letter of Recommendation for Gustavo Helman

The Middle School of Jakarta International School currently serves the needs of 580 students from over fifty-five nationalities. It is a demanding work environment in which administration, faculty, students, and parents possess high expectations. In my capacity as the Middle School Principal, I supervised Gustavo Helman during the past eighteen months.

Gustavo possesses excellent teaching strategies balanced with strong knowledge of curriculum. Gustavo is a proven teacher of Modern Languages. During his time at JIS, Gustavo taught Spanish and one section of Japanese this year. Gustavo is well schooled in the proficiency-based approach to teaching modern languages. He is a very intelligent and a deep thinker relative to the art and science of teaching.

His style with students is warm and friendly, and he possesses high expectations in class. The atmosphere in his classroom is positive. He has involved students in a variety of valuable projects and assignments. I appreciate his approach to the teaching of Spanish and believe the classroom environment he creates is very conducive for learning.

Over the past several years, the Middle School Modern Languages department has actively revised curriculum. They have worked to articulate their curriculum in a set of unit planners with clearly described outcomes, skills, assessments, and activities. Gustavo has contributed strongly to this process. His technological skills combined with his strong organizational skills and knowledge of teaching has assisted colleagues in this area. His work ethic is strong and he presents himself professionally.

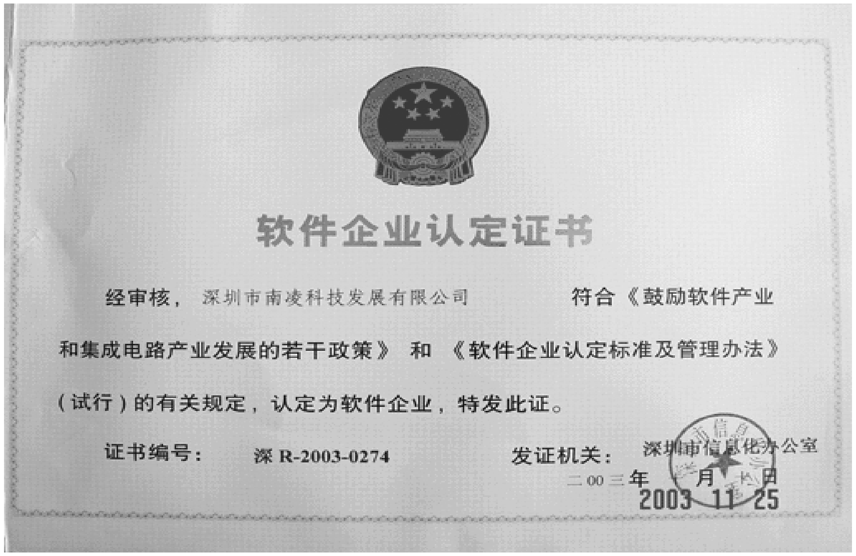
Gustavo pursues professional development opportunities. He is in the process of earning a doctorate, no small feat while teaching full time. In addition, he actively involves himself on the academic side of the profession. He presented a workshop at last year's EARCOS Teachers Conference and he is scheduled to present again, later this month, at the next ETC in Bangkok. He must be commended for his eagerness to pursue professional development opportunities.

Gustavo is departing Jakarta International School after two years for personal reasons. He has proven to be a solid contributor to our school and I have no doubt that he will positively contribute to other organizations in the future.

Geoffrey Smith
Middle School Principal
Jakarta International School

JAKARTA INTERNATIONAL SCHOOL
JIS
INDONESIA

P O Box 1078 JKS Jakarta 12010 Indonesia Telephone: (62-21) 769-2555 Fax: (62-21) 759-08843 http://www.jisedu.org



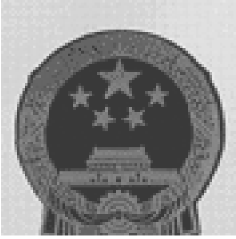
软件企业认定证书

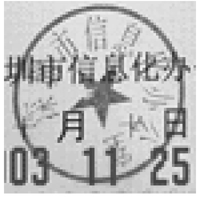
经审核，深圳市南凌科技发展有限公司　　符合《鼓励软件产业和集成电路产业发展的若干政策》和《软件企业认定标准及管理办法》（试行）的有关规定，认定为软件企业，特发此证。

证书编号：　深 R-2003-0274　　发证机关：深圳市信息化办公室

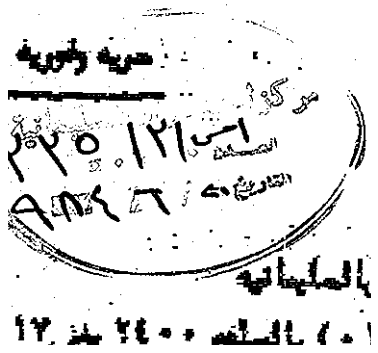
二00三年　月　日

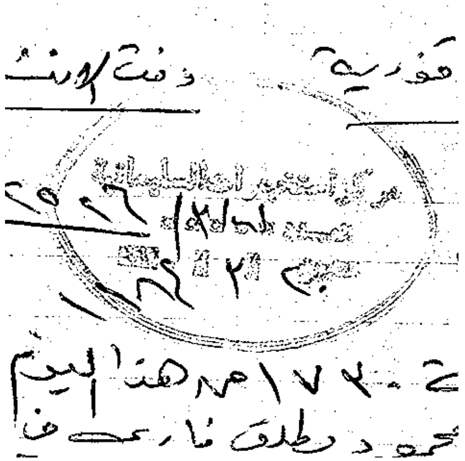
2003 11 25

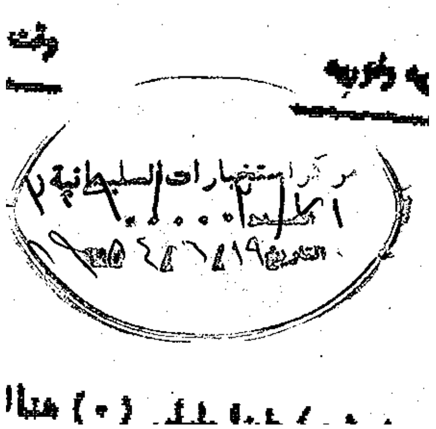


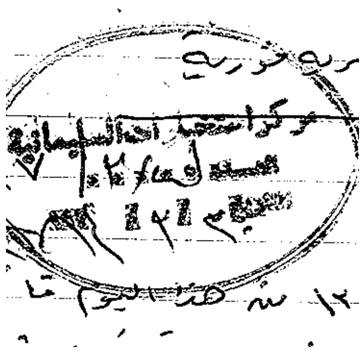


Demo



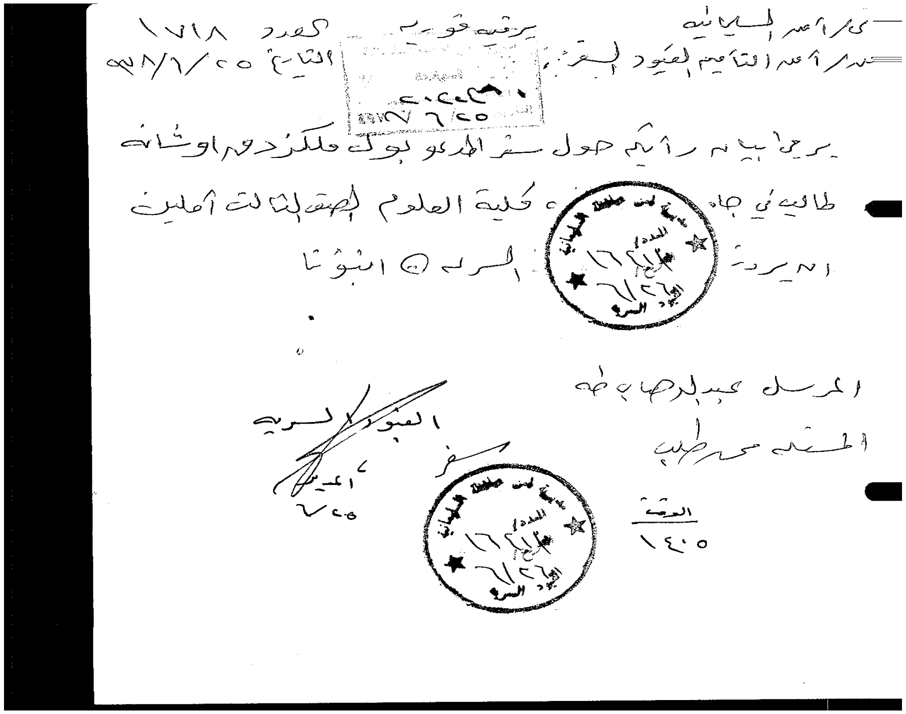

مركز استخبارات السليمانية


مركز استخبارات السليمانية ٢


مركز استخبارات السليمانية

Demo

Capability to detect multiple stamp instances



الى/ أمن السليمانية برقية فورية العدد ١٧١٨

من/ أمن التأميم العقود السرية التاريخ ٩٨/٦/٢٥

يرجى بيان رأيكم حول سفر المدعو بوك فلكرد وعراوشانه

طالب في جامعة كلية العلوم الصف الثالث أهلين

المربوط المرحلة ⑤ اشوشا

المرسل عبد الرحمن طه

المشغل محمد طيب

العقود السرية

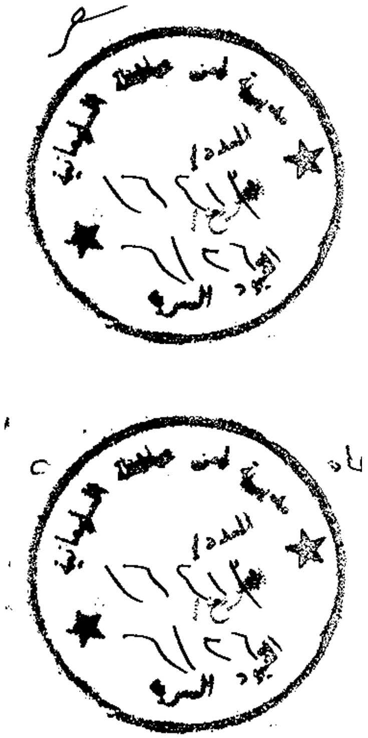
العقدة ١٤٠٥

مديرية أمن محافظة السليمانية

العدد/

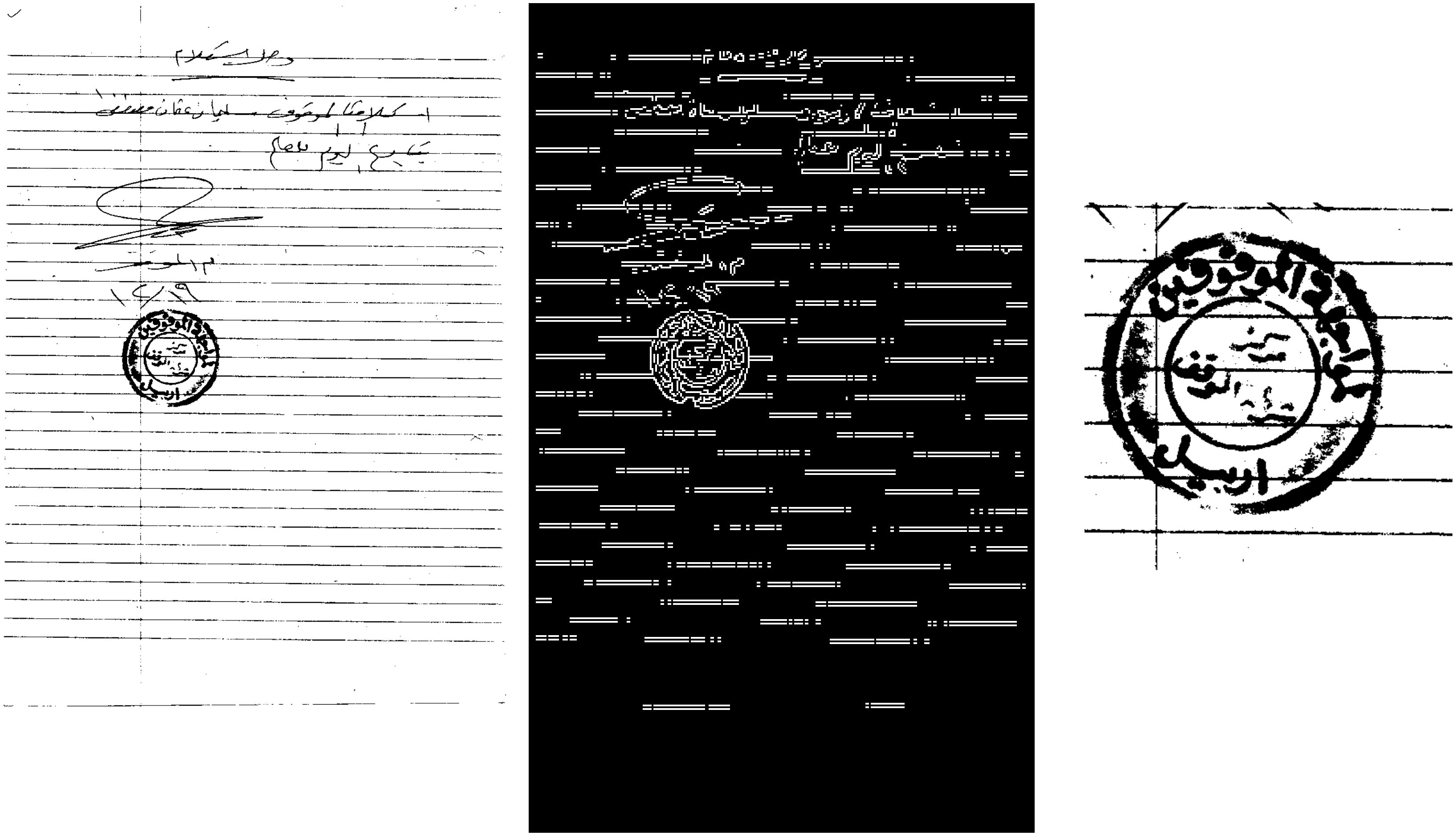
التاريخ

العقود السرية



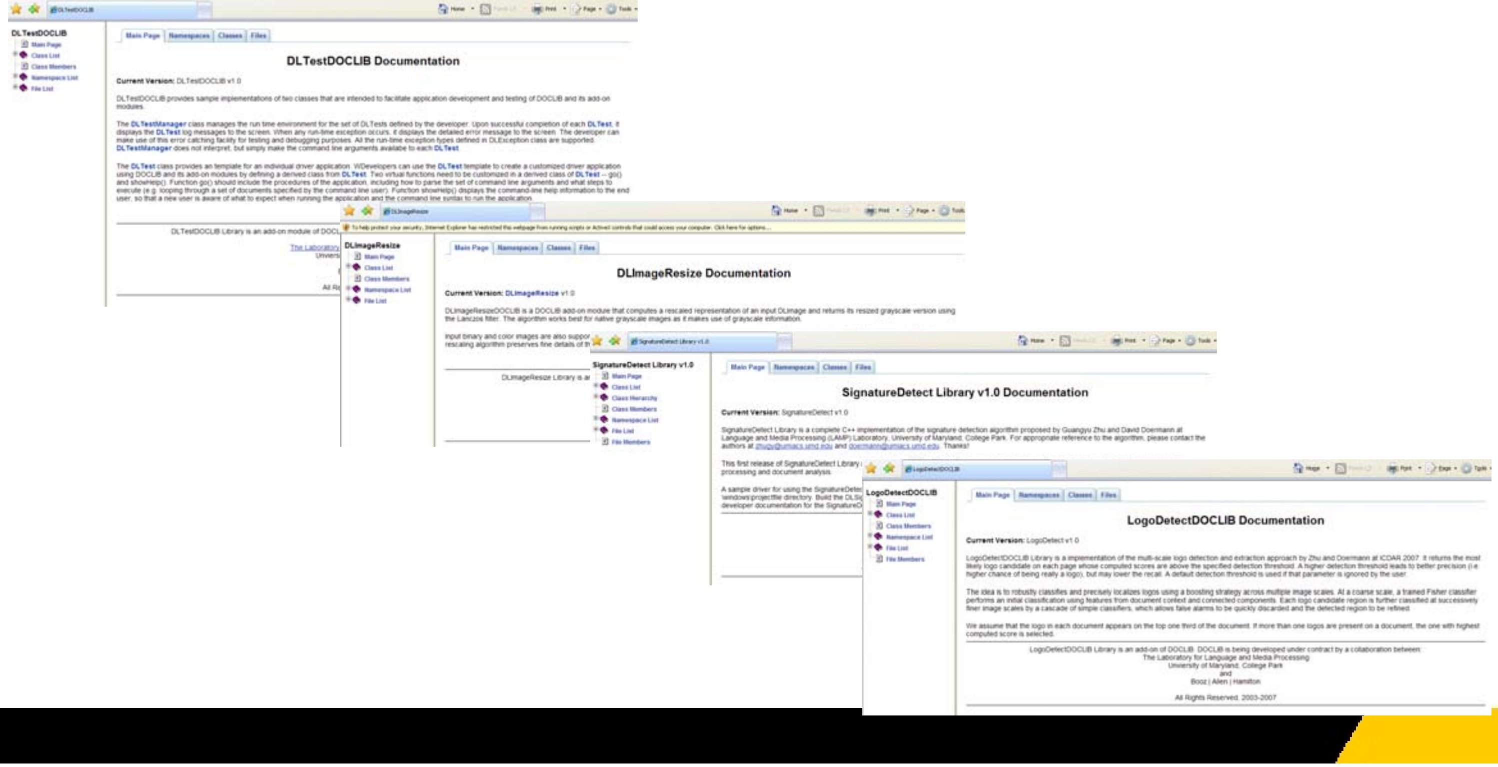
Demo

Capability to detect stamp instances in diverse backgrounds



Software releases

- Signature detection and logo detection code are released as Doclib add-on modules
- Production test on 32,000+ documents
- Signature matching and logo matching expected



DLTestDOCLIB Documentation

Current Version: DLTestDOCLIB v1.0

DLTestDOCLIB provides sample implementations of two classes that are intended to facilitate application development and testing of DOCLIB and its add-on modules.

The **DLTestManager** class manages the run time environment for the set of DLTests defined by the developer. Upon successful completion of each **DLTest**, it displays the **DLTest** log messages to the screen. When any run-time exception occurs, it displays the detailed error message to the screen. The developer can make use of this error catching facility for testing and debugging purposes. All the run-time exception types defined in DLException class are supported. **DLTestManager** does not interpret, but simply make the command line arguments avalabe to each **DLTest**.

The **DLTest** class provides an template for an individual driver application. WDevelopers can use the **DLTest** template to create a customized driver application using DOCLIB and its add-on modules by defining a derived class from **DLTest**. Two virtual functions need to be customized in a derived class of **DLTest** -- go() and showHelp(). Function go() should include the procedures of the application, including how to parse the set of command line arguments and what steps to execute (e.g. looping through a set of documents specified by the command line user). Function showHelp() displays the command-line help information to the end user, so that a new user is aware of what to expect when running the application and the command line syntax to run the application.

DLTestDOCLIB Library is an add-on module of DOCL

The Laboratory
Univers

All Ri

To help protect your security, Internet Explorer has restricted this webpage from running scripts or ActiveX controls that could access your computer. Click here for options...

DLImageResize Documentation

Current Version: DLImageResize v1.0

DLImageResizeDOCLIB is a DOCLIB add-on module that computes a rescaled representation of an input DLImage and returns its resized grayscale version using the Lanczos filter. The algorithm works best for native grayscale images as it makes use of grayscale information.

Input binary and color images are also suppor
rescaling algorithm preserves fine details of th

DLImageResize Library is ar

SignatureDetect Library v1.0 Documentation

Current Version: SignatureDetect v1.0

SignatureDetect Library is a complete C++ implementation of the signature detection algorithm proposed by Guangyu Zhu and David Doermann at Language and Media Processing (LAMP) Laboratory, University of Maryland, College Park. For appropriate reference to the algorithm, please contact the authors at zhugy@umiacs.umd.edu and doermann@umiacs.umd.edu. Thanks!

This first release of SignatureDetect Library
processing and document analysis.

A sample driver for using the SignatureDetec
\windows\project\file directory. Build the DLSig
developer documentation for the SignatureD

LogoDetectDOCLIB Documentation

Current Version: LogoDetect v1.0

LogoDetectDOCLIB Library is a implementation of the multi-scale logo detection and extraction approach by Zhu and Doermann at ICDAR 2007. It returns the most likely logo candidate on each page whose computed scores are above the specified detection threshold. A higher detection threshold leads to better precision (i.e. higher chance of being really a logo), but may lower the recall. A default detection threshold is used if that parameter is ignored by the user.

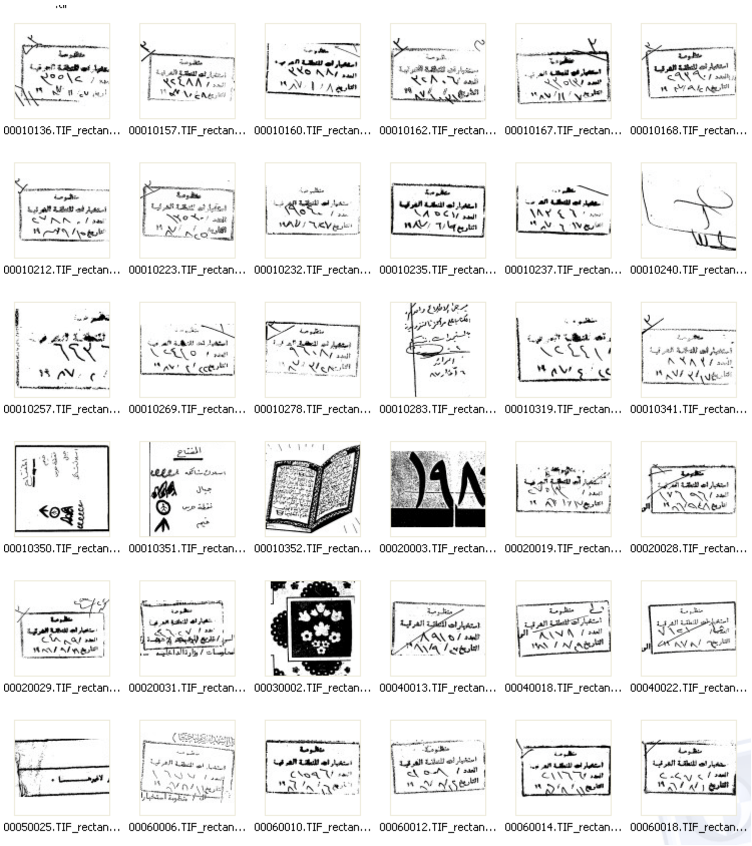
The idea is to robustly classifies and precisely localizes logos using a boosting strategy across multiple image scales. At a coarse scale, a trained Fisher classifier performs an initial classification using features from document context and connected components. Each logo candidate region is further classified at successively finer image scales by a cascade of simple classifiers, which allows false alarms to be quickly discarded and the detected region to be refined.

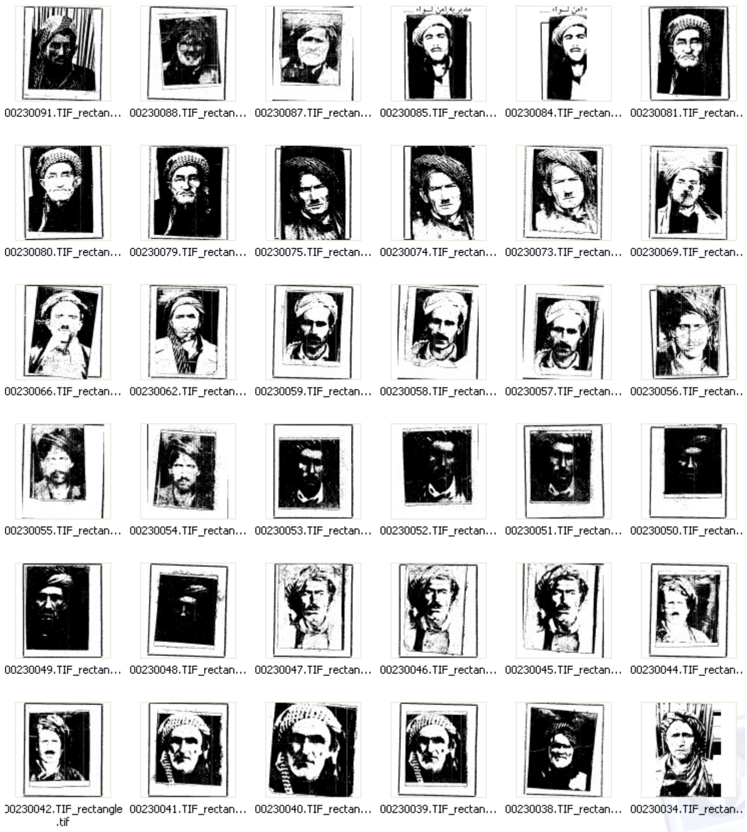
We assume that the logo in each document appears on the top one third of the document. If more than one logos are present on a document, the one with highest computed score is selected.

LogoDetectDOCLIB Library is an add-on of DOCLIB. DOCLIB is being developed under contract by a collaboration between:
The Laboratory for Language and Media Processing
University of Maryland, College Park
and
Booz | Allen | Hamilton

All Rights Reserved, 2003-2007

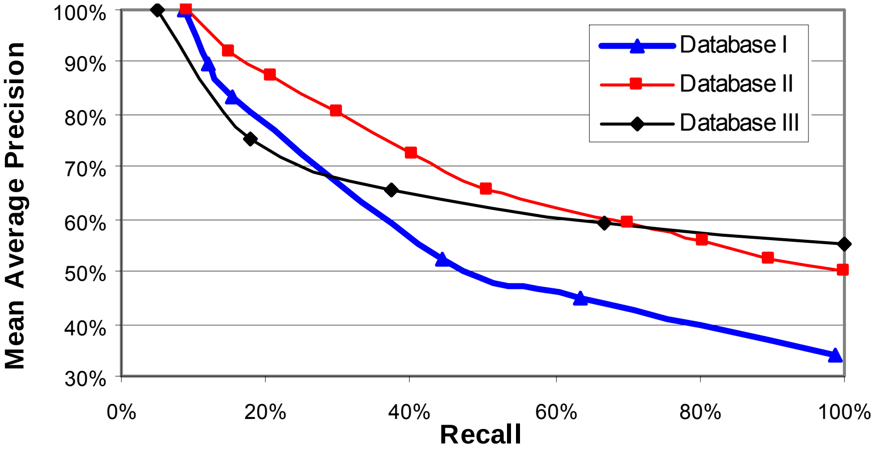
Demo


00010136.TIF_rectan...
00010157.TIF_rectan...
00010160.TIF_rectan...
00010162.TIF_rectan...
00010167.TIF_rectan...
00010168.TIF_rectan...
00010212.TIF_rectan...
00010223.TIF_rectan...
00010232.TIF_rectan...
00010235.TIF_rectan...
00010237.TIF_rectan...
00010240.TIF_rectan...
00010257.TIF_rectan...
00010269.TIF_rectan...
00010278.TIF_rectan...
00010283.TIF_rectan...
00010319.TIF_rectan...
00010341.TIF_rectan...
00010350.TIF_rectan...
00010351.TIF_rectan...
00010352.TIF_rectan...
00020003.TIF_rectan...
00020019.TIF_rectan...
00020028.TIF_rectan...
00020029.TIF_rectan...
00020031.TIF_rectan...
00030002.TIF_rectan...
00040013.TIF_rectan...
00040018.TIF_rectan...
00040022.TIF_rectan...
00050025.TIF_rectan...
00060006.TIF_rectan...
00060010.TIF_rectan...
00060012.TIF_rectan...
00060014.TIF_rectan...
00060018.TIF_rectan...


00230091.TIF_rectan...
00230088.TIF_rectan...
00230087.TIF_rectan...
00230085.TIF_rectan...
00230084.TIF_rectan...
00230081.TIF_rectan..
00230080.TIF_rectan...
00230079.TIF_rectan...
00230075.TIF_rectan...
00230074.TIF_rectan...
00230073.TIF_rectan...
00230069.TIF_rectan..
00230066.TIF_rectan...
00230062.TIF_rectan...
00230059.TIF_rectan...
00230058.TIF_rectan...
00230057.TIF_rectan...
00230056.TIF_rectan..
00230055.TIF_rectan...
00230054.TIF_rectan...
00230053.TIF_rectan...
00230052.TIF_rectan...
00230051.TIF_rectan...
00230050.TIF_rectan..
00230049.TIF_rectan...
00230048.TIF_rectan...
00230047.TIF_rectan...
00230046.TIF_rectan...
00230045.TIF_rectan...
00230044.TIF_rectan..
)0230042.TIF_rectangle .tif
00230041.TIF_rectan...
00230040.TIF_rectan...
00230039.TIF_rectan...
00230038.TIF_rectan...
00230034.TIF_rectan..

Experiment

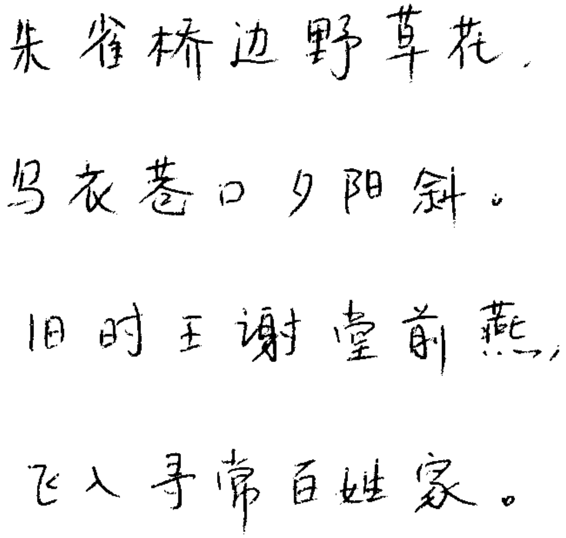
Test Databases	Total Images	Images with The Retrieved Stamp
Database 1	436	92 (Circular)
Database 2	193	68 (Elliptic)
Database 3	287	102 (Rectangular)



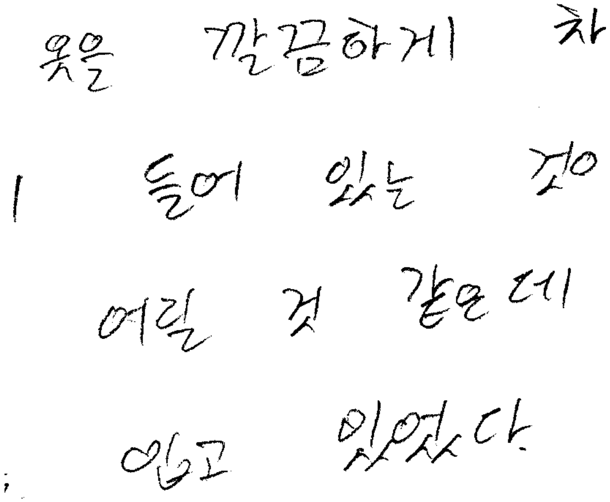
Script and ImageID

- 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.

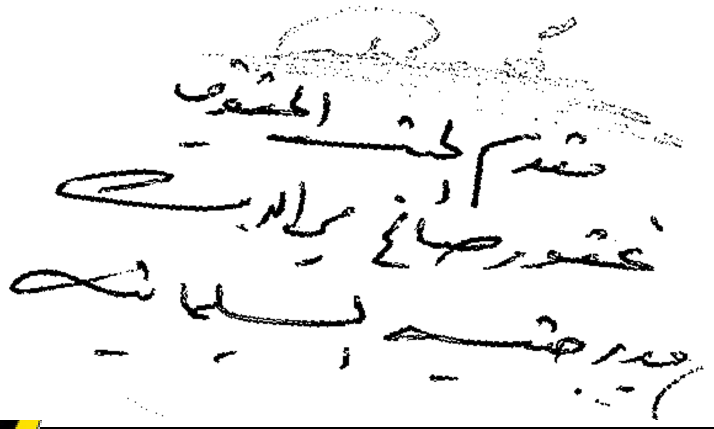
The Observation



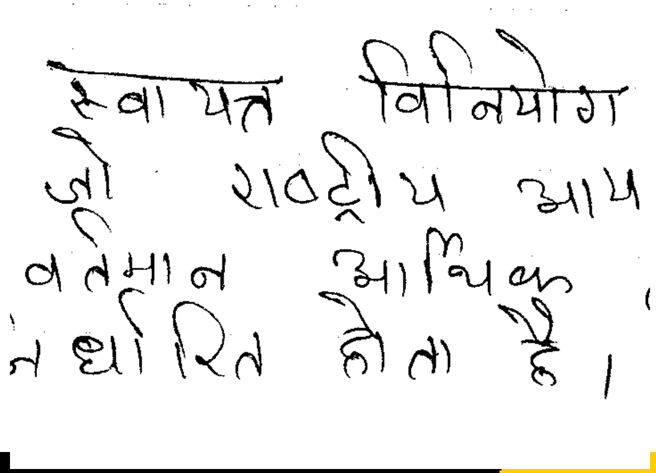
朱雀桥边野草花.
乌衣巷口夕阳斜。
旧时王谢堂前燕,
飞入寻常百姓家。



옷을 깔끔하게 차
| 들어 있는 것이
어린 것 같은데
입고 있었다



[illegible]



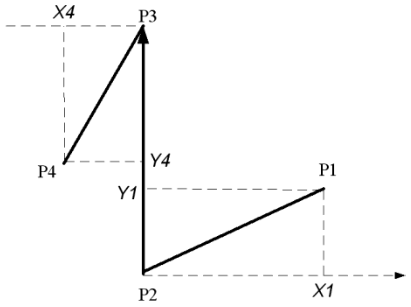
स्वायत विनियोग
जो राष्ट्रीय आय
वर्तमान आर्थिक
निर्धारित होता है।

The Observation (con’t)

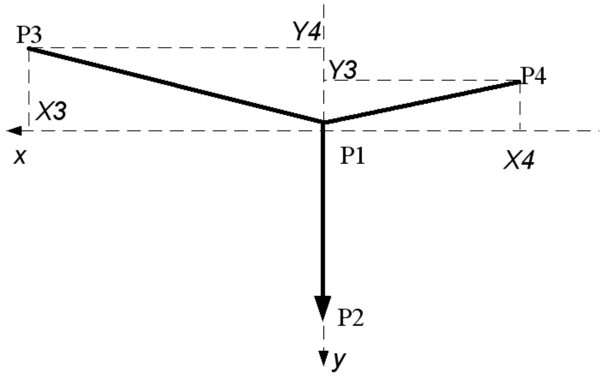
- The relationship of connected edges could be used for description;
- The dominant descriptors for different scripts could be different;
- The statistics of the descriptors could be used for discriminating different scripts.

The descriptor

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



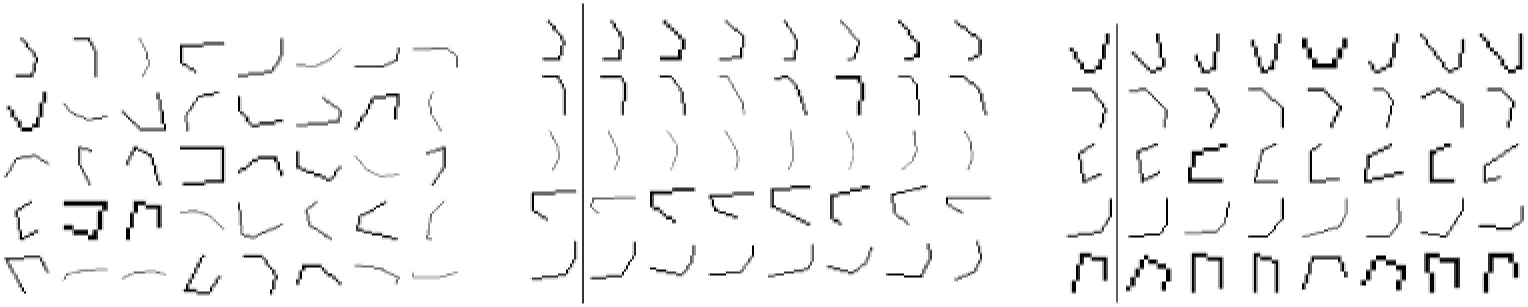
  - Y shape



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



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

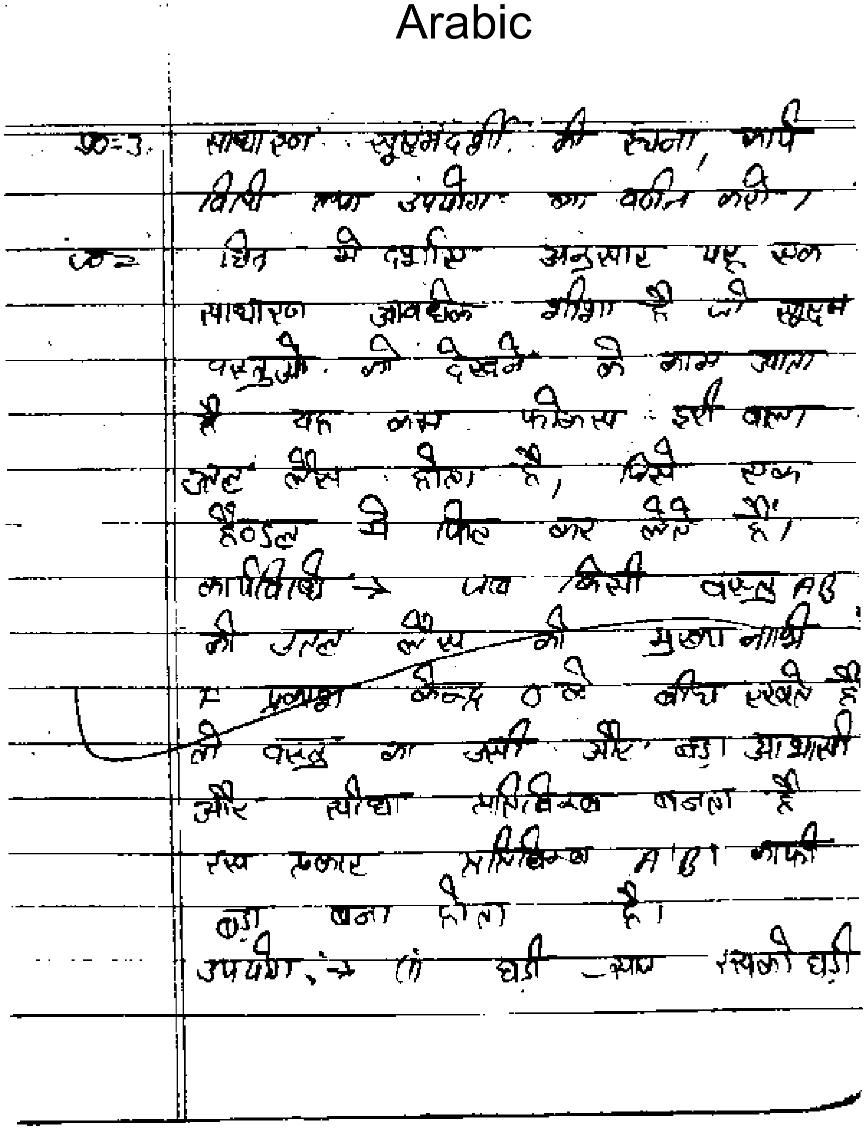
Result

- 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



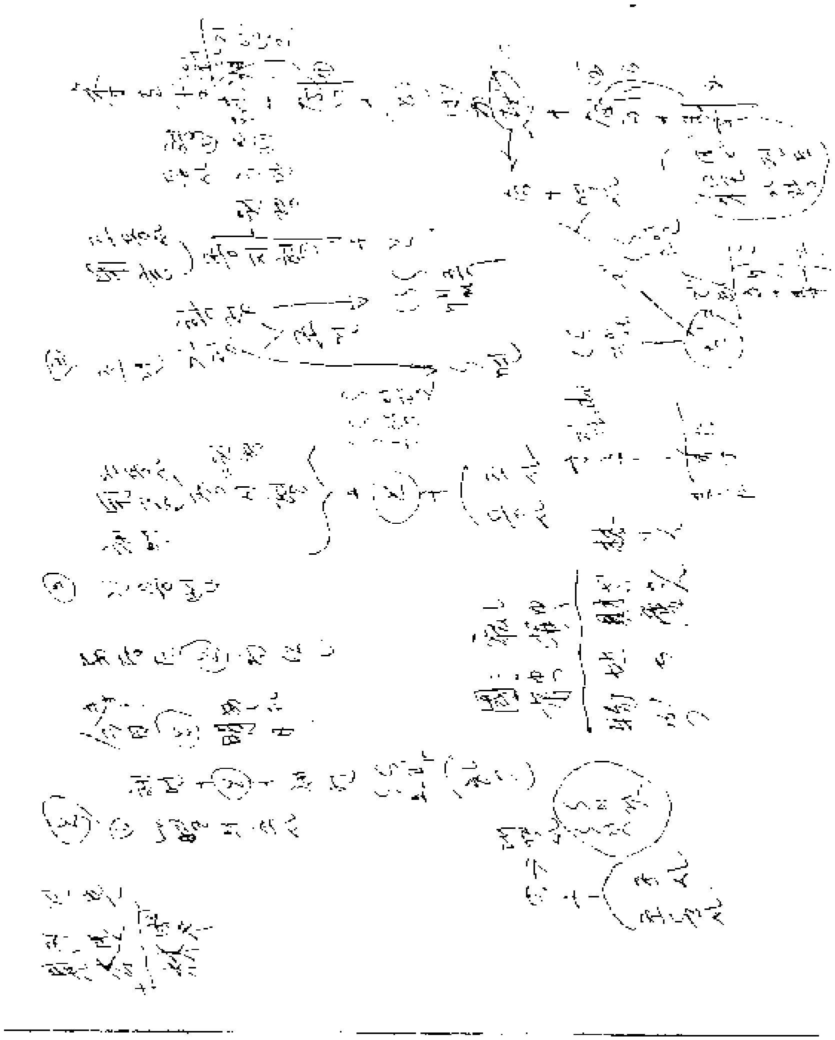
प्र०3. साधारण सूक्ष्मदर्शी की रचना, कार्य विधि तथा उपयोग का वर्णन करो।

उ०2 चित्र मे 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 मे फिट कर लेते हैं।

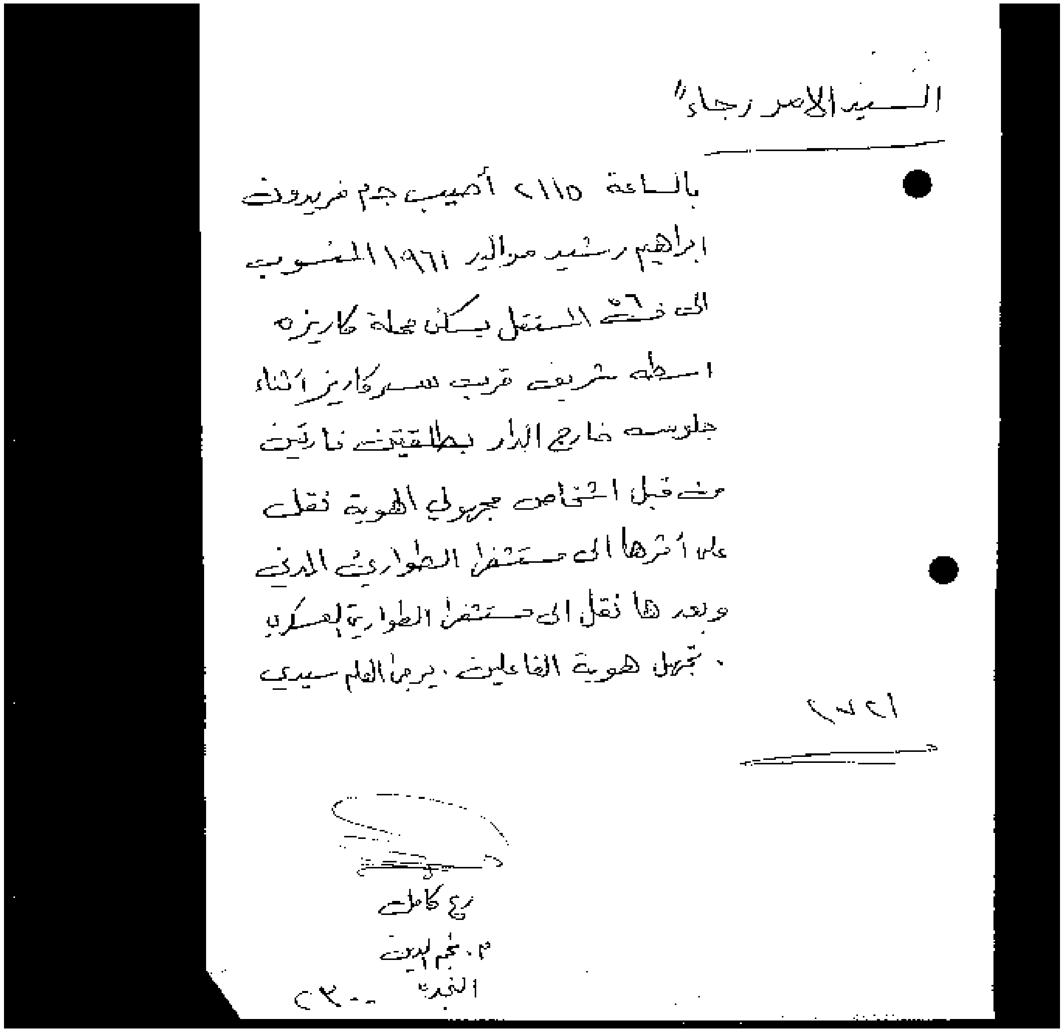
कार्यविधि → जब किसी वस्तु AB को उत्तल लैंस के मुख्य नाभि व प्रकाश केन्द्र O के बीच रखते हैं तो वस्तु का उसी ओर बड़ा आभासी और सीधा प्रतिबिम्ब बनता है इस प्रकार प्रतिबिम्ब A'B' काफी बड़ा बना होता है।

उपयोग → (i) घड़ी साज घड़ी

Chinese



Failure example (Korean)



السيد الآمر رجاءً

بالساعة ٢١١٥ أصيب جرم فريدون
ابراهيم رشيد مواليد ١٩٦١ المنسوب
الى فوج ٦ المستقل بسكن محلة كاريزه
اسطه شريف قريب سركاريز 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

١/٧/٢

ر ع كامل
م . نجم الدين
النقيب
٢٠٠

Image ID

- Determine which class:
  - Text, Image w/text, or image
- 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.

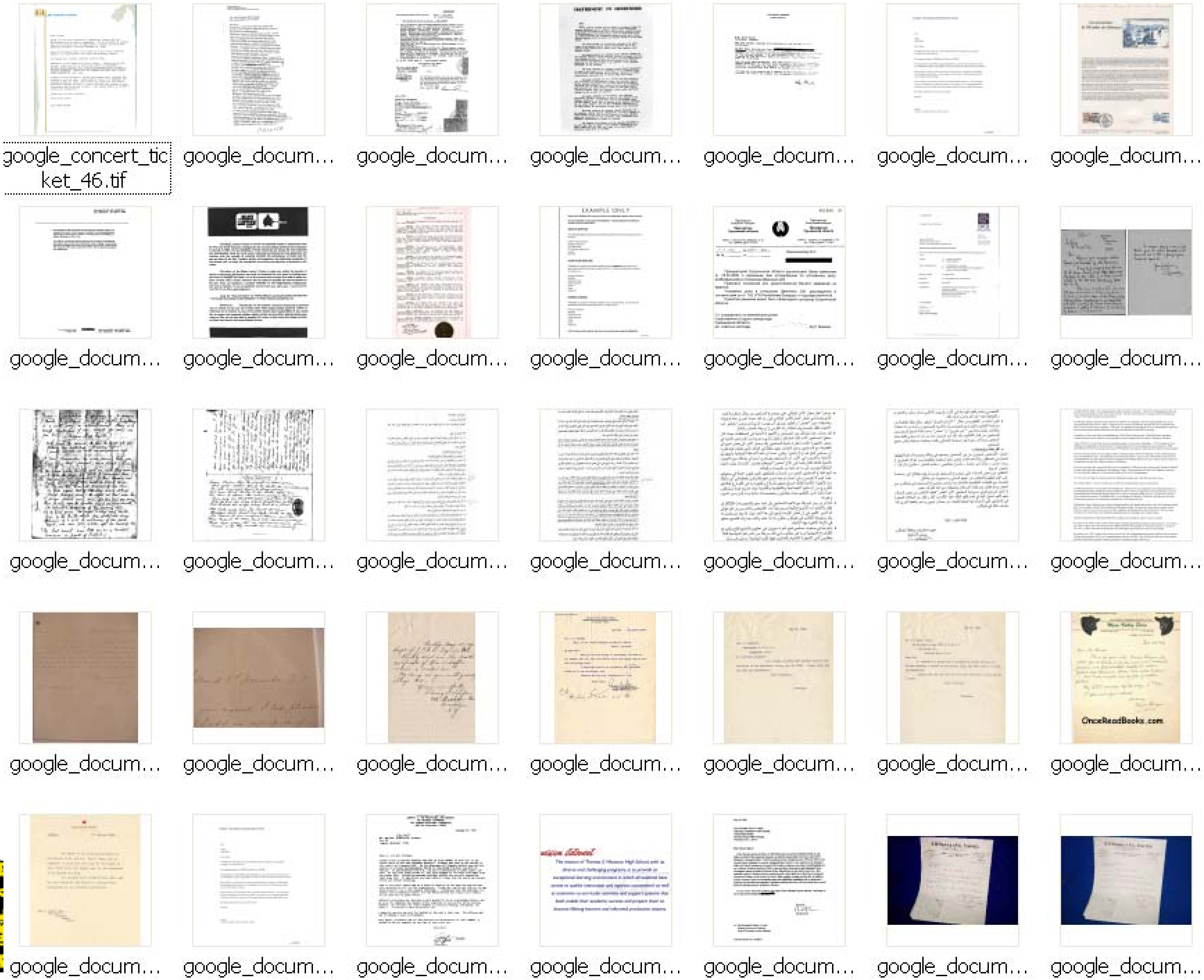
The Challenge

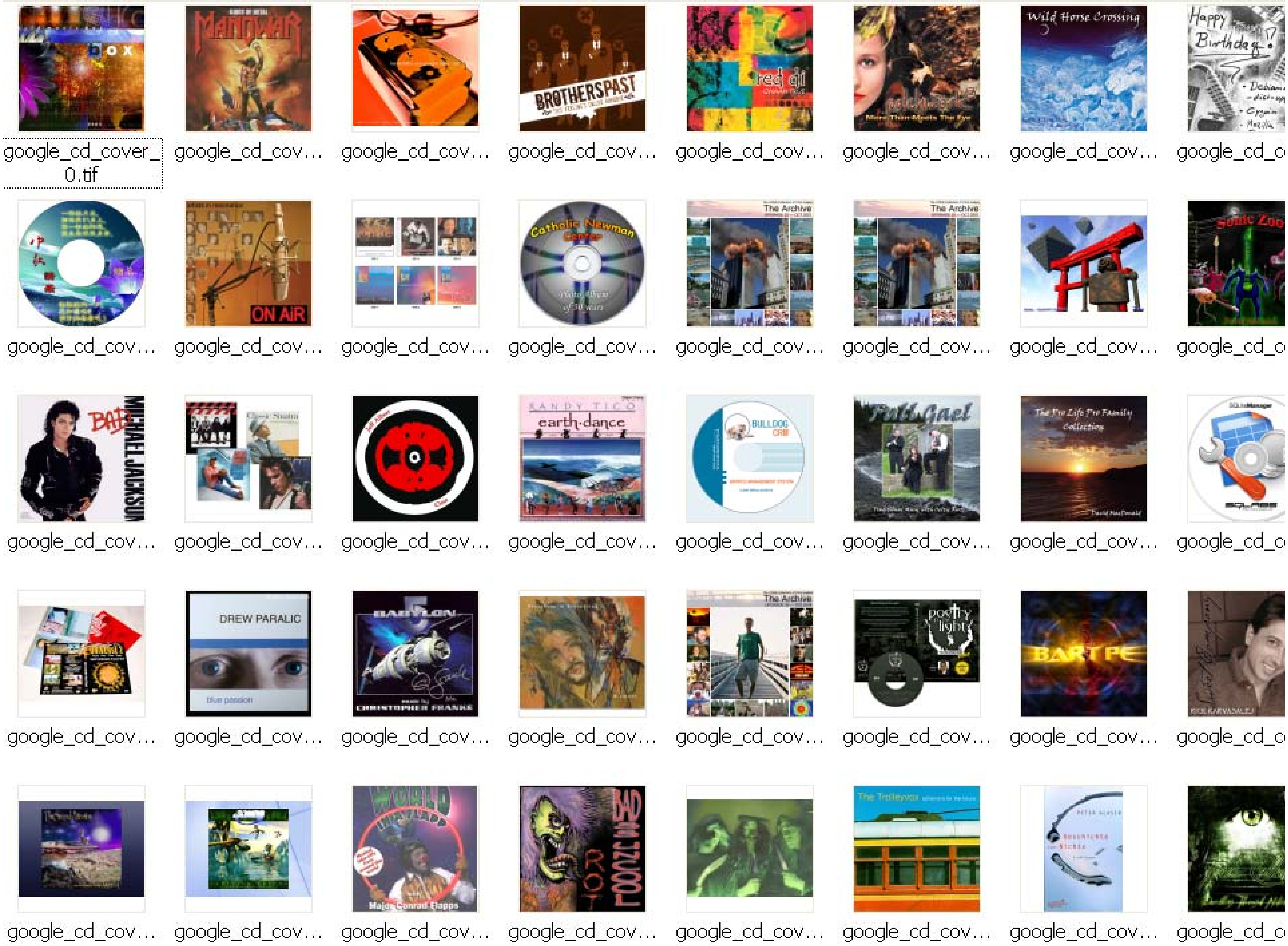
- Images are arbitrary
  - – Appearance model cannot be used for the classification.
  - – We use the same shape descriptor because the code book is generic.
- Ambiguity
  - – “images / text vs images”, e.g., Coke can.
  - – “doc vs images / text”, e.g. “publication cover” usually has figures.

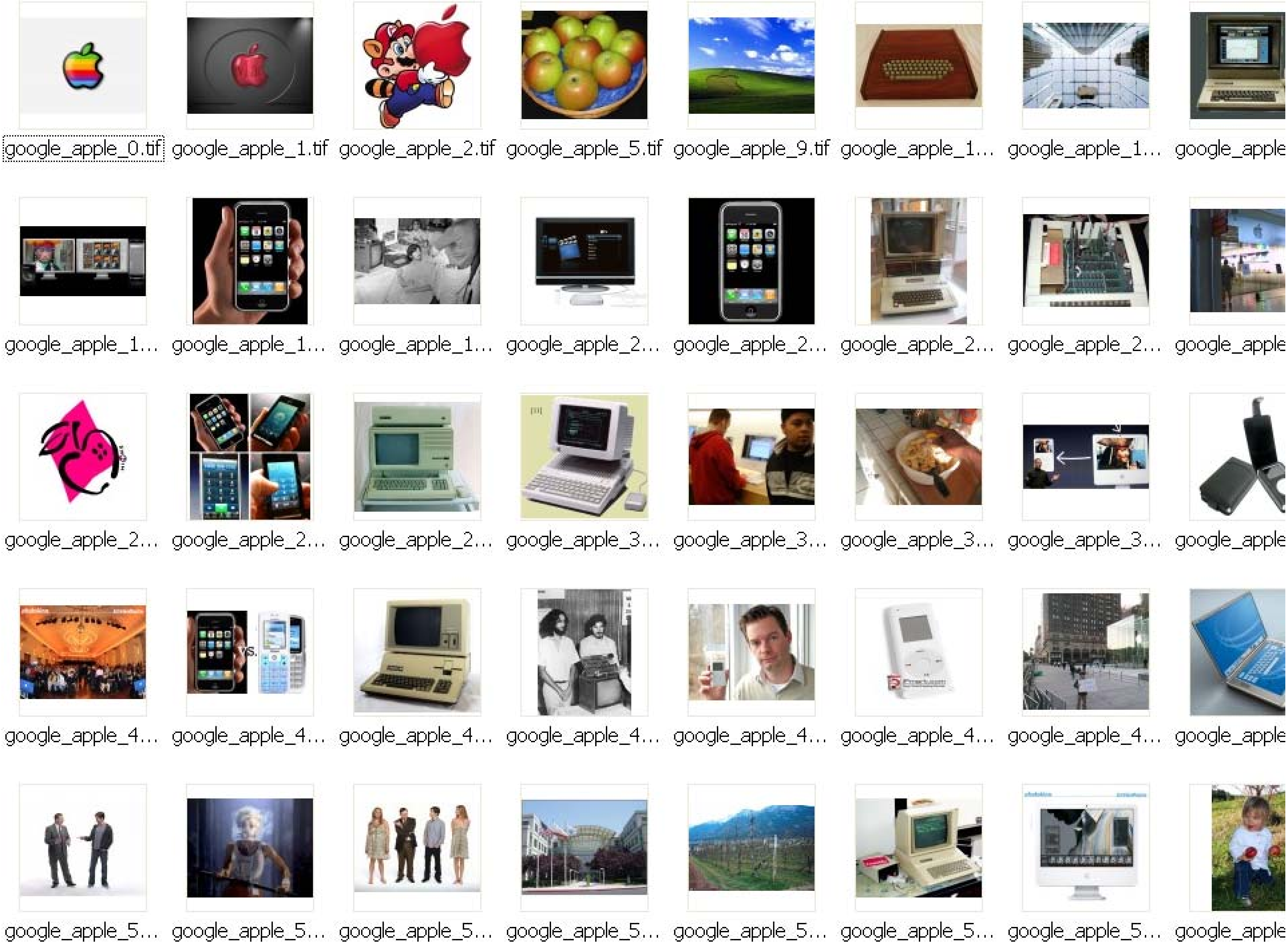
Dataset for ImageID

- 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…
OnceReadBooks.com
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…
google_docum…


BOX
MANOWAR
BROTHERSPAST
red qi
More Than Meets The Eye
Wild Horse Crossing
Happy Birthday
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
ON AiR
Catholic Newman Center
Photo Album of 50 years
The Archive
The Archive
Sonic Zoo
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_c
BAD
MICHAEL JACKSON
Classic Sinatra
RANDY TICO
earth·dance
BULLDOG CRM
Full Gael
The Pro Life Pro Family Collection
SQLManager
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_c
DREW PARALIC
blue passion
BABYLON 5
CHRISTOPHER FRANKE
The Archive
poetry light
BART PE
RICK KARVAJALEJ
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_c
WORLD
Majds Conrad Flapps
BAD SCHOOL
ROT
The Trolleyvox
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_cov...
google_cd_c

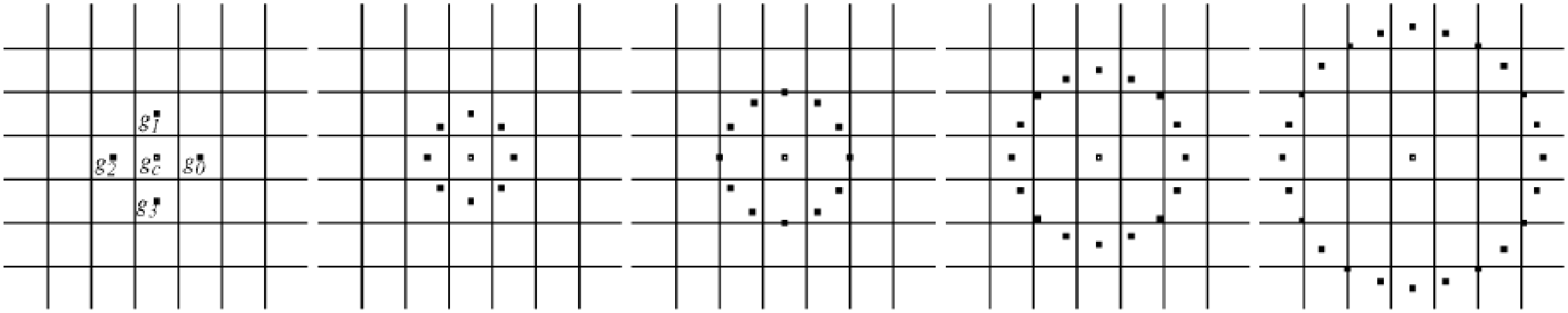

google_apple_0.tif
google_apple_1.tif
google_apple_2.tif
google_apple_5.tif
google_apple_9.tif
google_apple_1...
google_apple_1...
google_apple
google_apple_1...
google_apple_1...
google_apple_1...
google_apple_2...
google_apple_2...
google_apple_2...
google_apple_2...
google_apple
google_apple_2...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

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



(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)

The performance

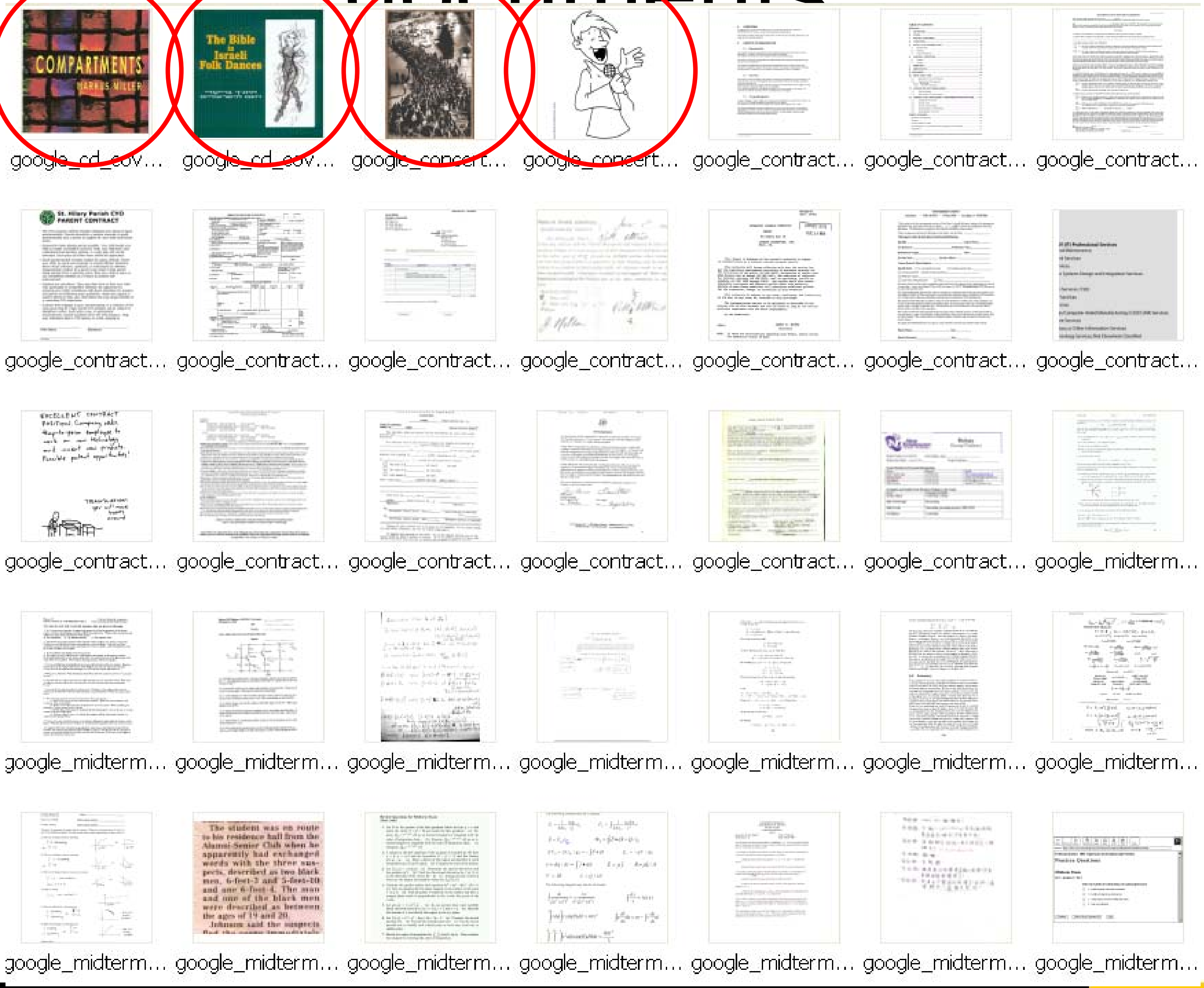
- Confusion matrix

	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

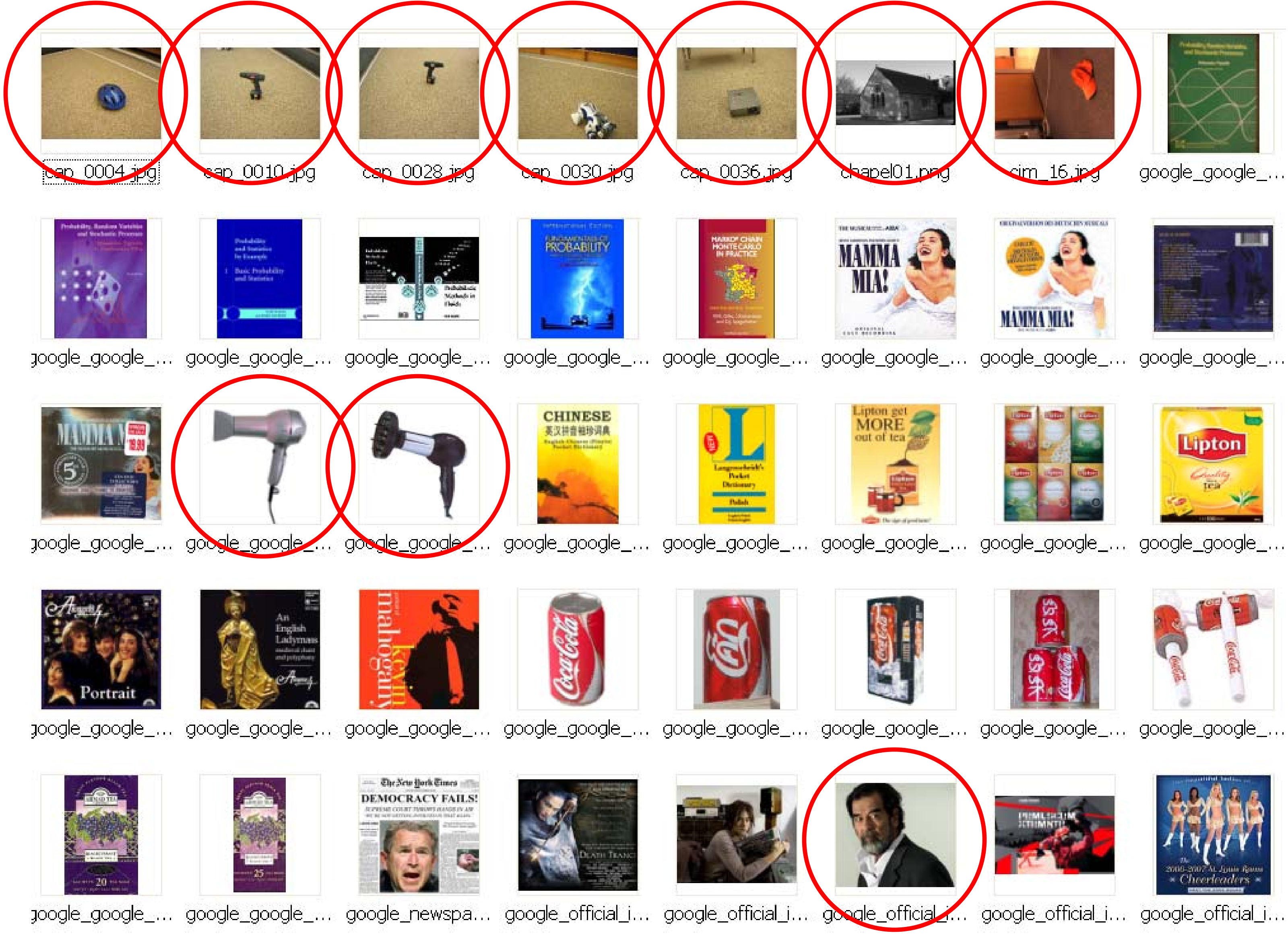
The Module

- 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

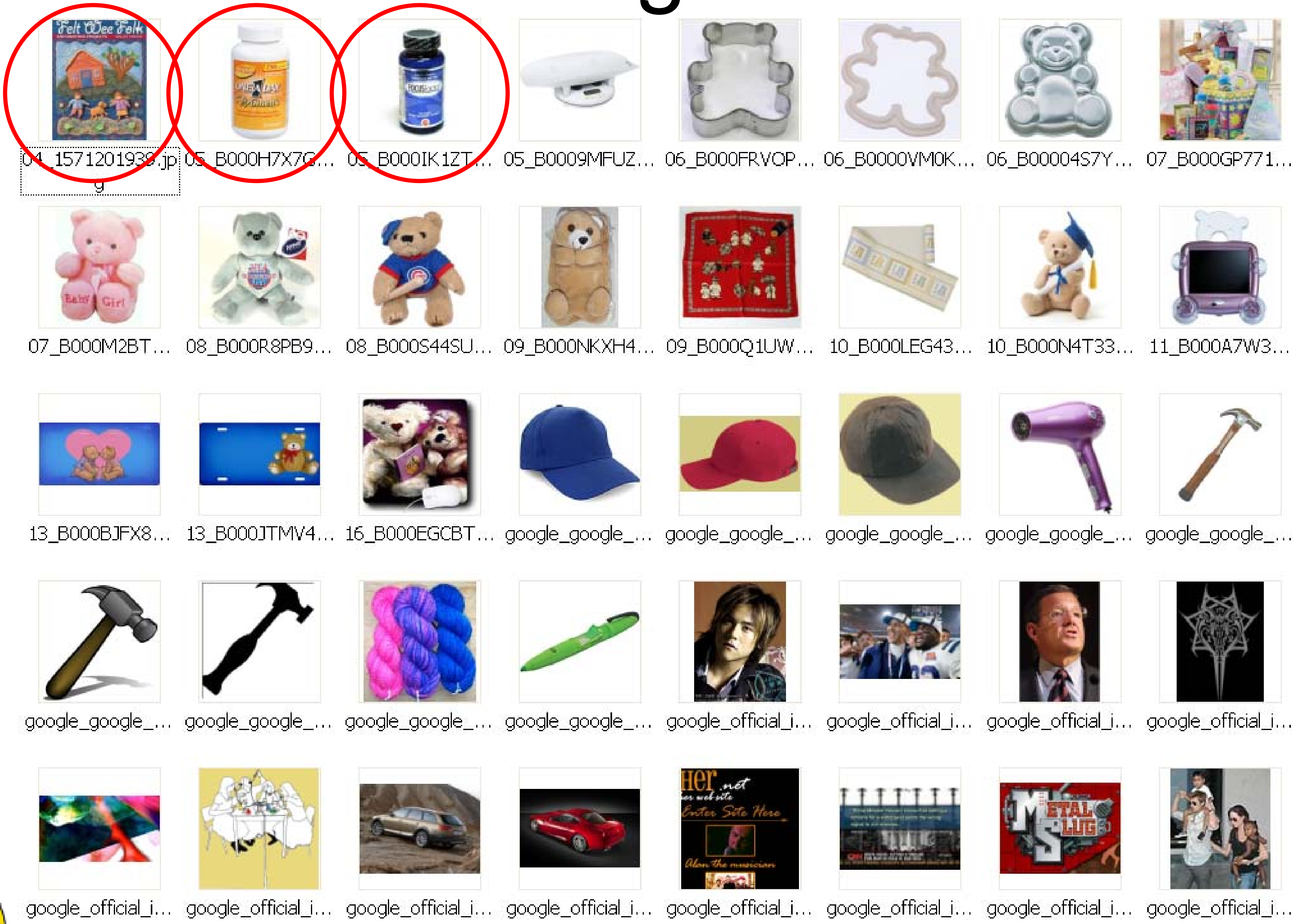
Results – classified as documents



Images w/ text



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…

Technical Presentations

- Page Segmentation (and rule line separation)

- Signature Detection
- Logo Detection and Recognition
- Stamp Detection
- Document ID/Script ID

- Page Layout Similarity

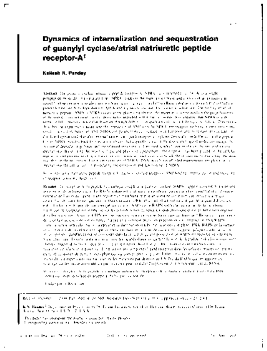
Document Ranking

- Video Research
  - Tracking and Analysis of People
  - Video Content Classification

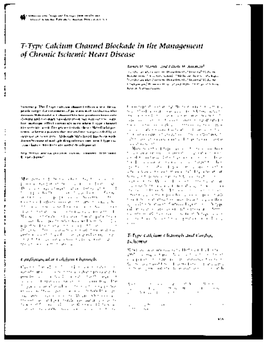
Motivation

- *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.

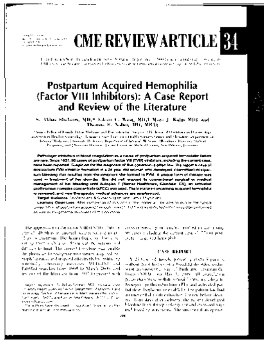
Layout Examples


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

1C

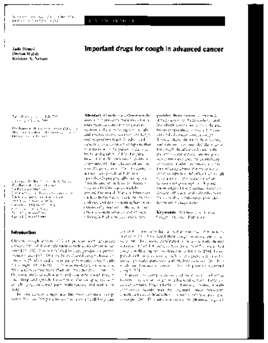

T-Type Calcium Channel Blockade in the Management of Chronic Ischemic Heart Disease

2C

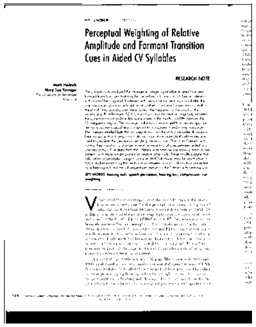

CME REVIEW ARTICLE 34

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

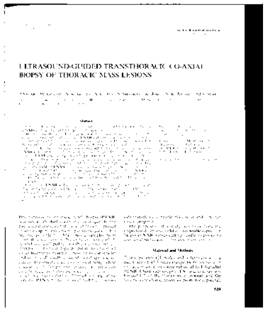
1r2C


Important drugs for cough in advanced cancer

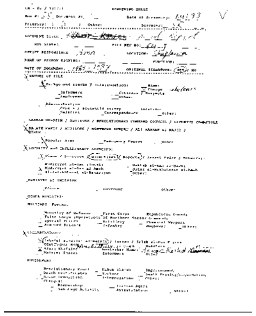
3C

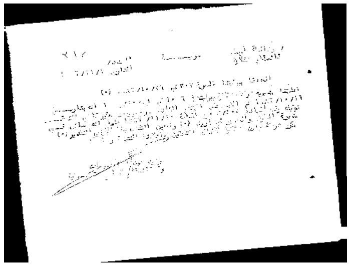

Perceptual Weighting of Relative Amplitude and Formant Transition Cues in Aided CV Syllables

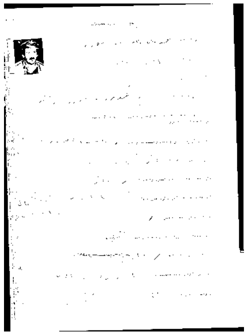
2c_asym

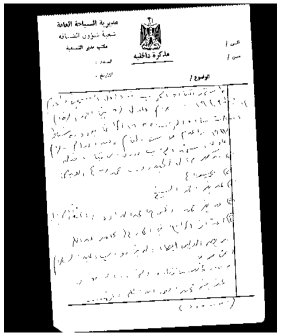

ULTRASOUND-GUIDED TRANSTHORACIC CO-AXIAL BIOPSY OF THORACIC MASS LESIONS

2c2c_asym


class1


class2

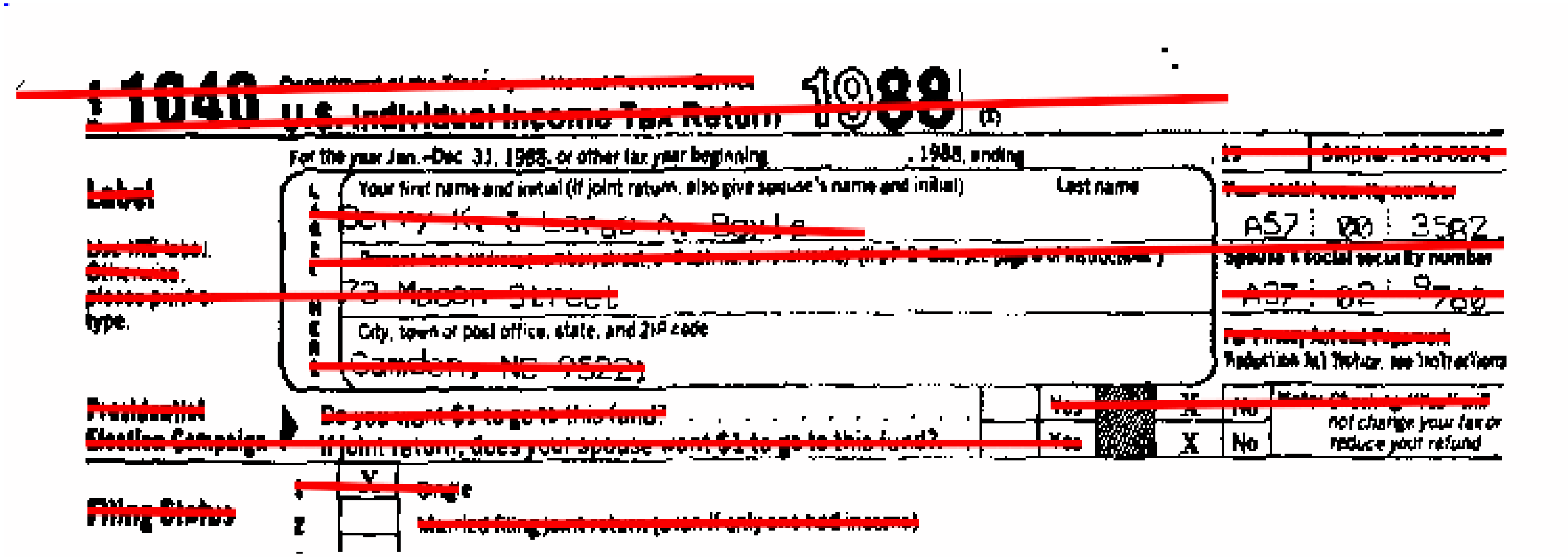

class3


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

A57 00 3582

Spouse's social security number

73 Mason Street

type.

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

Camden, NE 05221

A37 02 9780

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	▨	X	No	not change your tax or
Yes	▨	X	No	reduce your refund

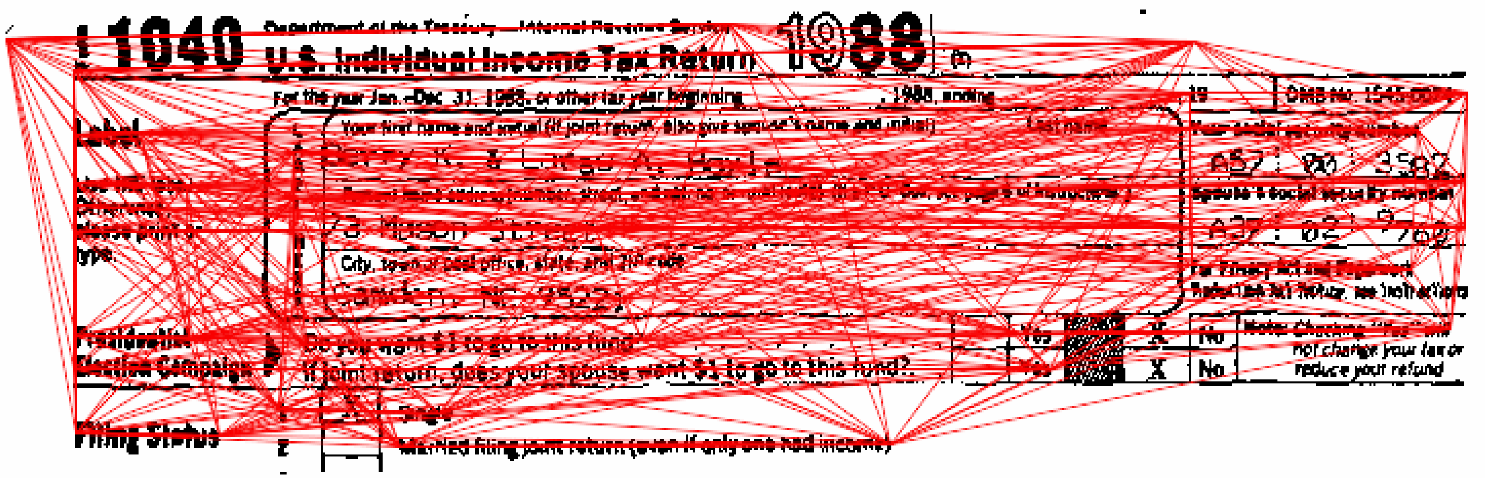
Filing Status

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. 1545-00

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? 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

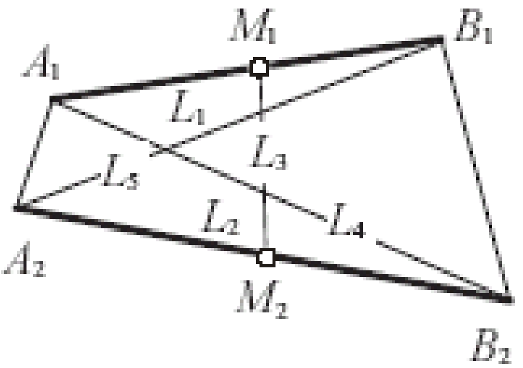
1 Single

2 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

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

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

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

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

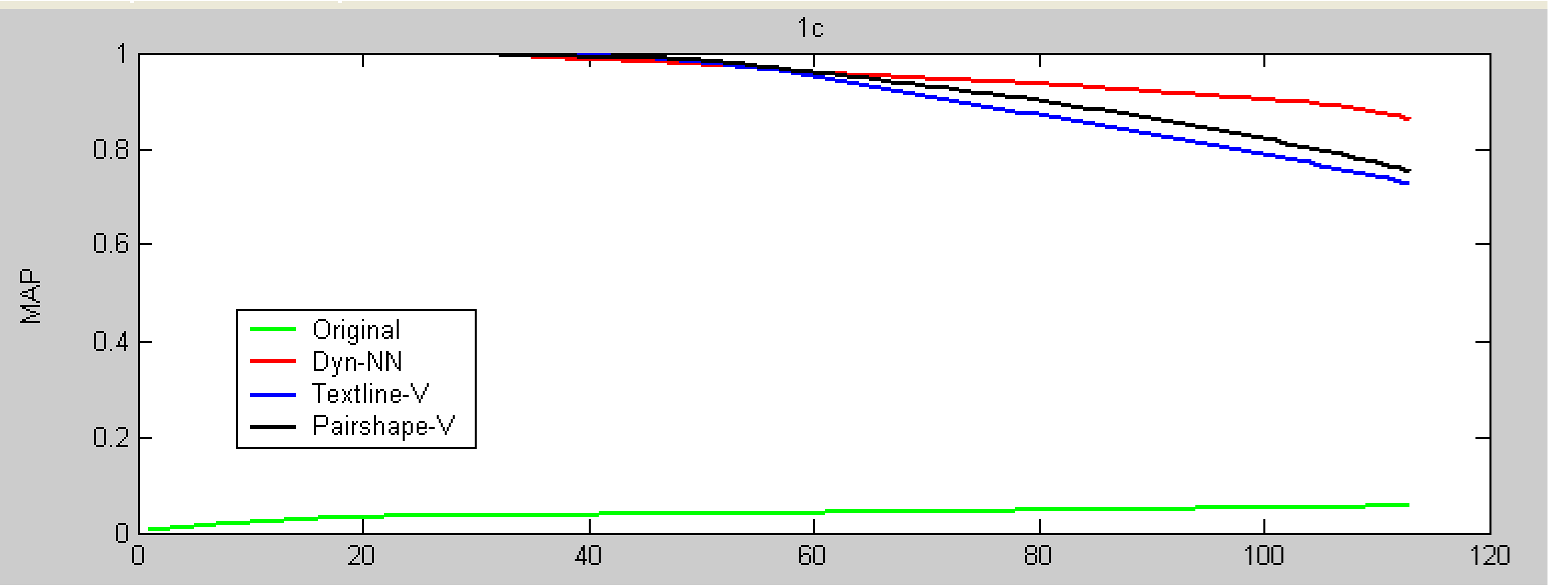
    N_t: wanted testing size
  - $\boldsymbol{I} \in [0, 1-N_t/N)$, the lower the better

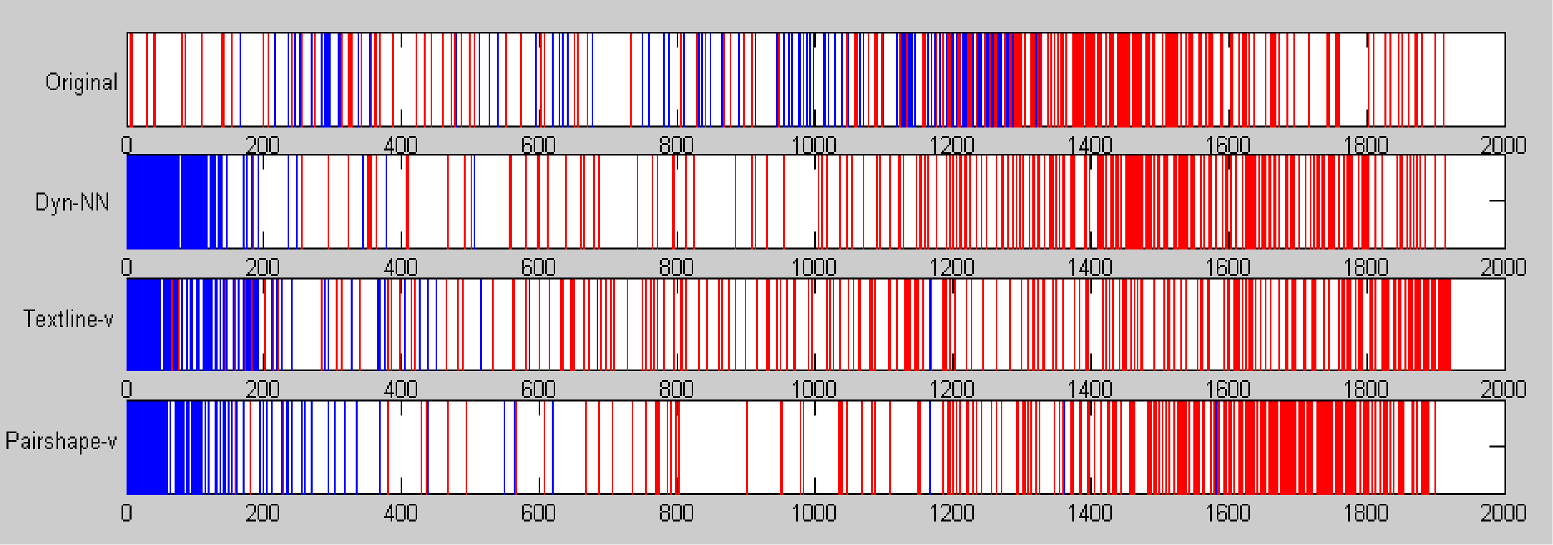
Experimental Results
--Confusion Matrix

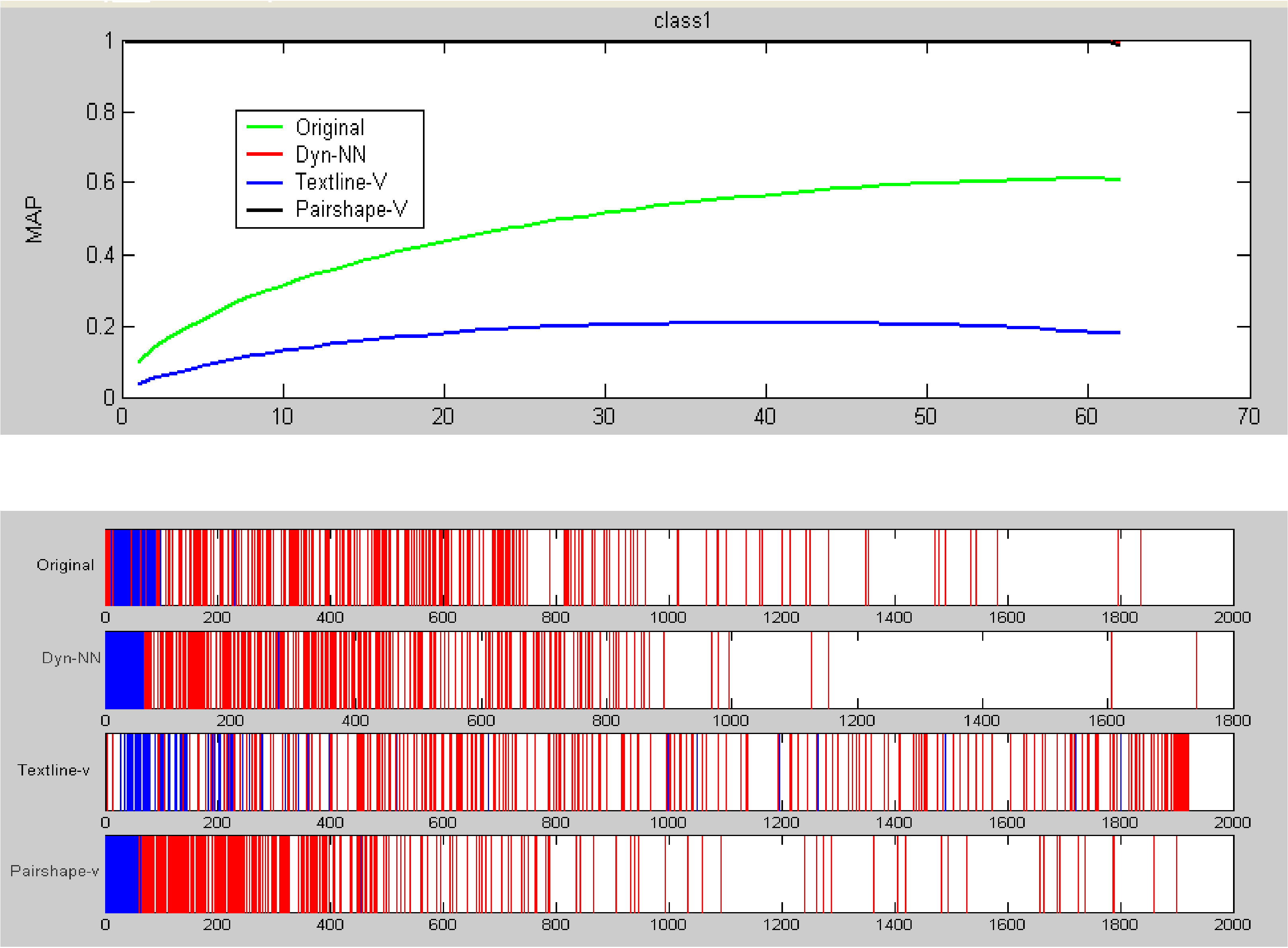
	1c	**2c**	**1r2c**	**3c**	**2c_asym**	**2c2c_asym**	**class 1**	**class 2**	**class3**	**class4**
1c (113)	87	8	16		2					
2c (144)		133	4	1		5	1			
1r2c (431)	9	168	246			8				
3c (23)				23						
2c_asym (6)					3	3				
2c2c_asym (45)		1				44				
Class1 (62)							62			
Class2 (264)	3					2	3	230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

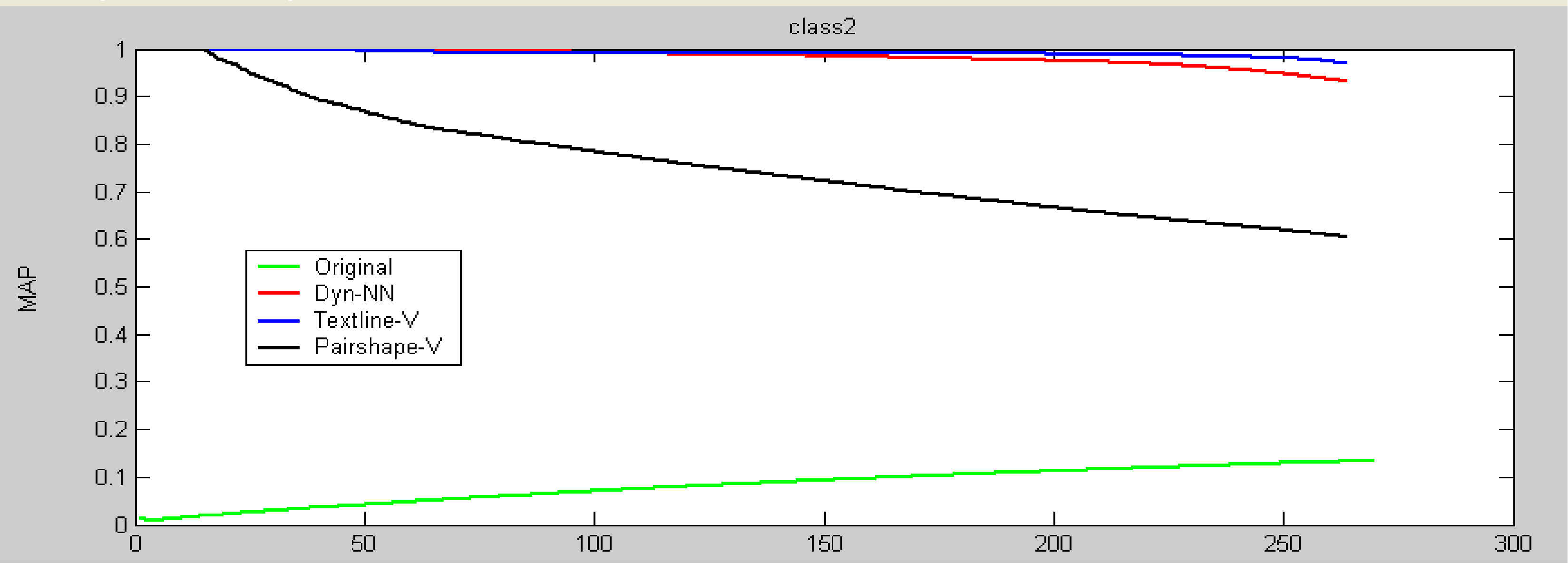
Experiments – ARR Results

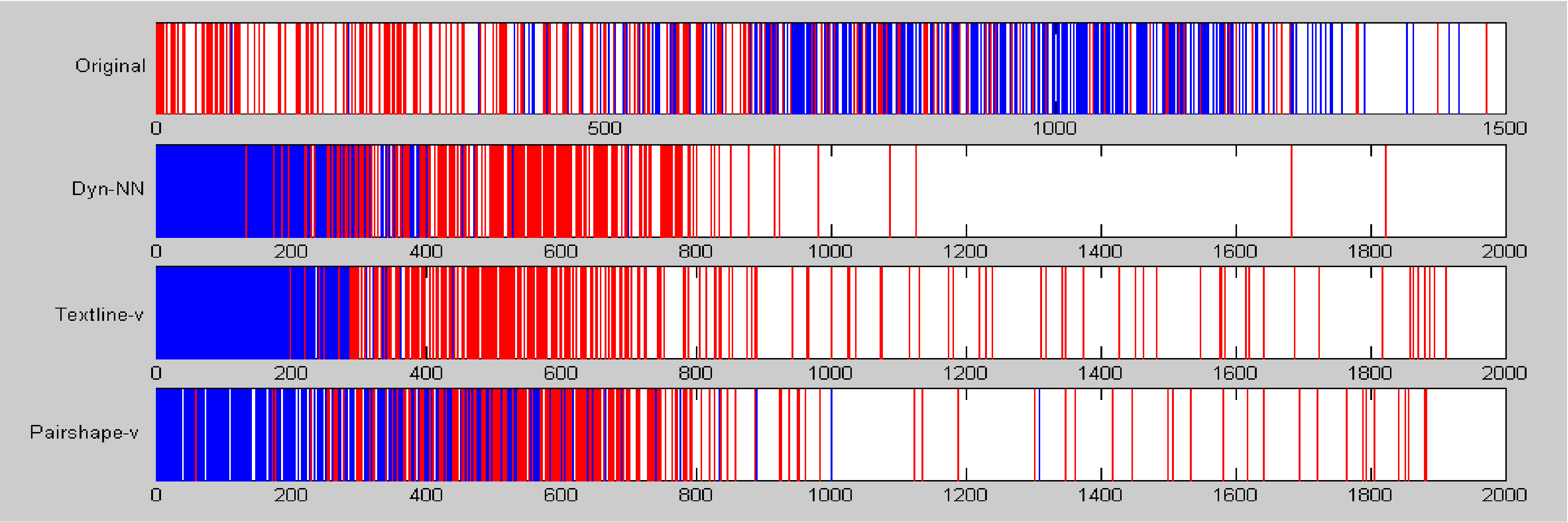
	Original	Dyn-NN	Text-V	Pair_V
1c	0.450	0.011	0.038	0.043
2c	0.062	0.010	0.324	0.087
3c	0.028	0.0002	0.504	0.013
1r2c	0.148	0.063	0.245	0.105
1r1r2c	0.159	0.010	0.103	0.045
1r2c2c	0.121	0.067	0.186	0.139
2c_asym	0.137	0.025	0.360	0.039
2c2c_asym	0.025	0.0002	0.097	0.010
class1	0.009	0.002	0.133	0.003
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

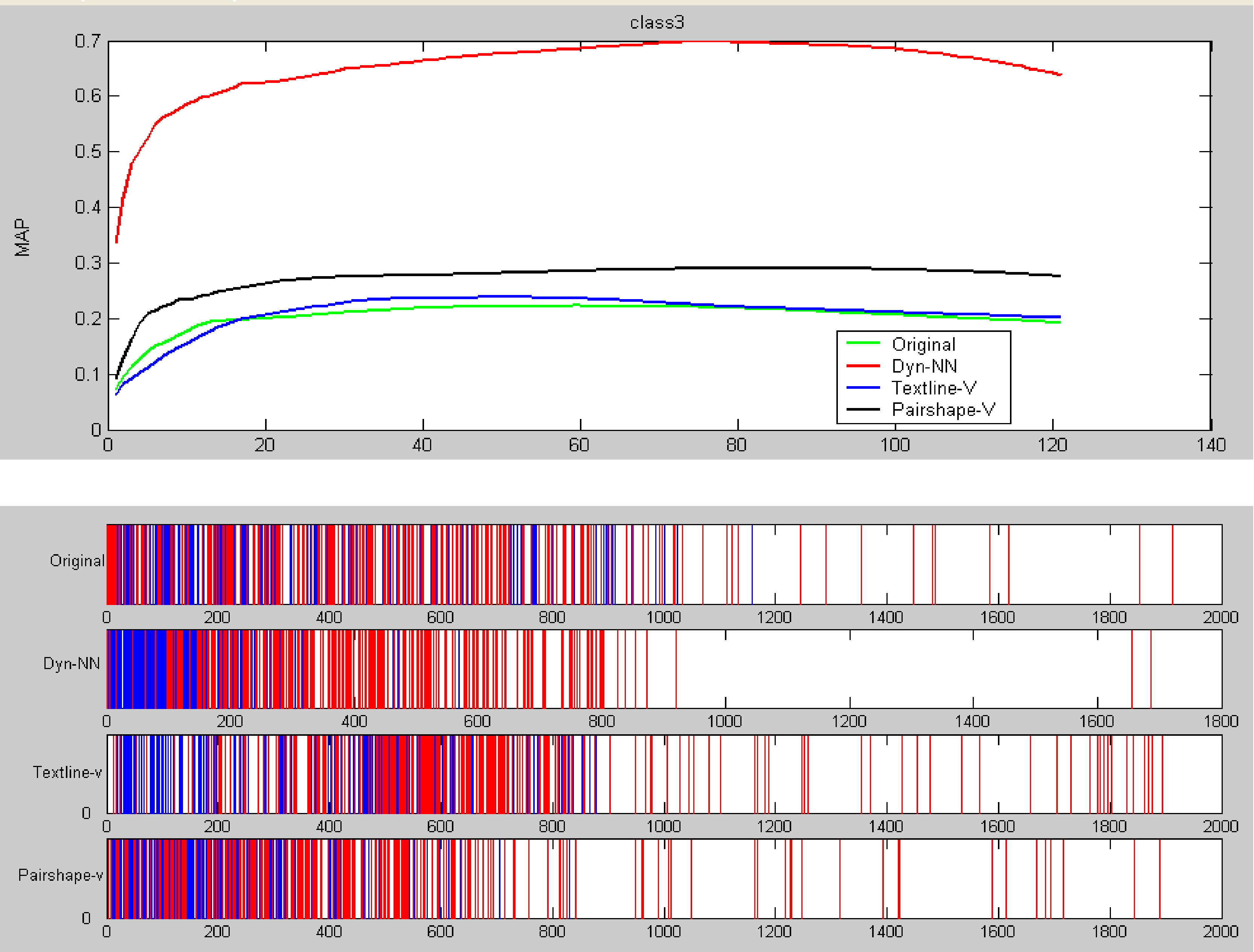

1c
MAP
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120
Original
Dyn-NN
Textline-V
Pairshape-V

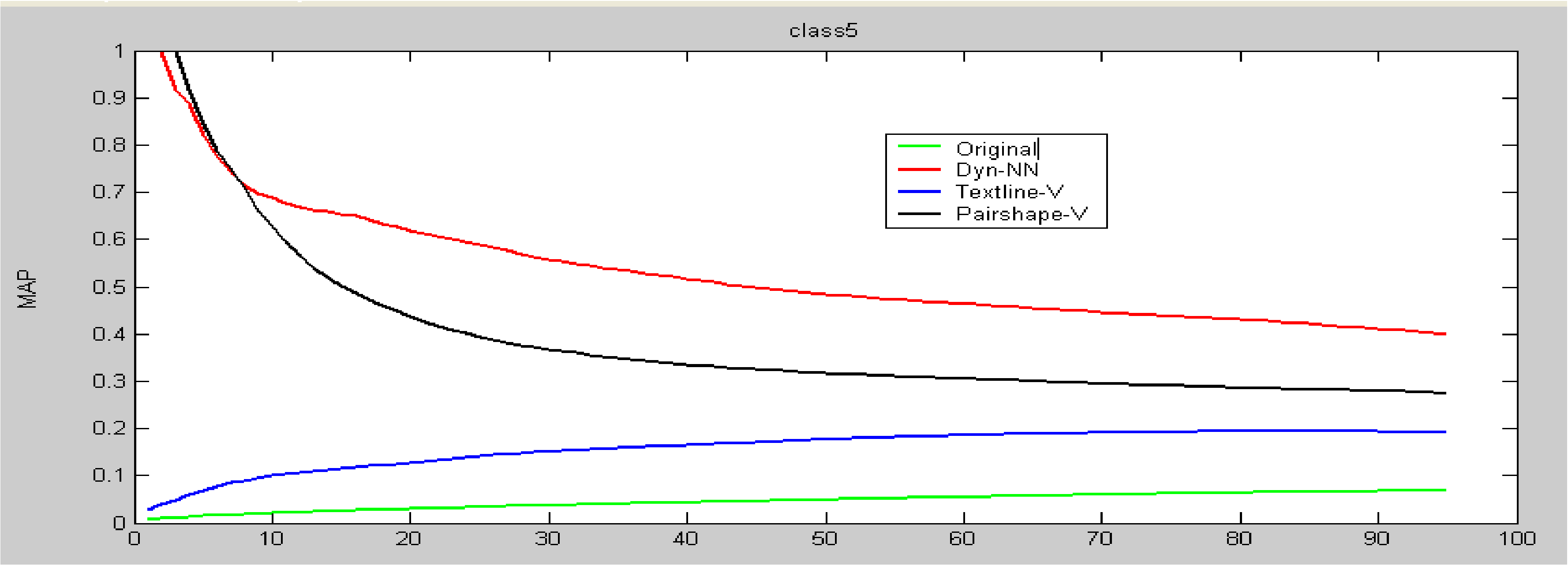

Original
Dyn-NN
Textline-v
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000


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
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
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

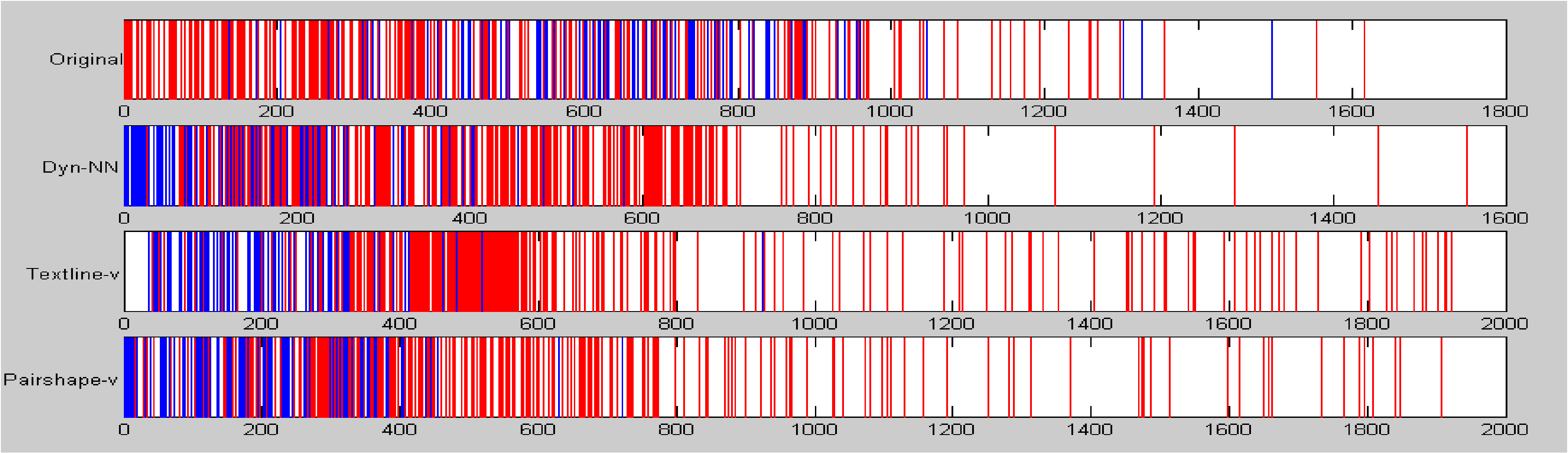
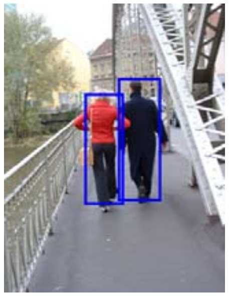
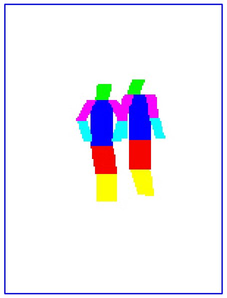
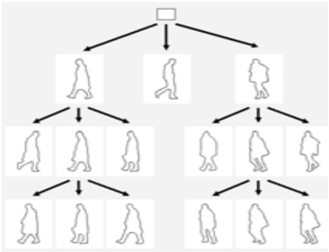

Original
0
200
400
600
800
1000
1200
1400
1600
1800
Dyn-NN
0
200
400
600
800
1000
1200
1400
1600
Textline-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Pairshape-v
0
200
400
600
800
1000
1200
1400
1600
1800
2000

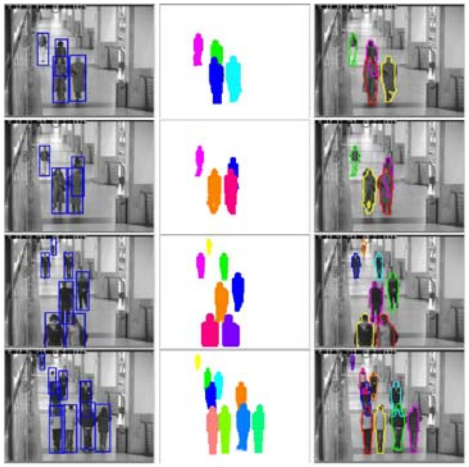
Image and Video Research

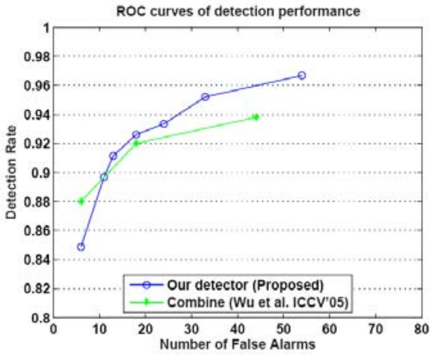






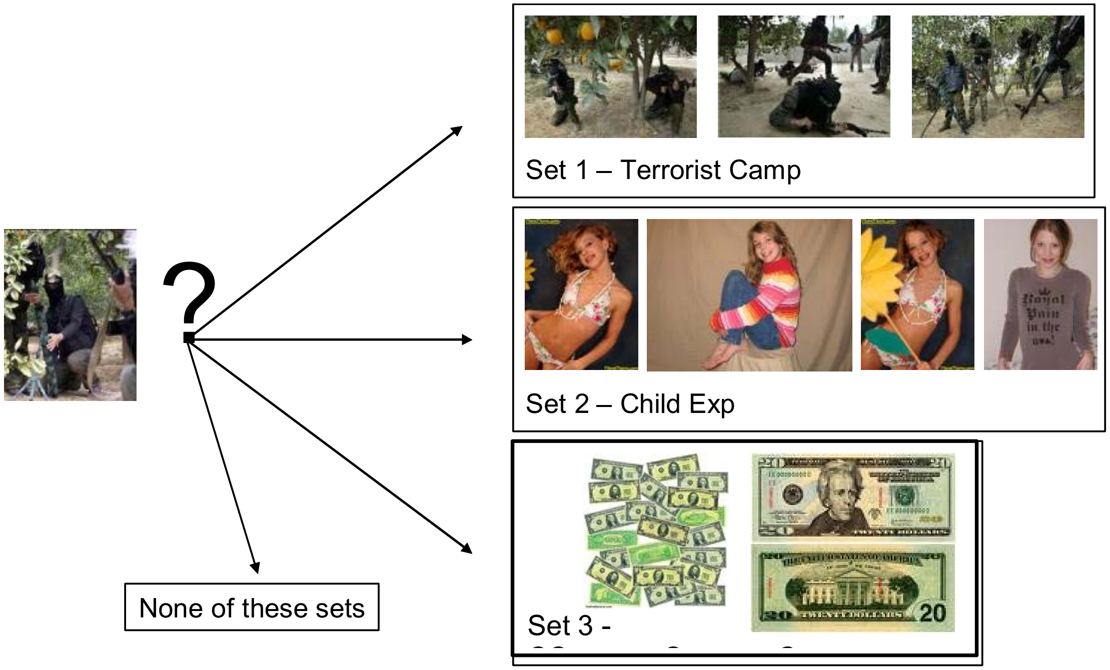
- Surveillance Video
  - People Tracking
  - Appearance Modeling
  - Pose Estimation
- Partial Image Matching
  - Robust to changes in view point
  - Able to match partial images





Forensic Image Search

- Consider a “search pack” which contains a “model” of a set of images of interest
- Hard Drive is searched and produces a report without revealing the search content



?

Set 1 – Terrorist Camp

Set 2 – Child Exp

Set 3 -

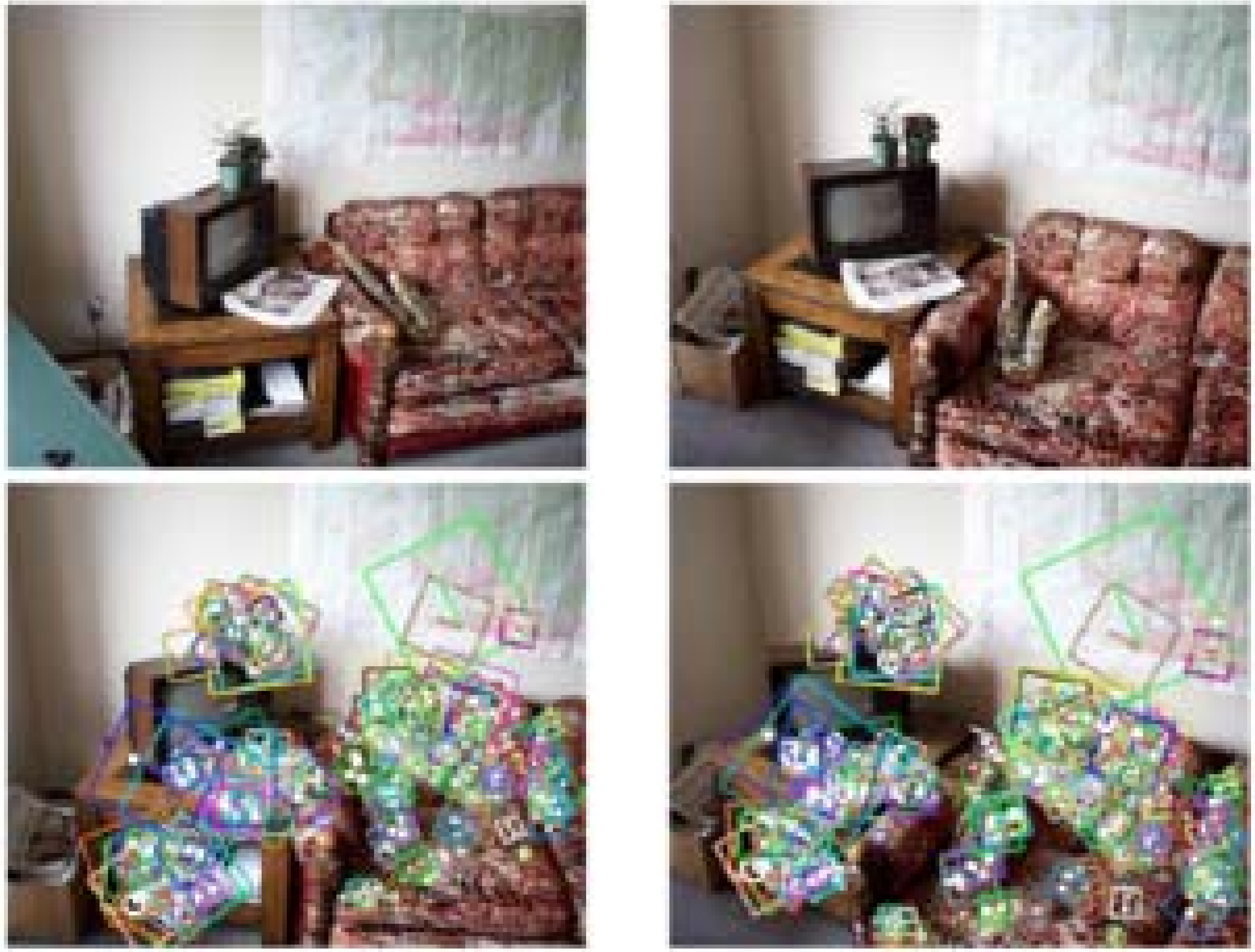
None of these sets

High Speed Image Classification

- The purpose of this project
  - To create a new content based image retrieval (CBIR) algorithm that will remove some of limitations of the state of the art
- The task description
  - A user provides a set of training images belonging to several known categories (called SearchPak) and a set of test images.
  - For a test image, the user wants to know if it is similar to one of the SearchPak categories and otherwise classify it as “non-SearchPak image” or “junk image”.

Using what features?

- Histogram, correlogram of color, edge, texture…?
- A good feature: keypoint
  - A feature based on neighborhood edge histogram that is scale and rotation-invariant
  - Independent of color
  - Approach is called SIFT (*Scale Invariant Feature Transform*)
  - Captures salient visual information
- Groups of keypoints are powerful description of objects in images and video



What Can we do with Keypoints?

- Searching (Video Google, Zisserman Oxford, K-means clustering)
- Mining (find most significant objects)
- Indexing (find anchor and cluster frames)
- Browsing
- Logo search
- Near-duplicate detection
- Face detection
- Building detection

Summary

- Focused on Integration with DocLib framework
- Need Software engineering support
- Detailed evaluation and evaluation tools as part of Prototypes.

Possible Research Extentions

- Increasing the speed of processing (software or hardware)
- Script independent word spotting
- Stamp and signature recognition
- Scene text recognition and super-resolution in video.
- Word level Script and Language ID