# **CLEAT:**

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



# LAMP History

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



## Mission

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





# Media Personnel

#### **RESEARCH FACULTY**

**DOCUMENTS** 

#### <u>VIDEO</u>

Stefan Jaeger

David Doermann

Daniel DeMenthon\*\*

#### **Graduate Students**

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

Undergraduate Researchers

Mike Roth\*

Faculty Research Associate

Tandeep Sidhu



## Affiliates

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



## **Research Focal Areas**

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



# Outreach

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



# Agenda

**Project Overview** 

- Introduction
- Goals and Objectives

#### **Data Collection and Ground Truth**

**GEDI and Evaluation Framework** 

#### **Evaluation and Research Components**

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

#### **Technical Presentations**

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



## **Project Overview**



# **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
Laver Separation	90% coverage at the pixel level
Segmentation (Print and Hand)	85% using implementation of existing methods
Logo and Stomp Detection	75% provision at 95% recall
Logo and Stamp Detection	75% precision at 65% recail
Signature Detection	75% precision at 85% recall
THE PARTY OF THE P	



#### • Phase 1 - March 31, 2007

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



### Data Collection and Evaluation

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

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



# **Document Image Acquisition**

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



### **Document Genres**

Genre	
Forms, Drawing, Tables et at.	
Forms	650
Drawing	80
Tables	100
Chemistry formulae	25
Math equations Figures Total	165 40 1060

Business documents and Memo letters
Business documents
Business documents degraded
Business documents with annotations
Memo letters

Journal and Conference Papers, Articles
English
German
Japanese
Total

	Newsletters and Flyers	
	Google images	2400
650		
80		
100	Structured Documents	
25	Phonebook	229
	Dictionaries (Chinese English, English	
165	Chinese)	1150
40	Yellowpage	80
1060	Total	1459
	Structured Documents	
50	Phonebook	229
	Dictionaries (Chinese English, English	
2700	Chinese)	1150
160	Yellowpage	80
900	Total	1459
3810		
	Handwritten	
	Chinese	146





Total



	1.			
C	000,000	MARIT	ANYENTISING THUTCH	DATE
	MAIL BOX (GENERIC PRICE)	ARRAFIAS	SENSTAPER.	10/38/89
	BEACH (GENERIC PRICE)	AUXABLE	SUNDAY SUPPLEMENT	11/5/89
	EAGLE	INDIANA.	SUNDAY SUPPLEMENT	33/5/09
	(NEW GENERIC PRICE)	ENDIANA	NEWSPAPER	10/30/#5
	ENGLE (NEW LOW PRICE)	DISTANA RANKERS	OCH .	12/89-1/90
	RED FACES	ARRANSAS MARKITS	008	1/99
	BEACH	PORLAND, CREGOWIAN	SUNDAY SUPPLEMENT	11/8/89
	MAIL BOX	PORTLAND &	NEWSPAPER	10/30/89
	BEACH	GRESON, EUGENE REGISTER GIAND	RONDAY ROPPLENDING	33/38.89
	GIRL GIV MEAD-ON	ARRANGAS NARRESS	008	1/89-12/89
	SIRL GUY MEAD-ON REG. & LIGHTS	OREGON MARKETS	SUNDAY SUPPLISHING	1/89-12/88
	"CORVETTE"	PORTLAND, OREGON	SUNDAY SUPPLEMENT	11/26/89
	*CORVETTE*	EDGENE, OREGON	SUNDAY SUPPLEMENT	11/26/89
	*CORVETTE*	ANNAUSAS MAANTES	SUNDAT SUPPLICENT	11/14/89
	ARTCH - FON BAICH.	AROARA	XENGPAPER	11/20
	*BEACH*	OREGON MARKETS	RENEFATER	11/10
	C.7./80 October 18, 1989			1,044254





			ADAMS
Collable, Martan F	Additional Internet A	AMAANA Jaar B. Alaine Suray Boker (Josef & Ceripal Propers	Automatic Automatic P
TRIBUTO JOLAN	AP-Doll Million of the forward with	photo: Bash und and	Alast hants in
Charles Public Public & Leadershy	at the set of a large	Administration of the second s	Standard Cogan Salasations and 20-ann Mademichant unitable
Anitals Research Australiant Languerray	time day 7 the & Despate Ray	ren Burniturniais	ALASS, Decel P.
mentals from unit als	multiple areas	AMANDON, LANS 2 OFFICER, STR. Construct For Administry Program.	210 Basseraria rati
Addition, pages A	Aller's day 1 Process	Programmer Contractingtion representation and and and and	41488, Decent /
20 August Antenine Physicand ada	24-09-0 mentionenal une ans	Advention, Barbaro & Annue Marchine Version Print Derman	Advances (*10 Wood Bin, (Propin
Bertrager / Garrier Party	Address of the local division of the local d	terra baj 27-stat	bert trut acts
Pagaria in design Padada Basmal anciato	Land Arch Friderica Partnerson man (SP-sett) will thursd unit also	Call And Appl. Constraints & Public Static Testings (upl. 200-1216) Indept. Static appl.	Mark Science for Sale Devices
Research Association of the second	10000, boarsed 1	49-1-140, fame	1045. mail
print management and they had give they an (or Round and with	Freed Same Region Change Services Administration Contempor Contended Contend 24 doct	Residue North Institut Plans Research 1211 Diving, Research Facility, 2P (2011	man chair fan Right Galait man Britt Campon Reconstro Cartes malformit genes cart
Respond Patter Generation & Pattern	sola d'Angurcak	www.paureau	ADMILIARIAN L DOCHEM
col-dian	Per Arts Of the Sprane Assessor	Dear Adm Agent College stude Televisies	Tax Municipy Ingl Sadara & Humah June July
ABSET, Parcela M. SOLID Statistical College & Australian	ARTS. Duart	Intelligence and only	ADAMS, including 2
TO Available Bay 24-7111 Transmitterations	United Intelligence Associate VI.	Fairl Ser. Aller Throughermon Drug Therein Anno Drugs and Mitchell	Dictory 21(61,an desp
American fait for Inia Grants	Table Ton Table and address of the Table T	Arme, ideal.	Allest investor 7
Grants Dise Big 3P 1101	AND ADD DOG TO ADD TO ADD	Regards Stree Testing Strengt ST-2711	Nation Respire Balantation Mathematics Res. 27 (01) 1007007-001008
And Annual Annua	and and and also	ACEDO GARCIA, Inc	ADDEL Johns D Mr.
Services 1778 Carlors Reconting Carlor	AMULA Mindum 1 Martin Program	and dealerships	AND Mahamatra Rep. 74 stort.
2P (Kis) NT®unaturitikk	Fill Parent both April 29 Service Public Review (Marcello, Parente, Marcello, Par	ADDVEDG-OFFIE Arbane Minte	(KIG-40-014) (Although und who
ABOW (NORSON, Rosarian W 1972) (Institute 1: Mathematica	which Baserial and all	ADD A LM Palance (kep	ADDED, AND D
the second second second	AMUL, Auro 1 AMU		NUMB COMPLEX
COLLARD CHIM	Martin and Tank and and	Nei Datr Auri i ne Makina Propen 1992 Aurus Gudeley neamag	Martin Salahar Palajahan Martin Salahar Jawaiga
AMOTT Brackers L.	Annual Computing Computing	Constant Spec Barne's Office	\$2485, Sect
hibrill-shlab-	1102 converdence (Max Burst 24-601). Mark Burstrag unter and	tohe-Rashal uncali-	Pegers (mail family beings (Pr. 27) (NT) 508-4764
Agent (Sen Extension Device Carval	695.5 Elect 5	Anna Par English	talans? Paunal unstals
for the state	PTT Notice Lines (Pres)	the Ch. Broot No. Washington DC	Cost Ser. / Pactores Late
Reaching and the second	HR 43543, Warnly Re One Not 7 the System Research	Minister, Autor & State Public Parcillor Parcillances	HANNING COLOR OF COLOR OF COLOR
10 20141 (011 494 (012)	and it. Mirchell, Belancing 7, 1942	tartent and all	THE OWNERS ADDRESS TO
hananik parantaka	fee drat tear if Mathematic	ACCERTAINANT, Ramon L. Dogt Assey Housey	Marthuast #
ABOR, HAPES, Hungal & MARP Societies (14) No. 5 (Springer Springer)	Engeneric Area 24 Alth. Harbert Color an.	Frank both top that the Breat and all	Adventions for the features for the second s
CONTRACTOR AND A CONTRACTOR CONTRACTOR AND A CONTRACTOR VELOCITIES	Annual Annual Coll & Balance Annual Annual Coll & Balance All I Taximum Read (Mr. 2021	Martin Add / Counseling Triff Bostonians Ring (Prof. 1)	27-273 REFORMACIÓN
		Character and Generation	g Josiment Vijes-Nake und sokolaase





Neuroscie (1992) - 1992 (1992)

### Internet Downloads

Genre	
Figure	
Good	240
Medium	755
Low	548
Form	
Good	66
Medium	69
Low	32
Letter-Memo	
Good	55
Medium	88
Low	31
LIST	
Good	6
Medium	34
Low	11

Newspaper	
Good Medium Low	22 37 17
Publication Cover	
Good Medium Low	130 425 128
Receipt	
Good Medium Low	10 50 20
Screenshot	
Good Medium Low	184 848 566
Table	
Good Medium	52 124
Low	42



# Maryland Datasets

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



1921541 9119 RINEIN 0 DIGDINS after para and I am 220 ty - Ety Date 2109 21/2-11 Ø Tamaiy 14112415 CHIROLD Ablue Sta 22229 21 wit 2122 Hora Zia 370 हैर हुटा मिले ater 192040 3 myeller Ela villa a LACESTEL 2 Elans 2 Statue Sica Et & EUR Sina and inter ty THE THERE 2119240 212501 522 (232 23.29 201211 502 20662 - 5 201 2001 CMPC/201 Elch 210 2102 242ell 2420 Tople IS 210219 (5)to zerbin 2994 Z CUS IS INT Port BY TE STOL 271127 19+71ET 220 alan R 3214

以戊治周的遗感

李周基

超紧伤不幸逝世),恶风待来须人感到兽叱的悲痛;他两中圈的 从年間成小社名、经济35段食作435、可禁减、35正大贡献、因为人的渡 他,赞扬他,爱戴他:然句由於"六四"事件众把丧了人身狗肉,扫导受 长的月的高了3,但转他建足没有介一个国委制起来就的淬使,处地挖展 终9,支持x逶腾3;

代互小簿超紧伤的"结决"北受的河:就死韩九的爆炮方式而言, 也无秘与编说的,意情,以凡法国的思想。我和赵马建落的足,尤先有 的教學人,一方南大溝精溝以代光週(書些也很大要),而另一方南却又開發 以九海圈,表:浅踏吹九海圈。车来,趋紧得天重似圈家的重要领案人, 豪运议察想很,高超重的魏吉光,达面,未经任何凡佛教许多则凡亲 彩,隐影烟之人的我自治,就把超紧厉长别教禁起来,剥夺他的人 身角は後はしょう、远離道で野以氏地風的明和湖南、花林、传 117



ł,

165 V ม่อเมริก 1 5 FALLOUN 2521 אים געוואי איי 72 F งรืององสามามาที่แก่อยูกแมะที่จะผมภาพรีเมล ผู้อย่างการมีขาง กระกำมากัด งรักงอานและอย่างไปเองกางการการนี้เจ รลามอาทากรุแอละแหลอยน อโอลเน้นกล่าเ งรักงอานและอย่างไปเองกางการการการการการการการการนี้การการการการการการการการ E מהרדיו שעשייו לשיראות בת היהיו של בסט שעוטייו הבר אבים על ג אותיה. עשיו אות אות הב ארומייה ב העשייה שריב בעכיב שיוער כי שיע ש עשייו יליד האות Or שבעורה שלאו מיצה האואמטו האריל בהת שה שועלו הצושיר עום שלה שירושר הלים Ġ 6 6 520501 שה נותו מהאל הייטע ואשום. EL ETTO the man We have a we alto 6.91 Frence EE, 3 facing 0 10 15 22(2)



85 forment	n. 先亡. そ近、 25×123 そこ
USUFD	v. 75253624
Incéssant	A. 2002 at
shrink	V. 2211=2+, 3-1=2+, 2/7+ 52+ (p.p. shrunken)
rashly	Nd. 924111. 101264111, 101264111
mássacre	U. \$5x23126
flee	V. Horutzt. 22/11/24 (p. fled)
tribe	m. 72. 35
abandon	V. 412124, 注11644
nomádic	A. 贵星纪二1. 开星从18轮云 和气
* Similize	V. うきもんにひ、うきろうこと、せきひ
chualry	n. 716824, 710+424
• refugie	m. 051324, Ilt24. 42
perpetual	a. 5746, 4764
tormoil	m. 圭唑,害亚 , 苄唑
87 undermind	V. ゆちんしけれた. Marth1 れをしんしみにた
* monastery	m. 7222 Att Att (A)
vast	A. 282466. Hunde . Wille
* proportion	M. ヨフノ. リレコノ. リーマ
realm	m. \$\$\$ . \$\$
/relm/	
bureaucrocy.	m. ことをおり、ことをやらえり、ことをうのう
+ exert	V. (ちし、うはうちこ) からんしかん 、(のかこ) ひろにち
< Taor	sm > EZ
Taoism	m. 22
• * calligraphy	m、 人10月1 、 人1毛 、 さきをん 、 35名から
conglomerati	on 7. 52 262

И. Брозский. Ð Trog berep ou beguy .. Год вечер он видит, заклювини в дверени ува веадника скачут в окрестноск полиях, как будо по кругу, Сквозь рону и гото, и досто не спогут друп друпа донар. To Spocul notopous, nounneys, yoyal, To encla l'ague boz dipagenero nue crab, и бысто по светения склону эсоний, В катоом из нас то в рону опер, еде случается Тыка Два всодника скогу в вегерией умени, не ролько от долега, от сердиза вотици, Jun guna One Occurrano, 30645, Borobo Gaves have herecune pape za payy numby? U jan unnorga un na cheje byboen сквозь рону и гар, сквозь пустой водойсе, we exan & bugy gouyuduuns nocrob, Kan Sygro usen neuen ne cerces kyepol!

U. Epodenus. Стихи под эпигразам. По, гро додволено Юпияску на Kangour open borois nor Эланон, нас и удаг. В катоби слудыке Бах, Бог 1150 berusen - Soraus. Бренноса - уден быхов... Сумеросани богов. И кадо небол высклуго, U, money Durs, rebronad. Ещё кас не раз распнуб Mexany norou pachag.

### Other "Documents"



# **IBM Cross Pad Data**

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



₿? FFX IISC MELLO TO RUN FOR MAYOR OF Bim, Bam, Bum - Ein Glockenton Bin Bam, Bun- Eis Clacker for FREMONT Fremont Councilman Gazy 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 in for mayor, but he said subsequently on several occupions the chances were still he would challenge Ball. After a long and difficult decision-making process, I have Jecided not to ven for mayor of Fremont, at this time, " Mel aid in a written statement. Mello's council seat does not expire unkil 1993. In a telephonic interview, the 40-year-old little instrance impany executive said he did not want to devote more time to city Taxsiness at the expense of his family and job. He said he currently spends about '30 hours a week on city-related and that being mayor would mean at bast 10 per week. flour, two trans of via, two tons of beans and hundricks of perios of super, times, cooling oil and used dollary

fliegt durch die Nacht, als fligh durce die Noclt, als hätt' er Vogelflügel, er fliegt Latt or Voul Right or flift in römischer Kirchentracht in rowiscler fircle backt wohl über Tal und Hügel. Er Wall other Tal und Hard Er GENERUSITY FINANCES MERCY MISSION TO FARMWORKERS A silust disater has shicken the people of Cultorino's Central valling, and two Frement by Hall staffers are spearlieading a June to help. The situation is stark. " Prope on oping to stear to death," wanted Frender Mayor Bir The French effort is focused on the area around Linervey, h of taliformal's largest elive precising plant and, according needy letterhand, the "Heart of the California Grange Area." The severe winker wifed out the atms crop themployment Scholson, Nathray behaven Freshe and Packershild, is running at an periode Work will not be available in the work industry there with November. San Francisco, I. Centre, and Edmond depend 12 tax of



### GEDI

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



#### **Software Architeture**

- Efficient Technology Transfer
  - software compatibility
  - balance of academia, governemnt, and industry needs
  - common framework for document processing
- Scalability
  - rapid prototyping of new methods
  - simple algorithm comparison
- Robustness and Stability
  - high quality standards
  - platform-independence
  - accommodation of frequently changing requirements



#### **DocLib Status**

Core DocLib components matured and stable (in use by a variety of government installations)\

Addons being integrated/implemented, primarily by developers

Freely available to government researchers

➢Core supported on Solaris, Linux and Windows



### Core vs Add-ons

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

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



### **Image Factory**



#### **Design Factors:**

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



#### **DocLib Architecture**

DocLib's architecture rests on two pillars:





### **Document Hierarchy**





### **Recent Modules**

- Thinning
- Rotation
- Deskewing
- XML i/o
- Degradation
- OCR Scansoft interface (Windows)
- Docstrum
- Logo detection
- Signature processing

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

# Logo and Stamp Detection



### Document/Non-Document Discrimination

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

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



### Non-Documents















### **Original Challenges**

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



### Introduction

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





#### Page Segmentation for Noisy Documents





\* Docstrum page segmentation technique is used



Overview of Our Approach

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



#### **Classification Results with Fisher Classifier**





### 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	Percen tage	Before Post- processing		re Post- essing 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**

	ATTORNEY PRIVILEGED
201428 17 1948 CC-4 8039 777 7980 8039 777 7981 8039 777 7981	Sentenber 21, 1981 803 - 783 - 1512 13
Ernust Papples, Esq. Sentar Yice-President ( Brown & Hilliamson, In P.D. Sou: 3500 Louisville, Eentucke 44	und General Counce C. 1932
Deer Rr. Papples The enclosed sum describing solutions an to contact these maters.	ary presents my elems coccentried yournal eriticles d PCD mayagroments, 2 appreciate the opportunity th
· ·	Sincerely yours, J. J. J
TD0/jac Biclournet Grant Sumary M Topre 1 Date function	Arrative II -
	ма g



After enhancement

1

#### Before enhancement



### Handwritten Line Detection



#### 4 ATTORNEY POINT AND A THREE IN CONTRACT POINT AND A TORNEY POINT AND A THREE IN A TORNEY AND A THREE AND A THREE

Notic to a reader of data related is concentration restancements of channel minimum in body divide is an understanding of the interface of uptote, environmental and physical statements, into the flood. Elect this is entire the cases interprint, at the electric is a substitution of parts in the entire two cases interprint, at the electric is a substitution between the little is beginning to determining the site of equilibrities between the cliffic is important to determining the site of equilibrities between the cliffic is a superior transmitter in the superscale presentation is a supers, bickness, and the body will effect itself values. If is determining a superstate is a superstanding the tests will block itself values. If is determined a superstate is and established also of the block is a superstate interview is an established in a size of the block is a superstate in the superstate is a superstate.

The full wring table firsts the edge further which impact on the entering of deterministics. It first he conters will used per digerate protect when collecte bias willows and already market moves do holders are edge to alternize the classifier conterts.

#### <u>Fights 1</u> subtrast Arresting the services arrests. The

- $L_{\rm e}$  . Suppliering and eccuracy of the mathematical star for containfaction of contrast planet without  $\mu$
- Continuing in telling (1971) INCO-Chang.
- a. reactor of character anded MT MOD.
- . Wartstips is daily eigensite unities patiette
- a, visibility pattern of the individual performant in the study:
- a. Partation in homes of claparation probations for
- many weight/surface area conjusted analy Minist prime antibility;

301029311

- a loss of the subjects.
- Convention held-block for sicciling metabolisms in cattering. Pacture 3, 6, and 6, contact with <sup>147</sup>, dat be work matter of a correction particle which income all work or verified which is blic.

# Challenges

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



- Significant Gap

- Regular lavout

	عد بيرية ا <sup>ع</sup> من محافظة السليما نيه
	( ( القيود السريه ) )
	للمسيحة الماسية المسيدة
	series 11
	إ قدم لمردو : بما ي لريو وجها ماج طب لد جل الدوال
	الأدين الحصرة فوا تصريح وقد وطالب هذها
	عد معد مالح البلا من من مع من مع مع مع مع مع
	Y ovv a start and a start and the
	الم لاف الم الم الم الم الم الم الم
	- slot
	الروسي لا المحالية المحالية المحالية المعالمة
	الدقت كاعبر مساحة من المارم المح معالية
	رالغان عرب مدم لعد الد المرجم
c	٧- ٢ مدم ورك بدافنا من الات الدول الرواس
	الا حوال من من الم
	analytic and a second
	A Direct for and the local state
	Ci in un interentions
	Till Company and
	Co portant
	Sarily"
	i Carriera de la companya de la comp
	//pta

# **Our Assumptions**

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



# Intuitive Approach

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



### The level set method



http://vision.ece.ucsb.edu/~sumengen/level\_set\_methods

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



### Results





ت خلقة (ما فع الاسركتر شراجة ا برتی 👘 ۲۰۰ میں بنوں شیو تا بیکٹ لاند کردز ۱ ماند ۱ می منظمة سک ستینا با ت سنطقه ۲ مسط لسه ا مركز مرك ستغدا لات جومات ، ساكت ف جه سال الاجرحسن في المعلم رما مه . ١ و مَنَ الشَيْقَ وَالَيْهِ مِنْ 17-907 applying من متول مشيرته بكلب مدتوكروز الجع نظمة بمؤسقيه لات منط · المة طري (ما / عود/ محامل ، *شنایت رنسیوتی ح*لاکادژ علاترہ نے کما پا 2 مردات ول عررا 56 40 فاردن ماي محصورت امام كمال ، ما ي مع شكون مول لايت مى مل برت ۱ مامی کمامل بی تحسیق رت وریخ د مامه ۱ مایی به سر سعیان المام بیکامل با سریلاملتورد ماده عل مسكمة بديات اماي المواجة كالمراسو كما مدة ملکان مآم للزش خربت مَدرا دِترکیا مل تعطم سيرتك = املى كمامل عدر ت ساويس سمسرد دلات، والبعة كاقل عدامة طاديل استرن ۱۰ - سر ریابة ۱ مایین کامل تعطیسی سر بوتس مار تلکیر ۵۰ قیل ، سیرت سے شامی مادام کاملہ ج بلاطرا سرن کامل ··· مسرون علم جرمة الحسل ومان م خط شباك تعط امان كالل •- صلح وسيرة معان كما علم مع خلم ترس عربة مع مددكة سرت ١٠ مسرا بيل مع مع يدتة معه المتأخذة عدر م جمية من من شر مسبع المحات الم تنطاق عليه علم المسر مروحة كماهل طفلة يتعدر علية تشرر مؤفرا وفاحل عليا رة المقة اعلاء مذاك يعل منغل والمتابع ويداد المسلما اد معلي مع الموتك معد وبعد ذالك تقديم رموية المالحك جسيط حول رماً ولا interio e an ar Borraci III a





# Evaluation

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



- 5000 text lines

## Comparison

#### QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100

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

HANDWRITTEN HINDI DOCUMENTS.

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100

HANDWRITTEN CHINESE DOCUMENTS.

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



# Comparison (cont')

QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100

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

HANDWRITTEN WITH BACKGROUND PARALLEL LINES.

#### QUANTITATIVE COMPARISON OF OUR APPROACH AND A CONNECTED COMPONENT BASED APPROACH ON 100 FREESTYLE

#### HANDWRITTEN DOCUMENTS.

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



### Robustness Test

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

動物研究的205、这么好的现象和1821年18、1241年18月1日 184、2014年18日(194)日本前上前上面1945年、2015年2014月1日 学校教育的2055年18日在1845年1845年1841年2014年19月1日

2 419/2026. 2202. 2010/210.2005 198 food Long 420.019 1000 40 1007 500. 199 food Long 420.019 1000 40 1007 500 単体相関のなべ、2014年の1月1日11日1日(1940年)2月1日(1945年) 1月1日、よう2月1日(1945年)1月1日(1945年)(1945年)(1945年) 学校開始(1945年)2月1日(1945年)(1945年)(1945年)(1945年)

 A. Looks Cooks Stellandow M. Look in the State State State Stellandow M. Look in the State State State Stellandow M. Look in the State State State Stellandow M. Look M. State A. State State

Source of the for a second file of the office of the second secon

n linder the set and the state of the set of

if any glass, it and and farely the party i ward entering designed of maring ion of the set being provided and the ton total a trialistic sociale as 4+ + survey a service it want & stylene tool anima by sten equiverily to any had the have a set of the and a set to set of Winit Stricks & good to stand of stil his strong and underso, strationic 1. in Anni 2. 129 . 128 dianay & mail 108 1011.01 Ases will Bernston And a regard s anipute lost wante side en postoriasi mass fedara van anyour Antine to of the far and





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

## **Publications**

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



# DVD

- Dataset without Ground Truth
- Slides from Today



# Agenda

Projeci Overview

- Introduction
- Goals and Objectives

Data Collection and Ground Truth

GEDI and Evaluation Framawork

Evaluation and Research Components

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

#### **Technical Presentations**

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

