LAMP Review and Planning Meeting
And
Related Video Analysis Research

December 18, 2008

Laboratory for Language and Media Processing
University of Maryland, College Park

AGENDA:
LAMP Review and Planning Meeting

9:30am December 18, 2008
AV Williams, Room 2120

9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation *David Doermann, Elena Zotkina, Wontaek Seo, Levon Mkrtchyan*
10:30	Text line and rule-line detection and removal for handwritten documents *Wael Abd-Almageed, Mohammed Rafaey, Jayant Kumar*
11:00	Voronoi++ - Extension of page segmentation for handwritten documents *Mudit Agrawal, David Doermann*
11:30	Clutter Detection and Enhancement *Mudit Agrawal, David Doermann*
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications *Xiaodong Yu, Daniel DeMenthon*
1:45	Video Processing @ LAMP – Introduction *Wael Abd-Almageed*
	Sports Video Summarization using Text Webcasts *Mohammed Rafaey, Wael Abd-Almageed*
2:10	Processing Video Collections on GPU Arrays *Ramani Duraiswami*
2:40	Kernel-based Learning on GPUs *Mohammed Hussein, Wael Abd-Almageed*
3:00	Understand Videos, Constructing Plot *Abhinav Gupta*
3:30	Discussion and Future Plans Potential Video Tasks Doclib Enhancements Tools and Datasets: GEDI Enhancements Proposed Public Data and Evaluation Framework – (w/ ARL)

LAMP Media Publications
2007-2008

L. Yi, Y. Zheng, D. S. Doermann and S. Jaeger. Script-Independent Text Line Segmentation in Freestyle Handwritten Documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1313-1329, August 2008.

T. Steinherz, D. Doermann, E. Rivlin and N. Intrator. Off-Line Loop Investigation for Handwriting Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. (ACCEPTED).

Jian Liang, Daniel DeMenthon and David Doermann. Geometric Rectification of Camera-captured Document Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 591-605, July 2008. (ACCEPTED).

Jian Liang, Daniel DeMenthon and David Doermann. Mosaicing of Camera-captured Documents Without Pose Restriction. *Computer Vision and Image Understanding*, 2008. (ACCEPTED).

Xu Liu, D. Doermann and Huiping Li. VCode - Pervasive Data Transfer Using Video Barcode. *IEEE Transactions on Multimedia,* 10(3), pages 361-371, April 2008. (ACCEPTED).

Xu Liu, David Doermann and H. Li. A Camera-based Mobile Data Channel: Capacity and Analysis. *ACM International Conference on Multimedia*, 2008.

Xu Liu and David Doermann. Mobile Retriever: Access to Digital Documents from their Physical Source. *International Journal on Document Analysis and Recognition* (ACCEPTED), 2008.

Guangyu Zhu, Xiaodong Yu, Yi Li and David Doermann. Unconstrained Language Identification Using A Shape Codebook. *The 11th International Conference on Frontiers in Handwritting Recognition (ICFHR 2008)*, pages 13-18, 2008.

Guangyu Zhu, Yefeng Zheng and David Doermann. Signature-based Document Image Retrieval. *The 10th European Conference on Computer Vision (ECCV 2008)*, pages 752 - 765, 2008.

Guangyu Zhu, Xiaodong Yu, Yi Li and David Doermann. Learning Visual Shape Lexicon for Document Image Content Recognition. *The 10th European Conference on Computer Vision (ECCV 2008)*, pages 745 - 758, 2008.

Mudit Agrawal and David Doermann. Re-Targetable OCR with Intelligent Character Segmentation. *DAS*, September 2008.

W. Abd-Almageed, M. Agrawal, W. Seo and D. Doermann. Document Zone Classification Using Partial Least Squares and Hybrid Classifiers. *ICPR*, 2008.

Xu Liu, David Doermann and H. Li. Camera Phone Based Tools for People with Visual Impairments. *The First International Workshop on Mobile Multimedia Processing*, pages in press, 2008.

Xu Liu and D. Doermann. A Camera Phone Based Currency Reader for the Visually Impaired. *The Tenth International ACM SIGACCESS Conference on Computers and Accessibility*, October 2008.

Xu Liu, D. Doermann, H. Li, K. C. Lee, Hasan Ozdemir and Lipin Liu. A Novel 2D Marker Design and Application in Object Tracking and Event Detection. *4th International Symposium on Visual Computing*, pages to appear, December 2008.

Guangyu Zhu, Yefeng Zheng, David Doermann and Stefan Jaeger. Signature Detection and Matching for Document Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

Xu Liu and D. Doermann. Computer Vision and Image Processing Techniques for Mobile Application. Technical Report: LAMP-TR-151, Center for Automation Research, University of Maryland, November 2008.

Stefan Jaeger, Huanfeng Ma and David Doermann. *Machine Learning in Document Analysis and Recognition: Combining Classifiers with Informational Confidence*. Chapter: . Springer, LNCS, 2007.

Guangyu Zhu and David Doermann. Automatic Document Logo Detection. *The 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, pages 864-868, 2007.

Sergey Tuljakov, Stefan Jaeger, Venu Govindaraju and David Doermann. *Machine Learning in Document Analysis and Recognition: Review of Classifier Combination Methods*. Chapter: . Springer, LNCS, 2007.

Xu Liu and D. Doermann. Mobile Retriever - Finding Document with a Snapshot. *CBDAR 07*, pages 29-34, September 2007.

Guangyu Zhu, Yefeng Zheng, David Doermann and Stefan Jaeger. Multi-scale Structural Saliency for Signature Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pages 1-8*. Minneapolis, MN, 2007.

Zhe Lin, Larry S. Davis and David Doermann. Hierarchical Part-Template Matching for Human Detection and Segmentation. *IEEE International Conference on Computer Vision (ICCV'07)*, pages 1-8, 2007.

Jimmy Lin, Michael DiCuccio, Vahan Grigoryan and W. John Wilbur. Exploring the Effectiveness of Related Article Search in PubMed. Technical Report: LAMP-TR-145/CS-TR-4877/UMIACS-TR-2007-36/HCIL-2007-10, University of Maryland, College Park, July 2007.

Xiaodong Yu, Yi Li, Cornelia Fermuller and David Doermann. Object Detection Using Shape Codebook. *British Machine Vision Conference (BMVC'07)*, December 2007. (accepted).

J. Hannuksela, P. Sangi, J. Heikkila, X. Liu and D. Doermann. Document Image Mosaicing with Mobile Phones. *International Conference on Image Analysis and Processing (ICIAP'07)*, pages 1-8, September 2007.

Ryan Farrell, David Doermann and Larry S. Davis. Learning Higher-order Transition Models in Medium-scale Camera Networks. *Workshop on Omnidirectional vision, Camera Networks and Nonclassical Cameras (ICCV'07)*, pages 1 - 8, 2007.

Sameer Kibey. Tools for Advanced Video Metadata Modeling. Technical Report: LAMP-TR-141/CAR-TR-1024/CS-TR-4857/UMIACS-TR-2007-11, University of Maryland, College Park, February 2007.

Zhe Lin, Larry S. Davis, David Doermann and Daniel DeMenthon. Simultaneous Appearance Modeling and Segmentation for Matching People under Occlusion. *Asian Conference on Computer Vision (ACCV'07)*, pages 404-413, 2007.

Zhe Lin, Larry S. Davis, David Doermann and Daniel DeMenthon. An Interactive Approach to Pose-Assisted and Appearance-based Segmentation of Humans. *Workshop on Interactive Computer Vision (ICV'07)*, pages 1-8, 2007.

Guangyu Zhu, Timothy J. Bethea and Vikas Krishna. Extracting Relevant Named Entities for Automated Expense Reimbursement. *The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 1004 - 1012, 2007.

LAMP Review and Planning Meeting

December 18th, 2008

Agenda

9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation
10:30	Text line and rule-line detection and removal for handwritten documents
11:00	Voronoi++ - Extension of page segmentation for handwritten documents
11:30	Clutter Detection and Enhancement
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications
1:45	Video Processing @ LAMP – Introduction
	Sports Video Summarization using Text Webcasts
2:10	Processing Video Collections on GPU Arrays
2:40	Kernel-based Learning on GPUs
3:00	Understand Videos, Constructing Plot
3:30	Discussion and Future Plans

Recent Graduates

- Xu Liu (PhD, 2008)
  - Computer Vision and Image Processing Techniques for Mobile Applications
- Burcu Karagol-Ayan (PhD, 2007)
  - Resource Generation from Structured Documents for Low-Density Languages
- Yang Yu (MS, 2007)
  - Human modeling and Tracking
- Sameer Kibey (MS, 2007)
  - Threat Modeling for video analysis
- Nagia Ghanem (PhD, 2007)
  - Petri Nets for Video Analysis

Current Researchers

- Faculty
  - David Doermann
  - Wael Abd-Almageed
- Faculty Researchers
  - Elena Zotkina, Wontaek Seo
- Graduate Students
  - Mudit Agrawal, Xiaodong Yu, Guangyu Zhu, Mohammed Rafaey, Jayant Kumar, Xu Liu
- Undergraduates
  - Zach Ollson, Orri Ganel, Levon Mkrtchyan

Publications

Publications
- 2007: 14
- 2008: 17

- List provided in Handouts
- Xu Liu. A Camera Phone Based Currency Reader for the Visually Impaired. *The Tenth International ACM SIGACCESS Conference on Computers and Accessibility*, October 2008.
  - Placed second at the ACM Student Research Competition associated with ASSETS2008 in Halifax.
  - Mobile currency reader for the visually impaired is being beta tested nationwide by the National Federation for the blind,
  - Will participate in the ACM Grand Finals in April.

LAMP Related Topics/Research

- IJDAR
  - Editorial Office, publishing over 40 journal papers per year
- ICDAR
  - Program Chair and reviewers from LAMP
- CBDAR
  - Chairing Camera Based Document Analysis and Recognition
- Processing Historic Documents with Library of Congress

External Impact of LAMP Funding

- **Visor**: Video Surveillance Online Repository
  - Rita Cucchiara, Imagelab – Dipartimento di Ingegneria dell'Informazione , University of Modena and Reggio Emilia, **Italy**
- http://imagelab.ing.unimore.it/visor/

External Impact of LAMP Funding

- ViPER also being used for
  - VIRAT DARPA Program
  - TrecVid Event Annotation
  - VACE Event Annotation
- GEDI – Used for ICDAR Competitions, as well as MADCAT
- DocLib Delivered as part of CDIP

External Impact of LAMP Funding

- Doclib and GEDI instrumental for BOBCAT-DI
  - Enhancements to GEDI
  - Evaluation tools Integrated
  - Datasets Generated

Transition previously developed test methods, metrics, procedures, and associated software developed to BOBCAT-DI as part of the assessment infrastructure, with focus on established metrics, such as word-error rate, character-error rate, and recall-precision in the case of image-based search or handwriting-print discrimination.

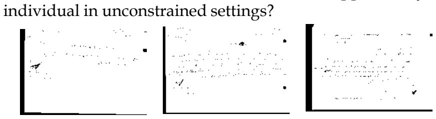
Recent Addons to DocLib*

- Signature, Stamp and Logo Detection
- ScriptID
- ImageID
- Line Detection (Second Talk)
- * Note Yet Vetted Through the System

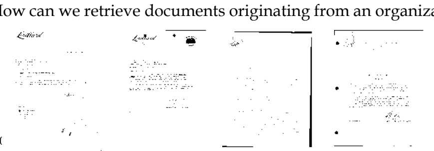
Problem Statement

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?



- ❑ How can we retrieve documents originating from an organization?

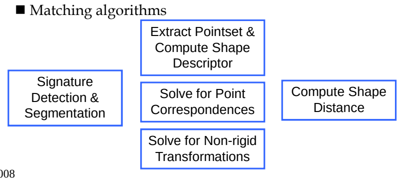


Motivation

- ■ 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

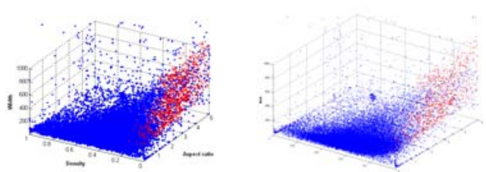


Challenges

- Detecting free-form objects in cluttered backgrounds is a challenging problem in computer vision
- 2D nature of off-line signatures
  - Difficult to recover tempo order of unconstrained off-line handwriting [1]
- Large intra-class variations of signatures
  - Intersession variability
  - Larger variations than other forms of handwriting
- Computation complexity

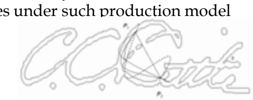
Intersession variability shown by Sabourin *et al.* [6]

Intra-class Variations of Signatures



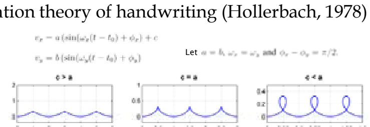
Overview of our approach

- 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



Signature production model

- We assume that
  - The pen moves in a cycloid fashion with reference to a sequence of shifting virtual baselines.
  - Local baseline changes as the pen moves its position with respect to the document.
  - Within a local curve segment, we consider that the baseline remains unchanged.
  - The locus of the pen maintains a proportional distance from the local center point (*focus*) to the local baseline (*directrix*).
- Oscillation theory of handwriting (Hollerbach, 1978)

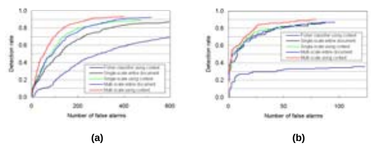


Signature production model

- This is equivalent to viewing signatures as piece-wise concatenations of small elliptic segments.
- The model imposes one additional constraint that limits the group of feasible second-order curves to smoother ellipses.

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.



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

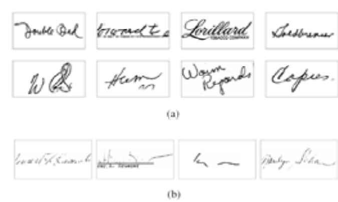
Evaluation

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



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

Overview of our approach

- We treat a signature as a shape
- Employ shape matching techniques for signature recognition
  - Shape representations
  - Shape matching algorithms
  - Measure of dissimilarities for shapes (shape distance)

Signature Detection & Segmentation

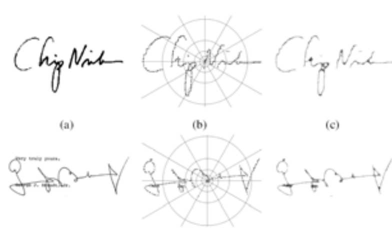
Extract Pointset & Compute Shape Descriptor

Solve for Point Correspondences

Solve for Non-rigid Transformations

Compute Shape Distance

Shape representation



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

Shape matching

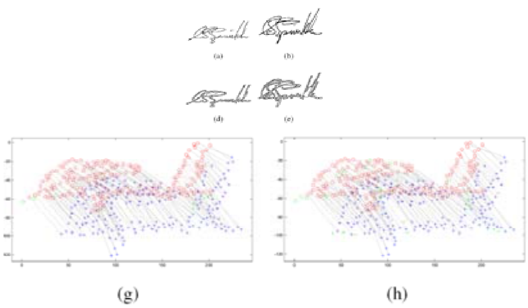


Illustration of signature matching using shape contexts and local-neighborhood-graph

Shape matching

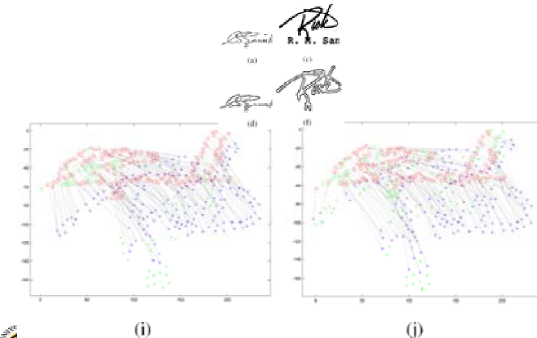
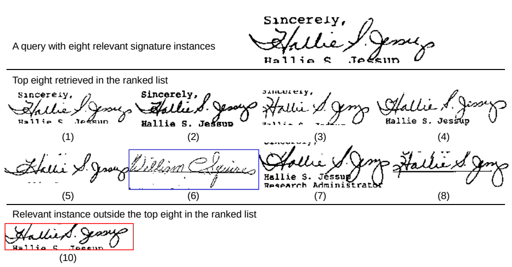


Illustration of signature matching using shape contexts and local-neighborhood-graph

Shape matching evaluation



A 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%

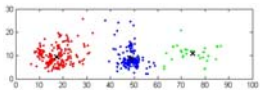
Current work

- Create groundtruth on Arabic documents for evaluation
- Preliminary results are very promising on Arabic datasets

Overview

- Propose a joint formulation for logo detection and extraction 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

Feature selection and extraction



Positions of logos in the Tobacco-800 dataset relative to the entire document.

We define context distance as

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

Table 3: Features used for classification.

Context distance	Aspect ratio
Spatial density	Area

Evaluation

- We use accuracy and precision as evaluation metrics

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

$$\text{Precision} = \frac{\text{\# of correctly detected logos}}{\text{\# of detected logos}}$$

- We consider a logo *correctly detected* if and only if the detected region contains more than 75% overlapping pixels with the groundtruth AND its area is less than 125% of the area of the groundtruth.

Table 4: Positions of logos in the Tobacco-800 dataset relative to the entire document.

	Accuracy	Precision
Improved spatial density [8]	39.3%	32.1%
Fisher classifier only, *i.e.* $\lvert S \rvert = 1$	59.2%	41.7%
Multi-scale approach with $\lvert S \rvert = 2$	57.0%	68.1%
Multi-scale approach with $\lvert S \rvert = 3$	84.2%	73.5%

Evaluation

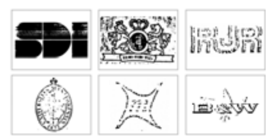
Examples of correctly detected logos from Tobacco-800.

Evaluation

(a) Over/under-segmented logos

(b) Non logos

Examples of incorrectly detected logos.



Examples of missed logos.

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

Classification of Script and Images

- 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
    - natural image
  - Dataset: ~3700 images from Internet.

Unconstrained Script Identification Using A Shape Codebook

Motivation

- Language identification remains a fundamental problem in document image analysis
- Increasing industrial demand for automatic processing of heterogeneous multilingual off-line document images
  - Google Book Search (Vincent, ICDAR 2007)
  - Global expense reimbursement (Zhu *et al*., KDD 2007)
- The performance of language ID is important in
  - Determining the correct OCR engine
  - Document indexing, translation, and search
- Prior research focused exclusively on machine printed text

Motivation

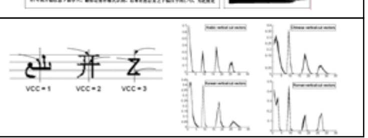
- Real document collections often contain a diverse and complex mixture of machine printed and handwritten text
- Language ID for handwritten document images is an open research problem

Challenges

- Categorization of unconstrained handwriting presents a number of fundamental challenges
- Handwriting exhibits much larger variability compared to machine printed text
  - Larger variations in the shapes of handwritten words due to style, cultural, and personalized differences
  - Freestyle handwriting text lines are curvilinear in nature
  - Gaps between neighboring words and lines are not uniform
- The approach needs to be robust in the presence of unconstrained document layouts, formatting, and image degradations

Related works on script and language ID

- Prior research focused exclusively on machine printed text
  - Statistical analysis of text lines

Upward concavity [Spitz, PAMI97]	
Projection profile [Suen et al., ICDAR98]	
Vertical component cuts [Lu and Tan, PAMI08]	

Related works on script and language ID

- Texture analysis

Gabor filters [Tan, PAMI98]

Wavelet [Busch et al., PAMI05]

- Template matching [Hochberg et al., PAMI97]

Rescaling to fixed size templates

Related works on handwritten language ID

- There exists very little literature on language identification for unconstrained handwriting
- Early experiment by Hochberg et al., IJDAR97
  - Use linear discriminant analysis on 5 simple features of connected components, including centroid locations and aspect ratio
  - Sensitive to variations across writers
  - Cannot robustly handle mixed content on document
  - Machine printed text, illustrations, markings, and handwriting in different orientations were manually removed from evaluation dataset

Language ID for handwritten document images

- Language ID for diverse handwritten content needs to be robust against
  - Presence of complex mixture of machine printed text and unconstrained handwriting
  - Unconstrained document layouts and large variations in font and style

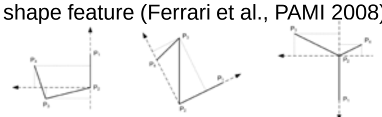
Our approach

- Effectively capture intricate differences between languages using segmentation-free shape features and a geometrically invariant shape descriptor

- Important problems to be addressed
  - Encoding shape information in a translation, scale, and rotation invariant scheme
  - Obtaining concise, structural indexing for large number of shape features extracted from diverse content

Our approach

- Encode local text structures using geometrically invariant shape codewords
- Learn a codebook of shape features by clustering and partitioning similar feature types
- Construct an image descriptor based on the frequency of occurrence of indexed features
- Local shape feature (Ferrari et al., PAMI 2008)



A shape feature $\mathcal{P}$ is compactly represented by an ordered set of orientations and lengths of s_i for $i \in \{1, 2, 3\}$, where s_i denotes line segment i in $\mathcal{P}$.

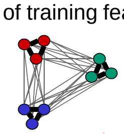
Learning the shape codebook

- Feature detection
  - Local line fitting on connected components formed from edges
  - Requires linear time
- Computing feature dissimilarity

$$D(P_a, P_b) = w_\theta \sum_{i=1}^{3} D_\theta(\theta_i^a, \theta_i^b) + \sum_{i=1}^{3} |log(l_i^a / l_i^b)|$$

- Constructing a graph representation of training features

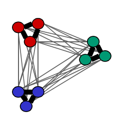
$$w(P_a, P_b) = \exp(-\frac{D(P_a, P_b)^2}{\sigma_D^2})$$



$\mathcal{G} = (V, E)$

Learning the shape codebook

- Creating structural indexing of diverse features

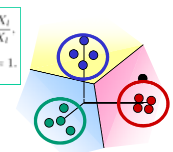


$\mathcal{G} = (V, E)$

$$\text{maximize } \epsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l},$$

$$\text{subject to } X \in \{0,1\}^{N \times K}, \text{ and } \sum_j X(i,j) = 1.$$

Normalized cuts algorithm



- Constructing image descriptor

$$D(P_k, C_k) < r_k$$

Nearest neighbor assignment with quantization w.r.t. cluster radius r_k

Experiments

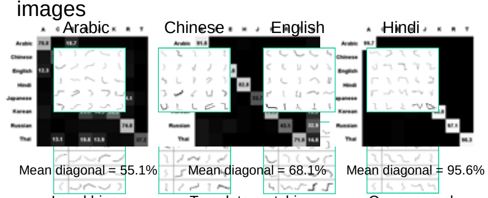
- Language identification for handwritten document images

Language	Arabic	Chinese	English	Hindi	Japanese	Korean	Russian	Thai
Pages	245	184	175	186	161	168	204	189
Writers	22	18	26	17	8	25	5	10

  - We constructed a 1,512 complex document image database for 8 languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) composed of mixture of handwriting and machine printed content

Experiments

- Language identification for handwritten document images



  - Our approach gives excellent results on all 8 languages, with a mean diagonal of 95.6%

Experiments

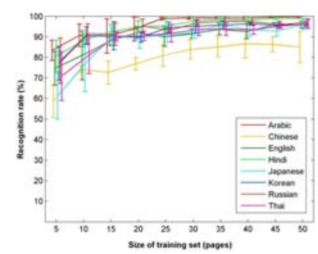
- Language identification for handwritten document images

	A	C	E	H	J	K	R	T
A	99.7	0.3	0	0	0	0	0	0
C	1.4	85.0	4.0	1.0	6.7	1.0	0.7	0.2
E	1.6	0	95.9	0.2	0	1.1	0.6	0.6
H	0.2	0.2	0	98.8	0.8	0	0	0
J	0	1.3	1.0	0.2	96.2	1.3	0	0
K	0	0.8	0.1	1.9	0.5	96.0	0.5	0.1
R	0.5	0	2.0	0	0	0	97.1	0.4
T	0	0.3	1.6	0.9	0.6	0.3	0	96.3

Mean diagonal = 95.6%

  - Our approach shows good generalization performance across large variations such as font types or handwriting styles

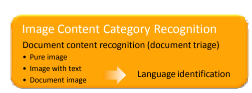
Experiments



- Our approach achieves excellent performance even using a small number of handwritten document images per language for training

Document Image Content Recognition Using A Shape Codebook

Background and Motivation



- High-level content interpretation becomes a fundamental vision problem
- Text-oriented image content recognition provides a reliable approach
- Two research problems are largely open
  - Image content category recognition
  - Unconstrained language identification

Examples of images with text returned by Google Image using the keyword "CD cover".

Image content category recognition

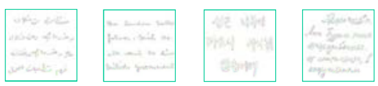
- A fundamental problem in computer vision and image analysis
- Focus on text content within images
  - Pervasive presence of text
  - Once text content and the language are recognized, images containing text can be processed by OCR systems and conveniently indexed
- Main challenges:
  - Diverse, unconstrained visual content
  - Large intra-class and inter-class variations
  - Diverse feature types, mixture of printed and handwritten text, fonts, and styles
  - Unconstrained layouts and formatting, and cluttered background
  - Computational complexity

State of the Art

- Most systems assume content is known
- Other techniques rely heavily "recovery"
  - Finding text –> implies document
  - Approach remains challenging for noisy or handwritten content in unstructured documents
  - Limits applications
- Global techniques (wavelets, texture) classify text, but ignore fine features

Our approach

- Intricate differences effectively captured using generic low-level vision primitives

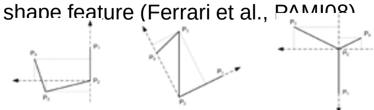


- Important problems to be addressed
  - Capturing shape information in a geometrically invariant fashion
  - Obtaining concise, structural indexing for diverse, and potentially large feature space

Our approach

58

- Encode using geometrically invariant lexical words
- Learn a lexicon of shape features by clustering feature types
- Partition the feature space
- Construct an image descriptor based on the frequency of occurrence of lexical words
- Local shape feature (Ferrari et al., PAMI08)

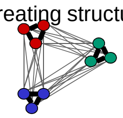


A shape feature $\mathcal{P}$ is compactly represented by an ordered set of orientations and lengths of s_i for $i \in \{1, 2, 3\}$, where s_i denotes line segment i in $\mathcal{P}$.

Shape lexicon

59

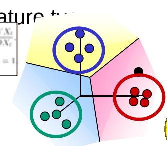
- Feature detection
  - Local line fitting on connected components formed from edges
  - Requires linear time
- Computing $D(P_a, P_b) = w_\theta \sum_{i=1}^{3} D_\theta(\theta_i^a, \theta_i^b) + \sum_{i=1}^{3} |log(l_i^a / l_i^b|)$
- Creating structural indexing of feature t



$\mathcal{G} = (V, E)$

$$\text{maximize } \epsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l},$$
$$\text{subject to } X \in \{0,1\}^{N \times K}, \text{ and } \sum_j X(i,j) = 1$$

Normalized cuts algorithm

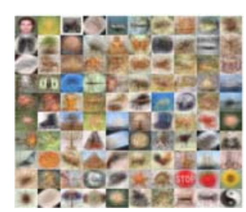


Text-oriented image content recognition

60

- Text-oriented high-level content interpretation
- Object category recognition and localization

- We focus on text and recognize an image as one of three content categories – pure image, image with text, and document image



Averaged image of each object category from Caltech 101 [L. Fei-Fei et al., PAMI06]. Picture courtesy of Antonio Torralba.

Experiments

61

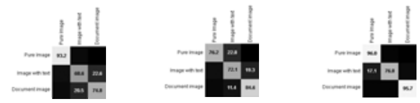
- Image content categorization

Pure images   Images with text   Document images

- We constructed a 4,500 image database by crawling Web images from Google Image search engine using a wide variety of text keywords

Experiments

- Image content categorization



Mean diagonal = 78.9%	Mean diagonal = 77.6%	Mean diagonal = 89.6%
Spatial envelope	Local binary patterns	Our approach

- We compare our approach with two well-known whole-image categorization approaches
- Spatial envelope recognizes natural images very well, but not on text
- LBP demonstrates balanced performances across content categories

Status

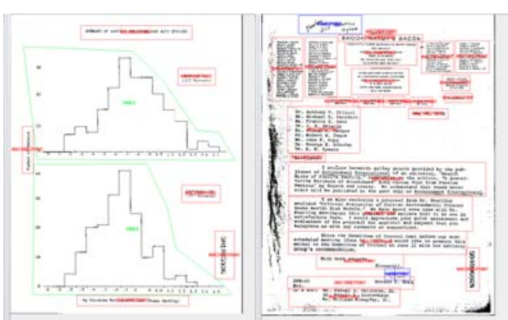
- Current Doclib Addons
- Interested in Larger Scale Testing

- Publications
  - Guangyu Zhu, Xiaodong Yu, Yi Li and David Doermann. Unconstrained Language Identification Using A Shape Codebook. *The 11th International Conference on Frontiers in Handwritting Recognition (ICFHR 2008)*, pages 13-18, 2008.
  - Guangyu Zhu, Yefeng Zheng and David Doermann. Signature-based Document Image Retrieval. *The 10th European Conference on Computer Vision (ECCV 2008)*, pages 752 - 765, 2008.
  - Guangyu Zhu, Xiaodong Yu, Yi Li and David Doermann. Learning Visual Shape Lexicon for Document Image Content Recognition. *The 10th European Conference on Computer Vision (ECCV 2008)*, pages 745 - 758, 2008.

Datasets

- Handwritten LanguageID dataset
- ImageID dataset
  - 10,000 images
- Tobacco Corpus Segmentation Collection
  - 25,000 pages
  - Identification of document type (memo, letter, etc)
  - Zone level ground truth around
    - Machine Print Block, Handprint blocks, Logos, Signatures, Stamps, Tables, Graphs/Graphics, Images
  - Optical Character Recognition of all Machine Print Author Information about the document

Tobacco GT



Agenda

9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation
10:30	Text line and rule-line detection and removal for handwritten documents
11:00	Voronoi++ - Extension of page segmentation for handwritten documents
11:30	Clutter Detection and Enhancement
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications
1:45	Video Processing @ LAMP – Introduction
	Sports Video Summarization using Text Webcasts
2:10	Processing Video Collections on GPU Arrays
2:40	Kernel-based Learning on GPUs
3:00	Understand Videos, Constructing Plot
3:30	Discussion and Future Plans

Agenda

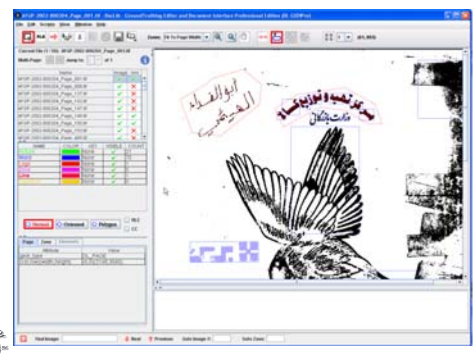
9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation
10:30	Text line and rule-line detection and removal for handwritten documents
11:00	Voronoi++ - Extension of page segmentation for handwritten documents
11:30	Clutter Detection and Enhancement
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications
1:45	Video Processing @ LAMP – Introduction
	Sports Video Summarization using Text Webcasts
2:10	Processing Video Collections on GPU Arrays
2:40	Kernel-based Learning on GPUs
3:00	Understand Videos, Constructing Plot
3:30	Discussion and Future Plans

DocLib Tools: Evaluation and Annotation

BOBCAT-DI Tasks

- Data Sets
  - Zone Classification and Segmentation GT
  - Character/Word level GT
- Tools
  - Modify UMD's GEDI to allow handwritten data representation
  - Develop DocLib Extensions/add-on routines
  - Extend ARL Image and OCR Toolkit (IOTK)
- Evaluation
  - Conduct Segmentation evaluations
  - Conduct Zone Classification evaluations

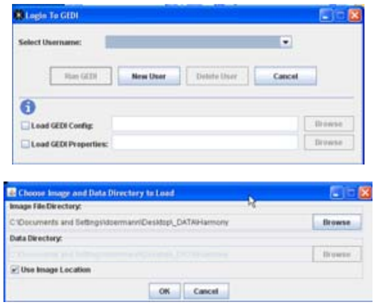
GEDI

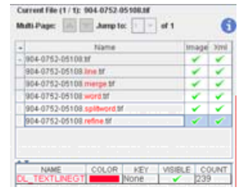


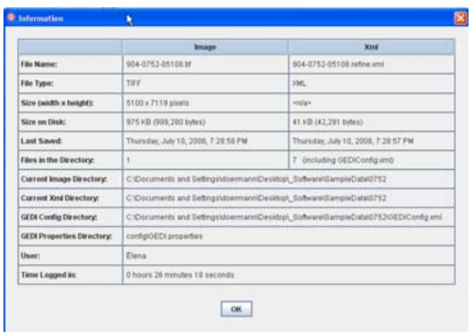
Recent Enhancements

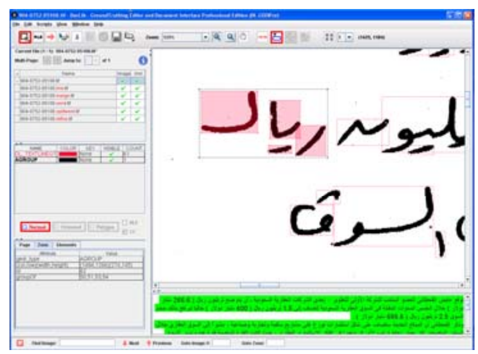
- Configuration
  - Log in and ability to load specified GEDIConfig and Properties files
- Image and XML Information
  - Browse panel (the top) shows total # of images in directory
  - Find panel: find image by name, lump to given #; search for zone
- Document Navigation
  - Selected zone doesn't loose selection on zooming;
  - Zone recentering on zooming
- Display Enhancements
  - Pseudo coloring by attribute value
  - Line thickness

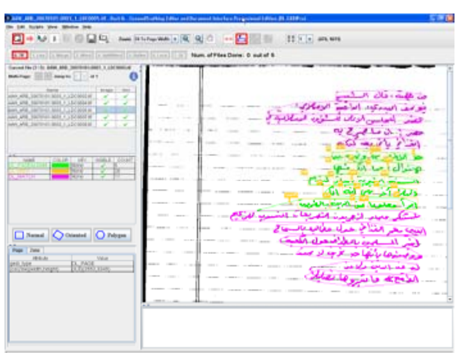
- Grouping capability
- Electronic Text Window – content of selected zone is highlighted here
- Document Generations
  - General parsing of image.XYZ.xml and grouping them; collapse and expand row
- Reading Order
- RLE and CC
- Network Listener - Enhanced
- Script panel
- GEDI Help











PETS Software

- Performance Evaluation Tool for Segmentation (PETS)
- Usage
  - PEZS -r { FILE | DIR } -g { FILE | DIR } -img { FILE | DIR }
  - [-o FILE -v DIR -m FILE -t NUM -detail -lid -rle -seg]
- Programming Language: C++
- Will be provided as add-on application in DocLib.

PETS General Concept

- Given two zones to be compared, calculate the matching score if there is at least one shared ON pixel
- Four types of result
  - MATCHED: location and zone type
  - DETECTED: location but not zone type
  - FALSE: Extra zone in Results
  - MISSED: Zone not matched from GT
- Threshold is set to determine which zones are matched for "detection"
- Zone types "can" be used for matching
- Software is integrated into DocLib
- Full match matrix is built to store the score of each pair of zones.

Matching Score and Result Types

- **I** = set of all ON pixel in Image
- R_i = set of all ON pixel in the result zone
- G_j = set of all ON pixel in the ground truth zone
- **T(s)** = function that count the elements of set s

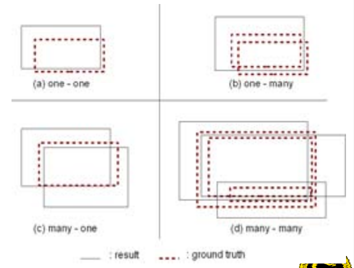
$$MatchScore(i,j) = \frac{T(G_j \cap R_i \cap I)}{T((G_j \cap R_i) \cap I)} \times 100$$

MATCHED
MatchScore(i,j) ≥ threshold
L(i) = L(j)

DETECTED
MatchScore(i,j) ≥ threshold
L(i) ≠ L(j)

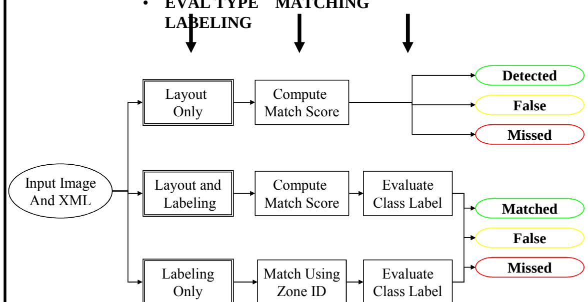
FALSE
MatchScore(i,all) < threshold

MISSED
MatchScore(all,j) < threshold



Evaluation Framework

- **EVAL TYPE MATCHING LABELING**



Segmentation and Classification

```
==================
Summary of Results
==================

- Total Number of Sample :  21786
- Overall Accuracy : 95.78%
- Average of Each Class Accuracy : 55.31%

01. Information on Classes
==========================
```

Label	Name of Class	Number of Sample	Accuracy
00	text_sm	20617	97.34%
01	ruling	201	61.69%
02	drawing	299	88.29%
03	table	76	46.05%
04	text_lg	51	64.71%
05	math	301	60.47%
06	halftone	144	83.33%
07	logo	13	0.00%
08	chm_drawing	80	51.25%
09	map	4	0.00%

Segmentation and Classification

02. Confusion Matrix
====================

Out\GT	00	01	02	03	04
00	20068(97.3%)*	70(34.8%)	11(3.7%)	14(18.4%)	12(23.5%)
01	69(0.3%)	124(61.7%)*	0(0.0%)	1(1.3%)	1(2.0%)
02	93(0.5%)	1(0.5%)	264(88.3%)*	23(30.3%)	4(7.8%)
03	46(0.2%)	0(0.0%)	5(1.7%)	35(46.1%)*	0(0.0%)
04	19(0.1%)	1(0.5%)	0(0.0%)	0(0.0%)	33(64.7%)*
05	284(1.4%)	2(1.0%)	8(2.7%)	2(2.6%)	1(2.0%)
06	38(0.2%)	3(1.5%)	6(2.0%)	0(0.0%)	0(0.0%)
07	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)
08	0(0.0%)	0(0.0%)	5(1.7%)	1(1.3%)	0(0.0%)
09	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)

	05	06	07	08	09
	106(35.2%)	5(3.5%)	7(53.8%)	0(0.0%)	0(0.0%)
	0(0.0%)	0(0.0%)	1(7.7%)	0(0.0%)	0(0.0%)
	9(3.0%)	18(12.5%)	0(0.0%)	9(11.3%)	4(100%)
	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)
	0(0.0%)	0(0.0%)	4(30.8%)	0(0.0%)	0(0.0%)
	182(60.5%)*	0(0.0%)	0(0.0%)	30(37.5%)	0(0.0%)
	0(0.0%)	120(83.3%)*	0(0.0%)	0(0.0%)	0(0.0%)
	0(0.0%)	0(0.0%)	0(0.0%)*	0(0.0%)	0(0.0%)
	4(1.3%)	1(0.7%)	1(7.7%)	41(51.2%)*	0(0.0%)
	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)*

Segmentation and Classification

03. Precision and Recall
========================

Class\Eval	precision	recall	detected	correct	total
00	98.89%	97.34%	20293	20068	20617
01	63.27%	61.69%	196	124	201
02	62.12%	88.29%	425	264	299
03	40.70%	46.05%	86	35	76
04	57.89%	64.71%	57	33	51
05	35.76%	60.47%	509	182	301
06	71.86%	83.33%	167	120	144
07	0.00%	0.00%	0	0	13
08	77.36%	51.25%	53	41	80
09	0.00%	0.00%	0	0	4

Name : PETS Performance Evaluation ToolS for zone segmentation and classification

Synopsis :

Unix/Linux platform command : PETS

Window platform command : PETS.exe

command r {<FILE>|<DIR>} g{<FILE>|<DIR>} i {<FILE>|<DIR>} [o <FILE>] [v <DIR>] [m <FILE>] [t <NUM>] [detail] [lid] [rle] [segonly|zoneclass] [az <FILE>|naz <FILE>]

Options :

r {<FILE>|<DIR>} : Location of Results File(s)

g {<FILE>|<DIR>} : Location of Ground Truth File(s)

i {<FILE>|<DIR>} : Location of Image File(s). Default location is the location of ground truth

o <FILE> : Name of File for Evaluation Results. Default is 'PETS Eval.txt'.

v {<FILE>|<DIR>} : directory where Xml output of GEDI format will be saved

lid : Zones which have same 'lineID' attribute in Ground truth will be merged to one zone

rle : runlength code will be add to visualization output

detail: enable detailed output for each zone

t <NUM> : set the threshold by user for determining a zone match based on pixecounts. Default is 80(%).

m <FILE> : result zones which are in a ground truth zone will be merged if it's types are in the <FILE>. First line of the FILE should have numeric data which is used as threshold for zone merging.

segonly : Evaluation will perform detection by not consider zone labels for matching.

zoneclass : Evaluation will rely on ZoneIDs for correspondence, considering only zone labels for results

az <FILE> : Zones which its types are in the <FILE> will be treated in the program, otherwise deleted from the result.

naz <FILE> : Zones which its types are in the <FILE> will be deleted from the result.

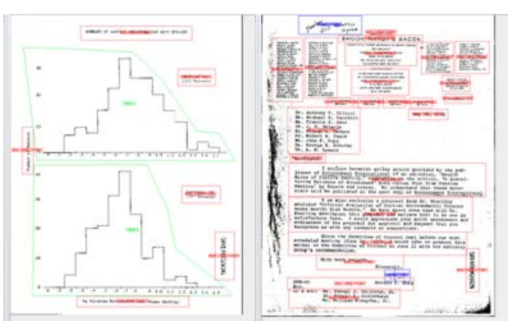
BOBCAT-DI

- Pilot Evaluations
  - Page Segmentation
  - Zone Classification
  - Line Detection
  - Clutter Detection and Removal

Datasets

- Handwritten LanguageID dataset
- ImageID dataset
  - 10,000 images
- Tobacco Corpus Segmentation Collection
  - 25,000 pages
  - Identification of document type (memo, letter, etc)
  - Zone level ground truth around
    - Machine Print Block, Handprint blocks, Logos, Signatures, Stamps, Tables, Graphs/Graphics, Images
  - Optical Character Recognition of all Machine Print Author Information about the document

Tobacco GT



Document Zone Classification

- Previous work
  - Many methods exist for finding particular types of zones (e.g. logo, signature, text, etc.)
    - Detection – integrated classification
      - G. Zhu, Y. Zheng, D. Doermann and S. Jaeger, "Signature Detection and Matching for Document Image Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.*
      - G. Zhu and D. Doermann, "Automatic Document Logo Detection", *The 9th International Conference on Document Analysis and Recognition (ICDAR 2007), pages 864-868, 2007.*
    - Segmentation – Requires subsequent zone classification
      - O. Okun, D. Doermann and M. Pietikainen. Page Segmentation and Zone Classification: The State of the Art. Technical Report: LAMP-TR-036/CAR-TR-927/CS-TR-4079, University of Maryland, College Park
- Best known zone classifier
  - Y. Wang, I. T. Phillips, and R. M. Haralick "Document Zone Content Classification and Its Performance Evaluation," PR, Jan. 2006.
  - Evaluation on University of Washington dataset
  - Accuracy: 98.45%

Document Zone Classification

- Limitations
  - U. Of Washington dataset is highly unbalanced
  - 88% of the samples is small text
  - No reported results on a balanced dataset
  - Assumes a latent reading order
  - Does not work on Harmony-like complex documents

Zone Classification using Hybrid Classifiers and PLS

- Objective:
  - Develop a novel zone classifier that
    - Works on complex documents
    - Robust against noise
- Feature Extraction
  - Structural features
    - Based on Run-length of foreground pixels
    - Encode pixel distribution
  - Textural features
    - Based on Local Binary Patterns (LBP)
    - Encode local texture information
  - Feature vector of 321 attributes
  - Dimensionality reduction?

Zone Classification using Hybrid Classifiers and PLS

- Classically, if we have *C* classes
  - Use e.g. Multiple Discriminant Analysis
  - Multiclass classifier
    - *C* one-against-all classifiers
    - $C(C-1)/2$ one-against-one classifiers; $f_{a,b}$
  - Voting scheme to determine the unknown zone class
- Limitations
  - If the test sample is not one of two underlying classes, an incorrect vote is cast
  - $C(C-1)/2 -1$ incorrect votes

Zone Classification using Hybrid Classifiers and PLS

- Construct $C(C-1)/2$ two-against-all classifiers -- indicator classifiers; $f_{a\text{-}b,all}$
- Construct $C(C-1)/2$ one-against-one classifiers; $f_{a,b}$
- Use indicator classifiers to determine which binary classifier to use
- Only one vote is cast per test sample
- Partial Least Squares for dimensionality reduction
  - Works like PCA but better in preserving discriminative structure
  - *Compute* $2*(C(C-1)/2)$ data projections

Zone Classification using Hybrid Classifiers and PLS

- Results
  - Test on UW dataset
  - 1690 documents – 24531 zones
  - 10 zone classes
    - Chemical drawings, small text and symbols, drawing, halftone, logo or seal, map, math, ruling, table and large text

	1-vs-1	Wang et al.	Hybrid
Unbalanced	93.1%	98.45%	97.3%
Balanced	88.2%	N/A	96.6%

Document Zone Classification

- Publication
  - W. Abd-Almageed, M. Agrawal, W. Seo and D. Doermann, "Document-Zone Classification using Partial Partial Least Squares and Hybrid Classifiers," ICPR 2008
- In progress
  - Use more texture features; Zernike moments
  - Use one-class Support Vector Machines
  - Test on Harmony dataset
  - Annotations for Harmony?
  - Submit paper to ICIP 2009

Document Zone Classification

- In progress
  - Improved classifier for Harmony/Anfal-like data
  - Use more texture features; Zernike moments
  - Use one-class Support Vector Machines
  - Annotations for Harmony?
  - Submit paper to ICIP 2009
- Future work
  - Integration into Doclib
  - SVM implementation?

Text Line Detection

- Objectives:
  - Very fast
  - Very accurate

Text Line Detection – Affinity Propagation[1]

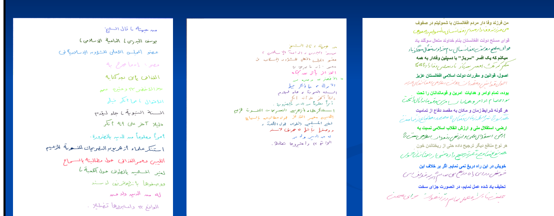
- Unsupervised clustering technique
- Advantages:
  - Operates in similarity space, rather than the feature space
  - Number of clusters need not be *a priori* specified
  - Can cluster models on non-Euclidean manifolds
- Affinity Propagation animation
  - (courtesy of the University of Toronto)

1. B. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, v. 315, no. 5814, 2007.

Text Line Detection using Affinity Propagation

- Algorithm
  - For each connected component, do feature extraction
    - Compute centeroid
    - Compute orientation
  - Compute pair-wise orientation similarities using a Gaussian kernel
  - Apply AP using orientation similarities to separate into oriented text layers
  - For each layer, compute location similarities to detect individual text lines

Text Line Detection



LDC data

Dari machine printer and handwritten data from ARL

Sample Results

Text Line Detection – Evaluation

- Algorithm evaluated on the LDC dataset[1]
- 1250 document images
- 21145 text lines

Precision	Recall	F_1 score
78.2%	84.15%	81.06%

1. Ground truth contained annotation inaccuracies (e.g. punch holes)

Text Line Detection – Evaluation

- Integrated into Doclib
  - 2 seconds per 2500x2500 document image
  - Single file/batch modes
  - Export results to GEDI XML
- Beats the Doclib[1] text line detection code (approx. F_1=50%)
- (Much) Faster than our previous text line detection[2]

1. Stefan Jaeger, Guangyu Zhu, David Doermann, Kevin Chen and Summit Sampat. DOCLIB: a Software Library for Document Processing. *International Conference on Document Recognition and Retrieval XIII, pages 1-9, 2006*
2. L. Yi, Y. Zheng, D. S. Doermann and S. Jaeger. Script-Independent Text Line Segmentation in Freestyle Handwritten Documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1313-1329, August 2008.*

Text Line Detection

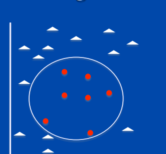
- In progress
  - Improvement to detect text in Harmony/Anfal-like documents
  - ICIP 2009 submission

Rule Line Removal

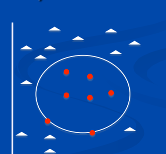
- Objective
  - Remove lines in complex, handwritten documents
- Classic approach
  - Use a binary classifier
  - Model rule line pixels as one class
  - Model everything else as another class
- Limitation
  - A binary classifier is not appropriate because modeling the universe is impossible
  - The "everything else" class will not work for handwritten documents where style varies

Rule Line Removal – New Approach

- One-Class Support Vector Machines
  - Model only class of interest (positive examples)
  - No need to model negative examples
  - PCA is enough for dimensionality reduction

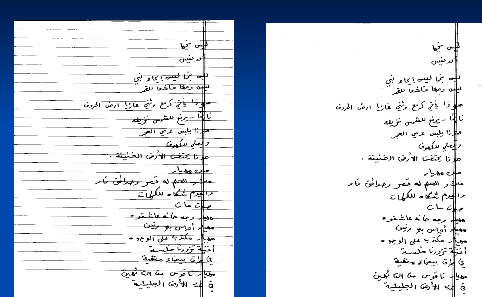


Binary SVM



One-class SVM

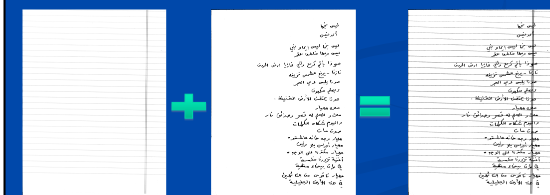
Rule Line Removal – New Approach



Sample Results

Rule Line Removal – Evaluation

- Very difficult because it needs pixel level annotation
- Synthetic dataset: 25 test images



Rule Line Removal – Evaluation

- Fp = foreground pixels present in the text image but not present in neither template nor the output image
- Fn = Foreground pixels present in both template and output images
- Tp = Foreground pixels present in the template but were removed from the text image
- Precision = Tp/(Tp + Fp)
- Recall = Tp/(Tp + Fn)
- F_1 = 2*Precision*Recall/(precision + recall)

Rule Line Removal – Evaluation

- Trained 12 one-class SVMs
- Test on synthetic dataset
- Compute average and median F_1 score

	F-1 Measure TYPE I						F-1 Measure TYPE 2					
AVERAGE	0.74222	0.653984	0.740112	0.661952	0.743116	0.636424	0.472492	0.32932	0.600408	0.373036	0.671212	0.3863
MEDIAN	0.7415	0.6589	0.7388	0.6597	0.7428	0.6543	0.4165	0.2404	0.5305	0.2885	0.5953	0.3205

Rule Line Removal

- In progress
  - Add rotation invariant features
  - Add text reconstruction
  - Evaluate on Harmony/Anfal-like data
    - Pixel-level ground truth?
  - Integrate into Doclib

Agenda

9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation
10:30	Text line and rule-line detection and removal for handwritten documents
11:00	Voronoi++ - Extension of page segmentation for handwritten documents
11:30	Clutter Detection and Enhancement
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications
1:45	Video Processing @ LAMP – Introduction
	Sports Video Summarization using Text Webcasts
2:10	Processing Video Collections on GPU Arrays
2:40	Kernel-based Learning on GPUs
3:00	Understand Videos, Constructing Plot
3:30	Discussion and Future Plans

Voronoi++
Extension of page segmentation for handwritten documents

Page Segmentation

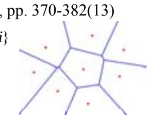
- Traditional Approaches
  - X-Y cuts
  - Smearing
  - Text Based Classifiers
  - Whitespace Analysis
  - Constrained Text-Line Detection
  - Docstrum
  - Voronoi based
- Challenges
  - Content overlapping between Regions

Voronoi Segmentation

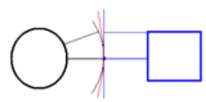
Segmentation of Page Images Using the Area Voronoi Diagram. Kise K.; Sato A.; Iwata M. CVIU, Volume 70, Number 3, June 1998 , pp. 370-382(13)

$V(p_i) = \{p \mid d(p, p_i) \le d(p, p_j), \forall j \ne i\}$

Voronoi Region for P:



Voronoi Area for Region G:



$V(g_i) = \{p \mid d(p, g_i) \le d(p, g_j), \forall j \ne i\}$

where

$d(p, g_i) = \min_{q \in g_i} d(p, g_i)$

Voronoi Area Approximation:

For every point on the boundary, generate V(P), and remove edges between identical components!

Steps

- Generate Voronoi Area Approximation
- Follow Contours and Sample
- Select Edges for removal based on features
  - Minimum Distance
    $d(E) = \min_{1 \le i \le m} d(p_i, q_i)$, p_i & q_i are pair of points constituting the ith edge between CCs
  - Area Ratio $a_r(E) = \dfrac{\text{max of areas of 2 CCs}}{\text{min of areas of 2 CCs}}$
- Delete an edge if
  - $d(E)/T_{d1} < 1$
  - $d(E)/T_{d2} + a_r(E)/T_a < 1$

  where $T_{d1} < T_{d2}$

  T_{d1} relates to inter-character spacing

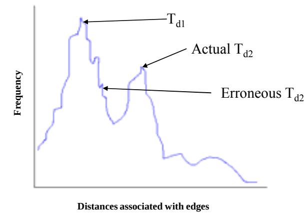
  T_{d2} relates to inter-word/line spacing

Examples

Problems for Handwriting

1. T_{d2} gets a local maxima after $T_{d1.}$ These are not consistent for Handwriting
2. Common Distance Threshold across a single page not the right measure for spurious edge deletion
3. Distance as the only parameter/feature per edge for deletion is not sufficient
4. Global information missing from local features

Problem 1: Spacing Thresholds



- **Solution: Assuming word-separation (T_{d2}) is at least c times than char-separation (T_{d1}) [where c > 1.5], smooth histogram with a window of size c *T_{d1} and then find T_{d2}**

Problem 2: Distance Thresholds

Choosing a common distance threshold across a single page doesn't suffice

Diacritics in Arabic handwritten text documents affect thresholds

The diacritics generally are at a higher distance from the word than word-separation boundary, giving them a separate region

Using a higher noise threshold leads to over-segmentation, as smaller characters do not participate in edge-formation, forming spaces between a single word

Solution:
- Let every component participate in edge-formation (noise-threshold independency)
- Component should not form an edge with its nearest neighbor (Docstrum idea)

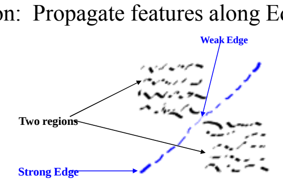
Problem 3: Distance Features

Distance and size features should not be the only features used

- Content features (zone texture, for example)
- Perceptual features (orientation, density, etc)
- Proximity to other edges

Problem 4: Local Features

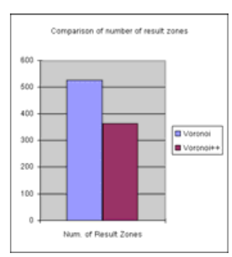
- Example: Edge Strength may cause some edges to fall below a threshold
- Solution: Propagate features along Edges



Restricting Over-segmentation

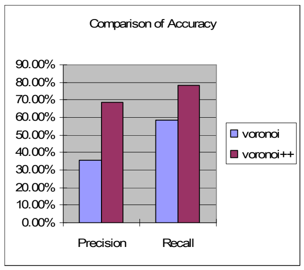
DataSet : 25 image from ZoneClassification-AMA

	Num. of Result Zones
Voronoi	526
Voronoi++	362



Accuracy Improvement

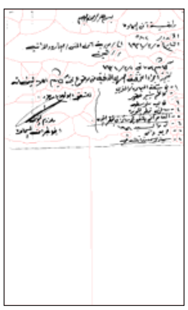
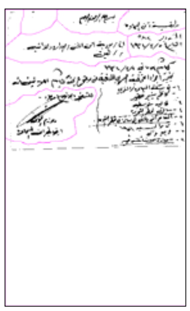
	Precision	Recall
voronoi	35.36%	58.49%
voronoi++	68.78%	78.30%



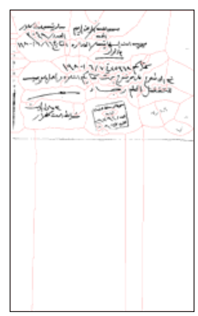
Parameters

Parameter	Description	Sensitive (Y/N)?
sr	Sampling rate	Y
nm	Size Th on noise CC	Y
Ch	CC height Th	N
Cw	CC width Th	N
Cr	CC aspect ratio Th	N
Az	Min area Th of a zone	N
Al	Min area Th	N
Br	Max aspect ratio Th	N
sw	Smoothing window	N
Td1	Inter char Th1	Y
Td2	Inter char Th2	Y
Ta	Area ratio Th	Y

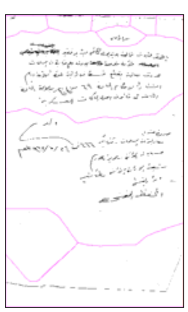
Examples

Examples




Examples

Agenda

9:30	LAMP Overview
10:00	DocLib Tools: Evaluation and Annotation
10:30	Text line and rule-line detection and removal for handwritten documents
11:00	Voronoi++ - Extension of page segmentation for handwritten documents
11:30	Clutter Detection and Enhancement
12:00	*Working Lunch*
1:00	Weakly Supervised Object Categorization for Real-world Applications
1:45	Video Processing @ LAMP – Introduction
	Sports Video Summarization using Text Webcasts
2:10	Processing Video Collections on GPU Arrays
2:40	Kernel-based Learning on GPUs
3:00	Understand Videos, Constructing Plot
3:30	Discussion and Future Plans

Clutter Detection and Enhancement

Motivation

- Problem 1:
  - Generic classifier to recognize documents with unstructured-Noise (focus on clutter and salt and pepper)
- Problem 2:
  - Classify *noisy* documents as
    - Salt-n-pepper
    - Clutter
- Problem 3:
  - Appropriate Noise Removal
- Goals:
  - Detect and Remove Clutter

Classification of Noise

- **Independent** Noise (e.g. salt and pepper, ink blobs)
  - Often an additive noise model (Structured)
    - recorded image *f(i,j)* is the sum of the *true* image *s(i,j)* and the noise *n(i,j)*:

$$f(i,j) = s(i,j) + n(i,j)$$

    - noise *n(i,j)* is often *zero mean* and described by its variance σ_n^2
- Noise **dependent** on image data (e.g. clutter, ruled lines)
  - Difficult to model
  - Mathematically non-linear and very complicated

Various Noise Measurements

- Signal to noise ratio (SNR)

$$SNR = \frac{\sigma_s}{\sigma_n} = \sqrt{\frac{\sigma_f^2}{\sigma_n^2} - 1}$$

where σ_s^2 and σ_f^2 are variances of true image and recorded image respectively

- Contrast ratio (CR)
  - Defined as the ratio of the mean value of the background to the mean value of the foreground
- Gaussian additive zero mean
  - This is characterized by the variance σ^2 of the values of the noise distribution.
- Input data error (IDE))
  - For percentage errors in pixel (e.g. for salt and pepper noise)

Types of Noise

- Salt and Pepper
- Marginal Noises
- Blobs
- Punched Holes
- Smudges
- Bleed/show-through
- Stamps
- Logos
- Pen Strokes
- Ruled Lines
- Borders
- …

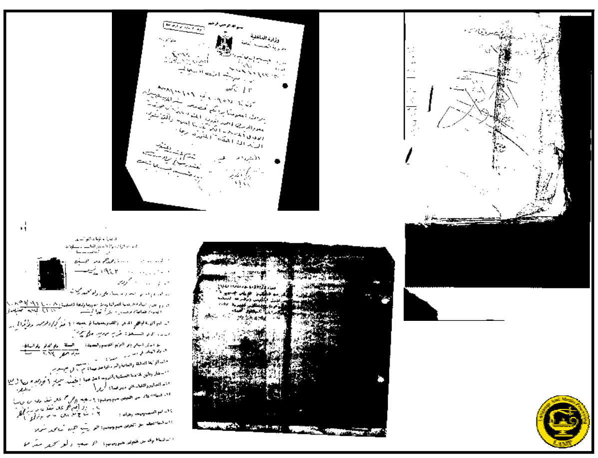
Unstructured

pattern *smaller than text-stroke width*

Foreground Pattern
larger than avg. text-stroke width
CLUTTER

Specific Patterns
(Across Text)

Structured



Unstructured Noise Removal

PRIOR ART

- Salt n Pepper Noise
  - Detection [Ali 1996]
    - Creates tiles of standard size from image.
    - Classifies each tile as clean text, noisy text, noisy and empty using a neural-network. Features are
      - # color transitions in horizontal and vertical directions
      - # black pixels in 8-neighborhood for each black pixel.
    - Neural network based
  - Removal
    - Morphological opening

Unstructured Noise Removal

PRIOR ART

- Clutter [Fan2002]
  - Specific algorithms for marginal noise removals
  - Identification and Removal based on
    - Length
    - Position
    - Neighborhood

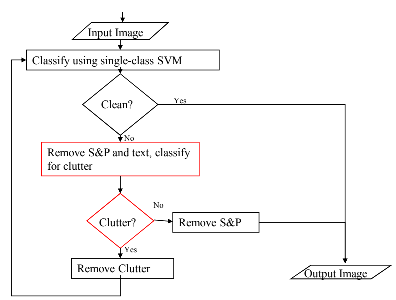
https://wiki.umiacs.umd.edu/lamp/index.php/MADCAT_PageEnhancementSurvey

Why we need better solution?

- Because current work relies upon the following features of noise
  - Position
  - Shape
  - Neighborhood
  - Disconnectivity from pure signal (independent)

General Approach

System Model



- Observations
  - Clutter often interacts with text content
  - Clutter typically has fundamentally different structure then content
- Challenges:
  - Can not remove large content that may also remove content
  - Blindly applying morphology is bad for Arabic Handwriting
- Approach
  - Use a generative model based on a distance transform to identify clutter regions
  - Fast, controllable

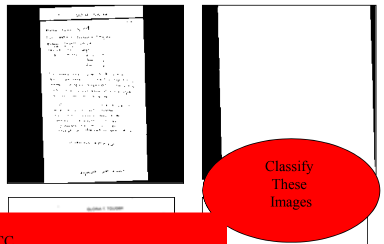
Approach

- Perform Distance Transform
  - For binary images, a distance transform specifies the distance from each pixel p to the nearest non-zero pixel
- Calculate the maximum value of Distance Transform over the foreground (zero pixels)
  - Strokes have approximately similar *depth (width)*
- Remove all content with D < dMax/k (k=2)

$$D_P(p) = \min_{q\in I}(d(p,q) + f(q))$$

where initially,

$$f(q) = \begin{cases} \infty & \text{if } q \in P \\ 0 & \text{otherwise} \end{cases}$$

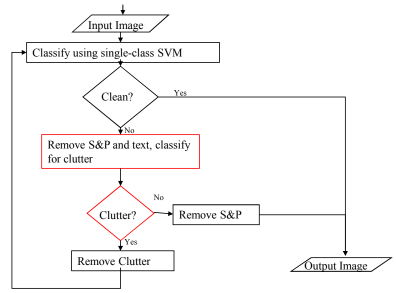


Classify These Images

Features
- # of CC
- Average size of CC
- Variance in size of CC
- Variance in position of centroids of CC
- Sum of Areas/Sum of Perimeters
- Ratio of CC after half-erosion to before erosion

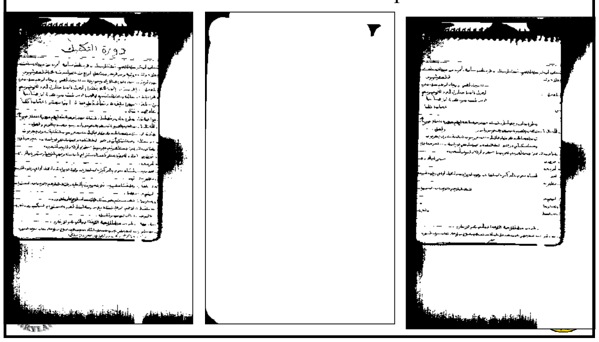
2-class linear SVM

System Model



Clutter Removal

- We can not delete entire clutter component…

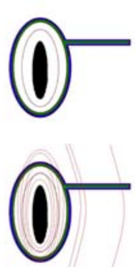


Restrictive Clutter Removal

- Clutter is often attached with ruled lines which are in turn attached with text
- Removal of clutter as a connected component may erase all the text attached to it
- Challenge is to remove clutter restrictively, without removing the text branches attached to it
- Solution should be independent of
  - Clutter structure or shape
  - Text-font(s)
  - Text-size
  - Amount of text attached
  - Direction of text

Approach

- Identify the clutter region, which may be touching text…
- Compute Distance Transform of the eroded image
- Compute the difference image between the two sets distance transforms
- Use the original distance Transform as a mask.



Algorithm

Frequency Sets

$D_P(x)$ = Distance Transform at a point x on set P
$D_{Q'}(x)$ = Distance Transform at a point x on set Q'

$$f(i) = distinct\left(\bigcup_{p \in P_i} \lceil D_{Q'}(p) \rceil\right)$$

$$where$$

$$P_i = \{\forall p \in P \mid \lceil D_P(p) \rceil = i)$$

$$\forall i \in [1, dtMax/2]$$

- Shrinking the clutter to its core cuts off the text-branches at some value of i, say i_{th}
- It is this value of i, there will be a *sudden dip* in monotonically decreasing f

i.e. change in the rate of change of f, reaches a maxima

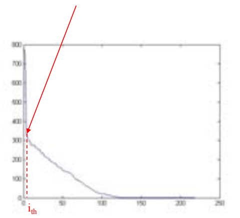
Algorithm

Determining text-branch threshold

- $f(x)$ = Number of distance contours ($D_{Q'}$) passing through clutter pixels (P) at i^{th} distance from clutter's boundary
- $f'(x)$ = rate of change of the function
- This rate of change slows down at i_{th}
- $g(x) = f''(x)$
- Hence, i_{th} = index of *first* maxima of $g(x)$

$$\frac{d}{dx}(g(x)) = 0$$

$$\frac{d^2}{dx^2}(g(x)) < 0$$



Algorithm

Shrinking and Dilating

- The clutter is shrunk to i_{th} from its boundary

$$\begin{aligned}&\mathit{if}\ \lceil D_P(x)\rceil < i_{th}\\&\quad I(x) = bg\\&\mathit{endif}\end{aligned}$$

- This cuts of text branches

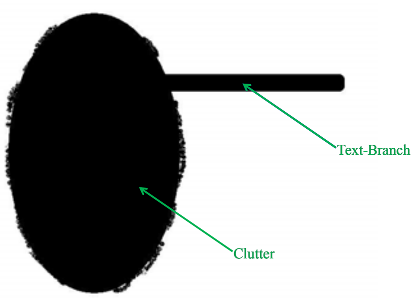
$$R = \{\forall x \,|\, I(x) = fg\}$$

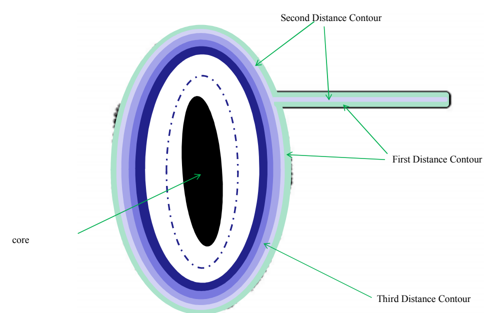
- Resultant shrunk clutter is dilated i_{th} to regain clutter shape and size

$$\begin{aligned}&\mathit{if}\ \lceil D_{R'}(x)\rceil < i_{th}\\&\quad I(x) = fg\\&\mathit{endif}\end{aligned}$$

- Clutter is removed by subtracting original image from this

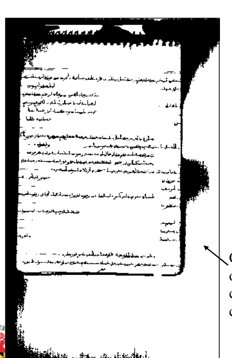
$$O_{clean}(x) = O(x) - I(x)$$



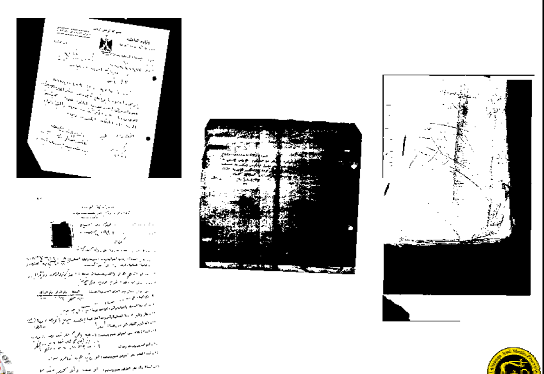


Distance Contours increasing inside, from boundary of clutter

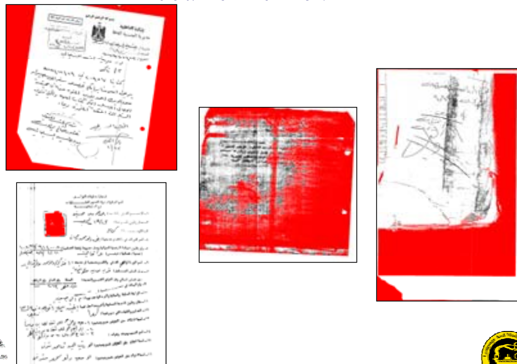
Attached text is not removed!



Clutter as one single connected component



Results Removal



Summary

Weakly Supervised Object Categorization for Real-world Applications

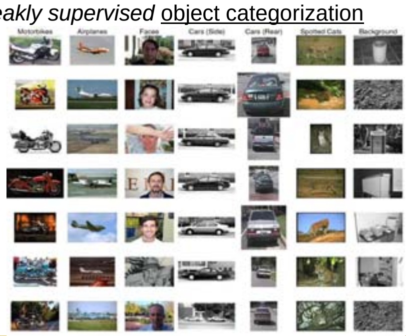
Xiaodong Yu
LAMP · UMIACS · ECE
University of Maryland, College Park, MD

Outline

- Background and Motivation
- Approaches
  - Object categorization for imbalanced image sets
  - Object categorization using Web image sets
- Summary

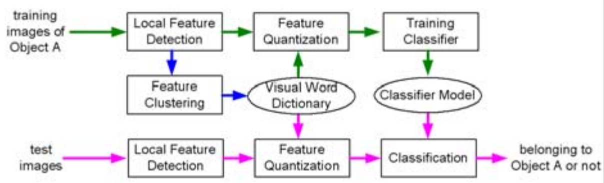
Background

- *Weakly supervised* object categorization



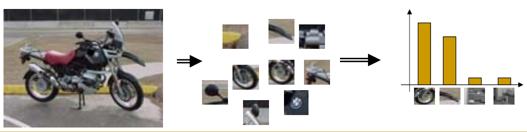
Sample images in the Caltech6 image set [Fergus et al2003]

A General Framework for Object Categorization of the State of the Art



Bag-of-words Image Representation

- Local features:
  - PCA-SIFT
  - SIFT on Regular Grids (RG-SIFT)
- Visual word
  - Representatives of clusters in the local feature space
- Bag-of-Words
  - Image = histogram of visual words



Limitations of Conventional Approaches (1)

- Training classifiers using manually labeled image sets
  - e.g., Caltech 101/256, PASCAL VOC
  - Limited scalability in real-world applications
    - Image labeling is a tedious and error-prone task
    - Too many object categories in the real-world to be labeled
      - More than 75,000 non-abstract nouns in English listed in the Wordnet
    - The number of training samples for each category is limited

Limitations of Conventional Approaches (2)

- Training and testing classifiers using balanced image sets
  - Balanced image sets:
    - airplane vs. Caltech background
  - Many real-world applications are imbalanced classification problems
    - airplane vs. non-airplane

Learning from Imbalanced Image Sets

- The problem of learning from imbalanced data sets
  - Given only positive training examples
  - No negative training examples are give explicitly but the number of them may be infinite
    - airplane vs. non-airplane
- Imbalanced data sets
  - Majority classes vs. minority classes
- Challenge:
  - Standard binary classification approaches will be biased towards the majority classes
    - Accuracy on majority classes are very good
    - Accuracy on minority classes are poor

An Application for Digital Forensic Image Mining

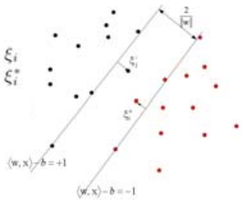
- Problem description
  - Given: a number of images of certain classes
  - Task: to identify images of these classes from a large collection of images on a suspect PC

Classifier 1: SVM

- Given labeled data set $\{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$ the goal of SVM is to find a hyperplane that separates two classes with maximum margin

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases}$$



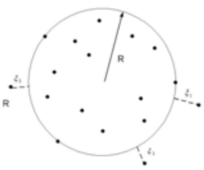
Classifier 2: SVDD

- Given a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ the goal of SVDD is to obtain a sphere enclosing these data with minimum volume

$$\min R^2 + C\sum_{i=1}^{N}\xi_i$$

subject to

$$\|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i \text{ and } \xi_i \geq 0$$



Preliminary Results (1)

- Image sets:

	training set	test set
gun	104	105
passport	64	65
people	76	76
truck	97	97
negative-test-2500	N/A	2500
google-negative-train	120	N/A

Preliminary Results (2)

Table 8: ROC EER for C-SVM.

	PCA-SIFT			RG-SIFT		
Visual Dict Size	100	200	400	100	200	400
gun	16.2	**10.2**	11.9	10.5	10.5	**7.8**
passport	5.3	6.5	**5.1**	9.1	**8.1**	**8.1**
people	32.9	27.6	**23.0**	13.2	11.8	**10.5**
truck	19.8	**15.5**	17.2	13.4	14.4	**11.1**

Table 9: ROC EER for SVDD.

	PCA-SIFT			RG-SIFT		
Visual Dict Size	100	200	400	100	200	400
gun	25.7	**25.1**	30.5	**24.8**	**24.8**	**24.8**
passport	25.8	**24.0**	26.8	17.6	**17.5**	20.2
people	**32.9**	35.5	36.8	**14.5**	14.8	14.7
truck	**20.6**	21.6	21.6	18.8	**18.6**	**18.6**

Problems

- The SVM is trained against a small negative image set
  - The negative sample space is not well represented
- The positive samples are heterogeneous
  - Scattered positive samples may not form compact clusters in the sample space and leads to poor results for SVDD

Future Work

- A systematical evaluation of SVDD and SVM
  - Test different techniques to deal with the imbalanced datasets
  - Real-world image sets
- A new negative training samples selection mechanism:
  1. Train a SVDD classifier
  2. Use this SVDD model to classify all the negative training samples
  3. Select the negative samples near the boundary of SVDD
  4. Train SVM using the selected negative samples together with the give positive samples

Object Categorization using Web Image Sets (1)

- Motivation:
  - To alleviate the scalability problem in the conventional object categorization approaches
- Ideas
  - Given an object category name, such as airplane
  - Submit it to text-based image search engines, such as Google/Yahoo!/MSN Image Search
  - Download the images returned by these image search engines
  - Train the classifier after optional pre-processing

Object Categorization using Web Image Sets (2)

- Benefits of using Internet images
  - Easy to obtain training samples for *any object category*
  - *A large number* of images are available on the Internet
  - Search for category name in *multiple languages*
  - Lots of *non-image resources* are also available along with the images on the Web
    - Text surrounding the images
    - HTML structures such as hyperlink, image file name
    - Human-labeled tags, e.g., Flickr, Amazon
  - It can be done *automatically!*
    - API for search engines
    - Scripts for web-based applications

An Application

- Semantic robot vision challenge (2007, 2008)
  - Give a list of object category names to a robot
  - The robot queries the category names on the Internet and downloads relevant images
  - The robot then moves around the house, takes photos and searches for the objects in the captured images

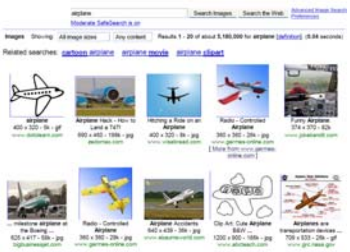
Univ. British Columbia

IMRA Europe

Univ. Maryland

Challenges (1)

- A large amount of noise (up to 85% [Fergus et al 2005]), i.e., images unrelated to the category of interest



Challenges (2)

- Imbalanced image sets
  - Easy to obtain positive samples
  - Difficult to obtain reliable negative samples
    - We are not given the negative object category names
    - Any category except the given one can be negative
    - One vs. the rest of the world
  - Problems
    - The negative space is too large to be properly sampled
    - A large negative training set will bias the classifier's decision boundary

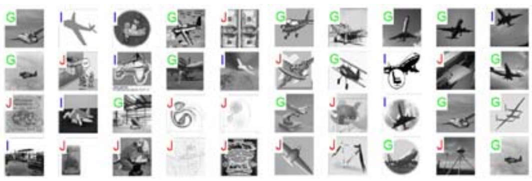
Proposed Approach

- Assumptions:
  - All positive samples are *alike*
  - Negative samples are different *in their own ways*
- Intuitions of Support Vector Data Description (SVDD)
  - Training classifiers using positive samples only
  - Enclose the dense cluster in training data with a sphere
  - Minimize the sphere volume to achieve better generability
  - Allow data points far away from the cluster center to be left outside of the sphere (i.e., outliers)
  - Use the boundary of the sphere as a classifier
    - New samples within the sphere – positive
    - New samples outside the sphere – negative

An illustrated example of SVDD

Experiment Results: Search Engine Improvement

- Top ranked images: airplane



Google raw images : 5 Good images

SVDD refined images (11 Good images)

Problems

- There could be more than one cluster in the positive samples
  - Homonymy
  - Sub-categories
- A solution is to cluster the dense clusters in the sample space and only send the selected images in the dense clusters to the classifier
  - Positive samples are alike *in some ways*
  - Negative samples are different *in their own ways*

Cluster Web Image

- Approach:
  - Cluster dense regions in the sample space of downloaded images while ignoring the rest
  - Use the images belonging to the dense clusters as positive examples to train the classifiers

[Jing and Baluja 2008]

BBC-S Algorithm

- BBC-S algorithm [Gupta and Ghosh 2008]:
  - Given the desired number of points in the clusters s
  - Iterate the three stages until convergence:
    - Each point is assigned to its nearest cluster representative;
    - Sort the distance from each point to its representative ascendingly and pick the first s points;
    - Update the cluster representatives, i.e., calculate the mean value of the data points for each cluster.
- K-Means is a special case of BBC-S
  - BBC-S = K-Means, if
    - $s = n$
    - Squared Euclidean distance

BubblePop Algorithm

- Initially use a large k
  - Improve the chance to include points near the dense regions
- Prune clusters ("bubble pops") if needed
  - Small clusters
  - Relatively sparse clusters
    - Compare the densest cluster and the sparsest cluster, if their density ratio is beyond a threshold, prune the sparsest cluster

Experiments

- Representative images of clusters produced by BubblePop from Web images

Summary

- A clustering approach to identify dense clusters in a noisy dataset
  - Automatically find the optimal number of clusters
  - Achieve more stable clustering results and eliminate the needs of multiple runs
- Applied to image classification using Web images
  - Reduce noise images
  - Improve classification performance
- Applied to Web image re-ranking
- Applied to visual summary of Web images for homonyms

Open Issues: Algorithms (1)

- BBC algorithm implicitly finds ball-shaped clusters in the given dataset, but in real-world applications, the data points may lies in more complex manifolds
  - Extend the BBC algorithm to work with geodesic distance
  - Extend the spectral cluster approaches to solve the "incomplete clustering" problem

[Ashlock and Kim 2005]

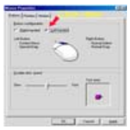
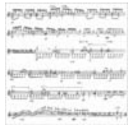
Open Issues: Algorithms (2)

- An efficient algorithms for BBC algorithms for large scale data set
  - Motivations:
    - In local feature clustering for visual words, often hundreds of thousands of local features are involved.
    - In video frame clustering, an one-hour video contains about 80 thousands of frames
    - In Web image clustering, there could be millions of images for a give topic
  - Potential solutions
    - Incremental BBC clustering
      - The state-of-the-art approach OPTIMAL [Li et al 2007] generate only one cluster from the Web images

Open Issues: Algorithms (3)

- Distractive images: human, face, document images, abstract images, screenshots, etc
  - If our goal is to classify general objects, these images of particular categories should be removed
  - Combine prior knowledge on distractive images within Web images into the framework of clustering






Summary

- Research contributions
  - An object categorization system for digital forensic image mining
  - An SVDD-based approach for learning object categories from Web images
  - A clustering algorithm, BubblePop, for the “incomplete clustering problem” with applications for
    - Web image clustering
    - Object categorization
    - Visual summary of homonyms

Beyond Image: Extensions to video(1)

- Retrieve video clips of a particular genre
  - Given: a set of video clips of a particular genre
    - e.g., news, football, etc
  - Goal: find video clips of this genre from a large video collection
  - Solution:
    - Formulate video retrieval as a problem of learning from imbalanced data set
      - Minority class: given class
      - Majority class: all negative classes
      - Employ techniques of learning from imbalanced data set to train the classifier

Beyond Image: Extensions to video(2)

- Detection recurring scenes in video
  - Example:
    - A anchor person in a news video
    - A room scene in an TV series
    - A pitcher in a baseball video
  - Solution:
    - Formulate it as a problem of near-duplicate image detection
      - recurring scenes form dense clusters
      - Non-recurring scenes scatter in the sample space
    - Apply BubblePop algorithms

Sports Video Summarization using Text Webcasts

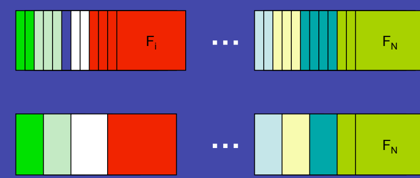
Mohammed Refaey
Wael Abd-Almageed

Video Summarization

A compact version of the video in term of a group of key-frames, key-shots, or key-scenes, which best semantically describe the contents of the underlying video.

video

Key-Framing

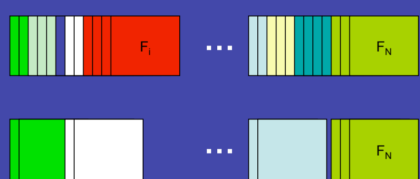


Video Summarization

A compact version of the video in term of a group of key-frames, key-shots, or key-scenes, which best semantically describe the contents of the underlying video.

video

Key Shots

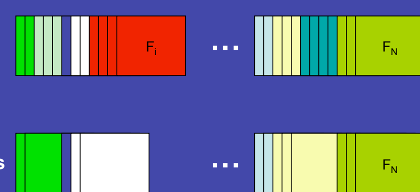


Video Summarization

A compact version of the video in term of a group of key-frames, key-shots, or key-scenes, which best semantically describe the contents of the underlying video.

video

Key Events



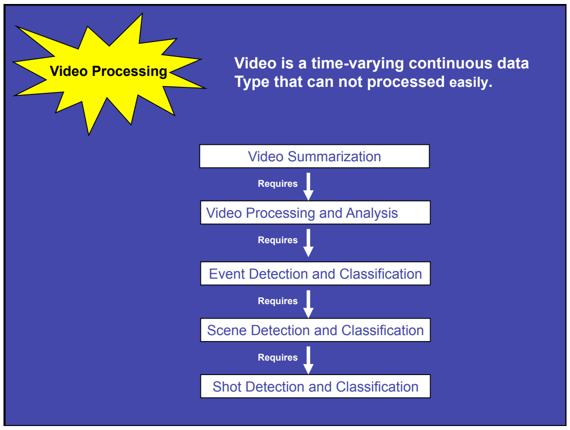
Summarization Applications

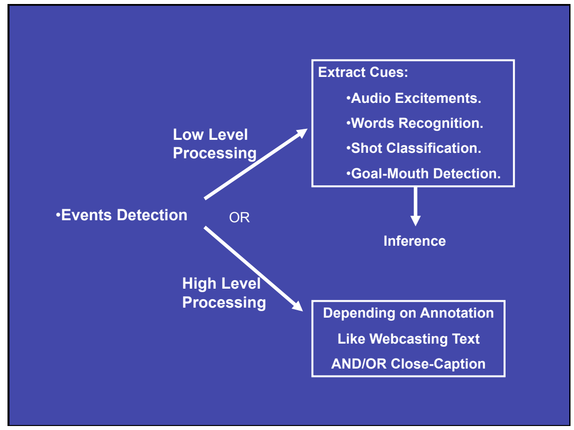
- Video Databases Browsing.
- Video On Demand.
- Video Compression.
- Video Indexing.
- Video Streaming over Limited-Bandwidth Internet.
- Broadcasting to PDAs
- Surveillance.
- Personal Video Recorders & TVs.

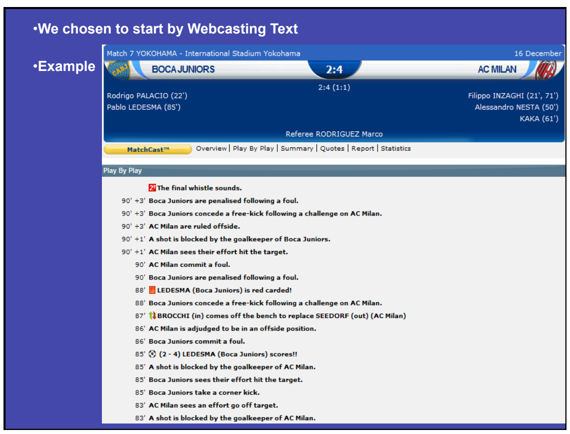
Domain

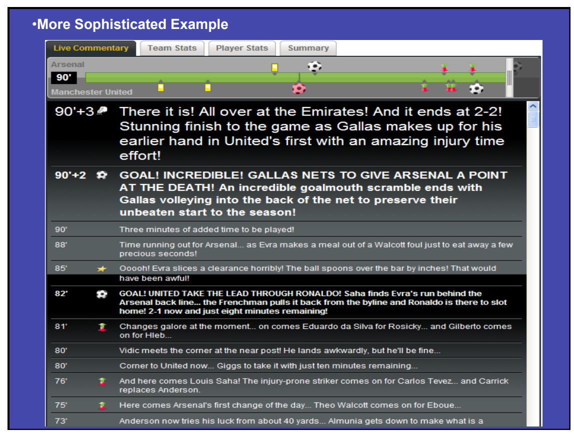
Our domain is the sports videos, why?

- Sports attracting many people.
- Different games are played at the same time, at different locations, with the case that the user cann't follow all of them.
- The user has no time to see the whole game in all competitions.
- The user's interest in the WHOLE game exponentially diminishes after the game is finished, he likes to watch the highlights only.

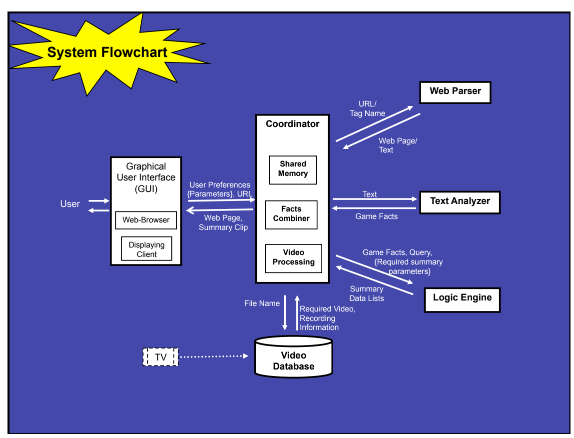

Video Processing
Video is a time-varying continuous data Type that can not processed easily.
Video Summarization
Requires
Video Processing and Analysis
Requires
Event Detection and Classification
Requires
Scene Detection and Classification
Requires
Shot Detection and Classification

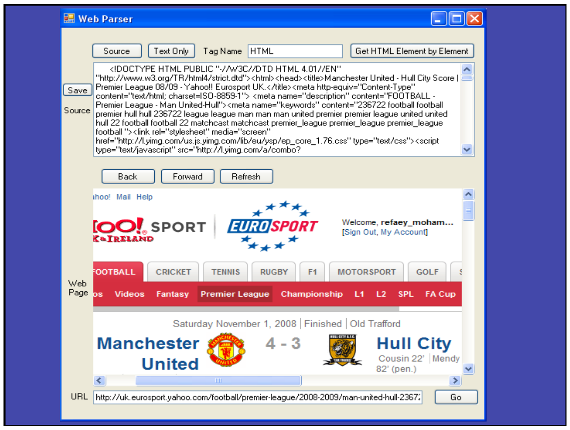

Extract Cues:
•Audio Excitements.
•Words Recognition.
•Shot Classification.
•Goal-Mouth Detection.
Low Level Processing
•Events Detection
OR
Inference
High Level Processing
Depending on Annotation Like Webcasting Text AND/OR Close-Caption

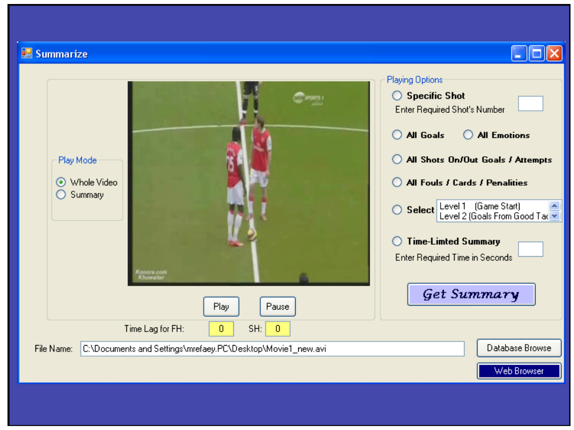

•We chosen to start by Webcasting Text
•Example
BOCA JUNIORS
2:4
AC MILAN
Play By Play

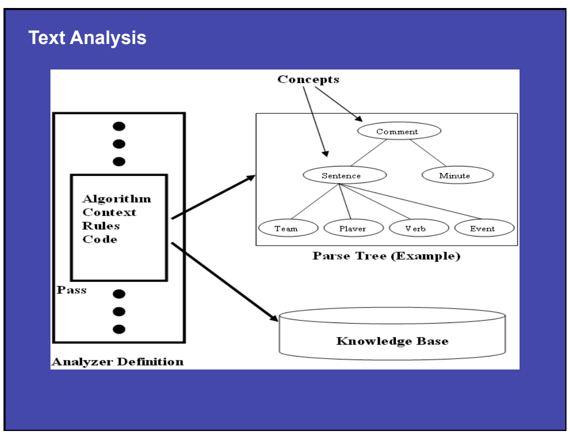

•More Sophisticated Example
Live Commentary
Team Stats
Player Stats
Summary
90'+3 There it is! All over at the Emirates! And it ends at 2-2! Stunning finish to the game as Gallas makes up for his earlier hand in United's first with an amazing injury time effort!
90'+2 GOAL! INCREDIBLE! GALLAS NETS TO GIVE ARSENAL A POINT AT THE DEATH! An incredible goalmouth scramble ends with Gallas volleying into the back of the net to preserve their unbeaten start to the season!


COURTSIDE LIVE
04:12
Suns
Mavericks
PLAY-BY-PLAY

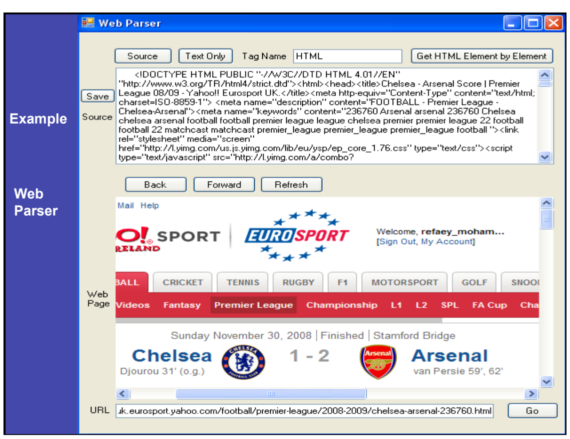

System Flowchart
Web Parser
URL/ Tag Name
Web Page/ Text
Coordinator
Shared Memory
Facts Combiner
Video Processing
Graphical User Interface (GUI)
Web-Browser
Displaying Client
User
User Preferences (Parameters), URL
Web Page, Summary Clip
Text
Game Facts
Text Analyzer
Game Facts, Query, (Required summary parameters)
Summary Data Lists
Logic Engine
File Name
Required Video, Recording Information
TV
Video Database

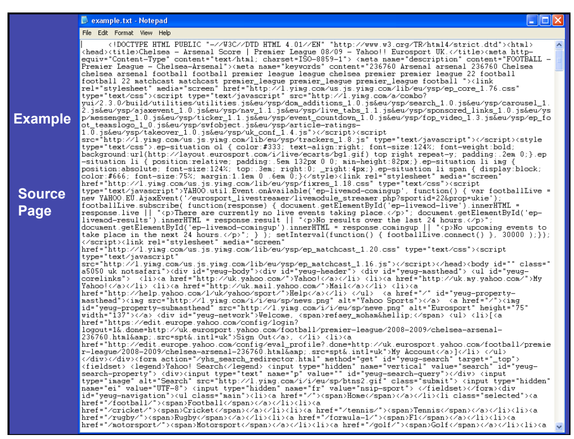

Web Parser
Source
Text Only
Tag Name
HTML
Get HTML Element by Element
Save
Source
Back
Forward
Refresh
Web Page
Manchester United
4 - 3
Hull City
URL
Go


Summarize
Play Mode
Whole Video
Summary
Playing Options
Specific Shot
Enter Required Shot's Number
All Goals
All Emotions
All Shots On/Out Goals / Attempts
All Fouls / Cards / Penalities
Select
Time-Limted Summary
Enter Required Time in Seconds
Get Summary
Play
Pause
Time Lag for FH:
SH:
File Name:
Database Browse
Web Browser


Text Analysis
Concepts
Comment
Sentence
Minute
Team
Player
Verb
Event
Algorithm
Context
Rules
Code
Pass
Parse Tree (Example)
Knowledge Base
Analyzer Definition


Example
Original
Web Page
Chelsea
1 - 2
Arsenal


Example
Web
Parser
Web Parser
Source
Text Only
Tag Name
HTML
Get HTML Element by Element
Save
Back
Forward
Refresh
Web Page
Chelsea
1 - 2
Arsenal
URL
Go


Example
Source
Page
example.txt - Notepad

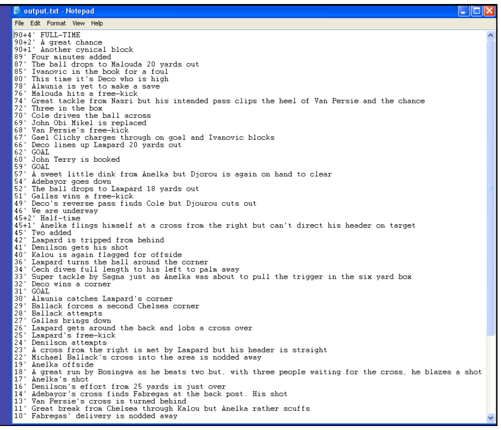
Example

Text Only

```
example.txt - Notepad
File Edit Format View Help
Yahoo! My Yahoo! Mail Help
  Welcome, refaey_moham
[Sign Out, My Account] Yahoo! Search     HomeFootballCricketTennisRugbyF1MotorsportGolfSnookerMore
Alpine Skiing
Athletics
Basketball
Boxing
CyclingExtreme Sports
Handball
Horse Racing
Ice Hockey
Mixed Martial ArtsNFL
Olympics
Poker
Sailing
SwimmingVolleyball
Winter sports
See all sportsYahoo! Sport
News
Photos
Videos
Message Boards
Fantasy SportsTV Guide
Eurosport Player
Blogs
Your opinion NewsPhotosVideosFantasyPremier LeagueChampionshipL1L2SPLFA CupChampions LeagueUEFA CupWorld
CupMore
League Cup
England
Transfers
Internationals
African Cup
Euro '08
Serie ALa Liga
Ligue 1
Bundesliga
Message Boards
Early Doors
Fantasy BlogFOOTBALL LiveSunday November 30, 2008 Finished Stamford Bridge  1 -  2 Chelsea
Djourou 31´ (o.g.) Arsenal
van Persie 59´, 62´
Van Persie wins it for Arsenal
Robin Van Persie struck twice in three second-half minutes as Arsenal came from behind to defeat Chelsea
2-1 at Stamford Bridge. MoreLive CommentaryTeam StatsPlayer StatsSummary
Time:90'
Chelsea Arsenal
90+4' FULL-TIME. Arsenal record their second successive victory over another member of the 'Big Four' and
they go fourth, seven points behind Chelsea. 90+2' A great chance to wrap things up but Denilson's first
touch is far too heavy and Cech pounces. 90+1' Another cynical block from Ivanovic, this time on Bendtner,
and he is lucky to escape a second yellow. 89' Four minutes added time to be played. 87' The ball drops to
Malouda 20 yards out and the Frenchman wafts a right foot at it. The result is a shot way over the
crossbar. 85' Ivanovic in the book for a foul on Clichy. 83' Nicklas Bendtner, in his pink boots, replaces
Adebayor. 81' Young Miroslav Stoch replaces Deco. 80' This time it's Deco who is high, wide and not very
```

Example

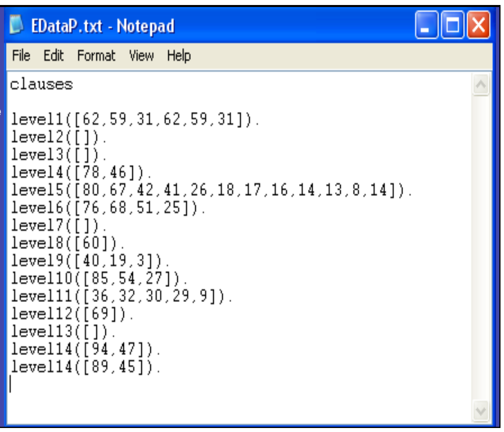
After Text Analysis



```
output.txt - Notepad
File Edit Format View Help
90+4' FULL-TIME
90+2' A great chance
90+1' Another cynical block
89' Four minutes added
87' The ball drops to Malouda 20 yards out
85' Ivanovic in the book for a foul
80' This time it's Deco who is high
78' Almunia is yet to make a save
76' Malouda hits a free-kick
74' Great tackle from Nasri but his intended pass clips the heel of Van Persie and the chance
72' Three in the box
70' Cole drives the ball across
69' John Obi Mikel is replaced
68' Van Persie's free-kick
67' Gael Clichy charges through on goal and Ivanovic blocks
66' Deco lines up Lampard 20 yards out
62' GOAL
60' John Terry is booked
59' GOAL
57' A sweet little dink from Anelka but Djorou is again on hand to clear
54' Adebayor goes down
52' The ball drops to Lampard 18 yards out
51' Gallas wins a free-kick
49' Deco's reverse pass finds Cole but Djourou cuts out
46' We are underway
45+2' Half-time
45+1' Anelka flings himself at a cross from the right but can't direct his header on target
45' Two added
42' Lampard is tripped from behind
41' Denilson gets his shot
40' Kalou is again flagged for offside
36' Lampard turns the ball around the corner
34' Cech dives full length to his left to palm away
33' Super tackle by Sagna just as Anelka was about to pull the trigger in the six yard box
32' Deco wins a corner
31' GOAL
30' Almunia catches Lampard's corner
29' Ballack forces a second Chelsea corner
28' Ballack attempts
27' Gallas brings down
26' Lampard gets around the back and lobs a cross over
25' Lampard's free-kick
24' Denilson attempts
23' A cross from the right is met by Lampard but his header is straight
22' Michael Ballack's cross into the area is nodded away
19' Anelka offside
18' A great run by Bosingwa as he beats two but, with three people waiting for the cross, he blazes a shot
17' Anelka's shot
16' Denilson's effort from 25 yards is just over
14' Adebayor's cross finds Fabregas at the back post. His shot
13' Van Persie's cross is turned behind
11' Great break from Chelsea through Kalou but Anelka rather scuffs
10' Fabregas' delivery is nodded away
```

Example

Game Facts



```
EDataP.txt - Notepad
File Edit Format View Help
clauses

level1([62,59,31,62,59,31]).
level2([]).
level3([]).
level4([78,46]).
level5([80,67,42,41,26,18,17,16,14,13,8,14]).
level6([76,68,51,25]).
level7([]).
level8([60]).
level9([40,19,3]).
level10([85,54,27]).
level11([36,32,30,29,9]).
level12([69]).
level13([]).
level14([94,47]).
level14([89,45]).
```

Shot Detection

Shot is:

A contiguous-group of frames continuously recorded by a single camera of an event without interruptions.

Shot Boundary Types :

Cut : is an abrupt change in the contents of the frames during transition from one shot to the next.

Fade In: Starts with a black (or a single color) Frame and lightens to gradually reveal an Image of new shot.

Fade Out: is the same effect but in reverse to Fade In.

Dissolve: is a Fade In/Fade Out with the two actions happening simultaneously.

Wipe: A pattern sweeps across the screen, progressively replacing the outgoing shot with the incoming one.

Proposed Method

We have developed an algorithm using the **Fuzzy Logic** to detect these transitions.

Shot classification

Is the process of assigning a label for each shot in the underlying video.

Shot Types

- Long: A large part of the field is appear and the players are appears so small entities.
- Close-Up: The camera is concentrated on one person. To show emotions…etc.
- Medium: as the Close-Up but the camera is more distant to the degree that the whole body of the person is appear . To show players skills.
- Out Of Field: to see something happen out of field or the fans.

Proposed Method

We have developed an algorithm using the **Neural Network** as a classification method.

- **Current Work:**
  - **Personalization**
- **Future Work:**
  - **Event Ranking**
  - **Event Localization**
  - **Extension to Closed-Caption**

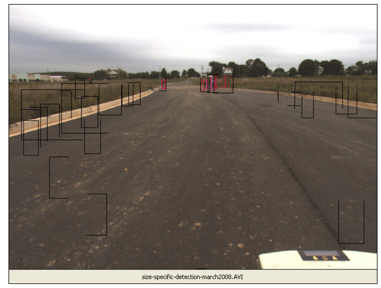
Processing Video Collections on GPU Arrays

Ramani Duraiswami

Kernel-Based Learning on Graphics Processors

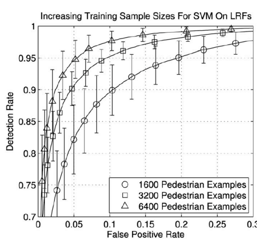
Mohamed Hussein
Wael Abd-Almageed

Human Detection in Videos



- Main stream research direction:
  - Can we find better classifiers?
  - Can we use better features?

Are We Using Enough Samples?



Munder and Gavrila, PAMI 2006

"Interestingly enough, classification errors are reduced by approximately a factor of two whenever the training sample size is doubled; no saturation effects are yet observed. Notice, furthermore, that the performance differences caused by increasing the number of training examples exceed the differences between different feature extraction methods."

Availability of Huge Labeled Datasets

Kernel-Based Methods

- Best published human detection techniques use SVM's [Dalal et. al. 2005, Felzenzswalb et. al. 2008]
- Other kernel-based methods are widely used in various machine learning problems
  - Classification, regression, clustering, dimensionality reduction
  - Examples: Kernel-PCA, Kernel-LDA, AP, GP, LLE, …

Challenge # 1: Kernel Matrix Size

- Given N samples, kernel matrix size is N x N
- Increasing N makes kernel matrix too large
  - Solution: create matrix elements on demand
  - Typically, computation of kernel values is the main bottleneck
- Goal # 1: Do not compute kernel values on demand
  - Use sparse representation instead

$$N \text{ samples}: x_1, x_2, \cdots, x_N$$

$$\begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1N} \\ k_{21} & & & \\ & & \ddots & \vdots \\ \vdots & & & \\ k_{N1} & & \cdots & k_{NN} \end{bmatrix}$$

$$k_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$$

Challenge # 2: Computational Time

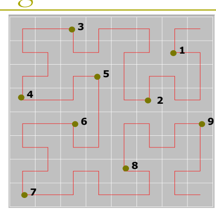
- Aside from kernel matrix formation, rest of computation typically is $O(N^k)$, where k is 2-3
- Goal # 2: Use parallelization
  - Speed computation of kernel values
  - Speed other computations based on it as well

Graphic Processing Units

- Why GPUs
  - Relatively cheap and widely available
  - Massively parallel devices, Currently, up to 240 core processor on a single card
  - Programmable for general purposes
- Limitations:
  - Limited memory size (currently 1GB at most)
    - Using sparse representations can take care of this
  - Good performance requires regular memory access pattern
    - Conventional sparse representation does not satisfy this

Space Filling Curves

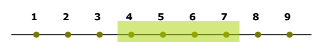
Maps points from N-D to 1-D



Good locality preserving properties

Orders on the curve	1	2	3	4	5	6	7	8	9
Indexes on the curve	2	9	19	26	30	35	42	51	58

Band-Limited Kernel Matrix



Using a fixed neighborhood *r* on the 1-D space

$$N \text{ samples}: x_1, x_2, \cdots, x$$

$$\begin{bmatrix} k_{1,1} & \cdots & k_{1,r} & 0 & \cdots & 0 \\ \vdots & k_{2,2} & & k_{2,r+1} & \ddots & \vdots \\ k_{r,1} & & & & \ddots & 0 \\ 0 & k_{r+1,2} & & \ddots & & k_{N-r+1,N} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & k_{N,N-r+1} & \cdots & k_{N,N} \end{bmatrix}$$

$$k_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$$

Kernel Matrix Computation on GPUs

Applications

- Any application that uses kernel methods for clustering, classification, regression, or dimensionality reduction can benefit from this approach
- Examples
  - Object classification: e.g. human, face, animal, car, train, flag, …
  - Background classification: e.g. indoors vs. outdoors, urban vs. rural, summer vs. winter, …
  - Activity Recognition: e.g. action in sports videos

Understanding Videos, Constructing Plots

Abhinav Gupta

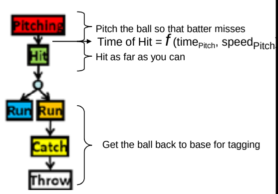
Collaborators:
Praveen Srinivasan (**Univ. of Penn**), Jianbo Shi (**Univ. of Penn**)
Larry S. Davis **(Univ. of Maryland)**

Inference from Videos

- Car swerve, Pedestrian comes on road
- Car swerves to save the pedestrian comes on the road

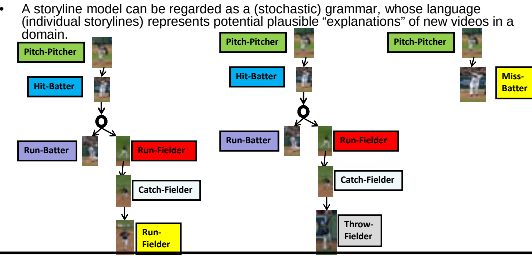
Beyond Action Recognition

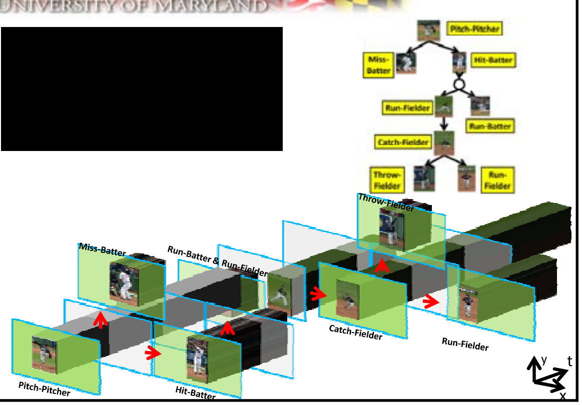
- Inferences from videos involves not only action recognition but also extracting intentions and inferring causalities.
  - Useful for video mining specially unusual event detections (based on intentions of the actor)
- Understanding the causal relationships among them provides information about the semantic meaning of the activity in video (Storyline)
  - The entire set of actions is greater than the sum of the individual actions.
- Storyline of a video captures the set of actions in the video and the causal relationships between those actions
  - Serves as contextual model for recognition of individual events.



Storyline Model

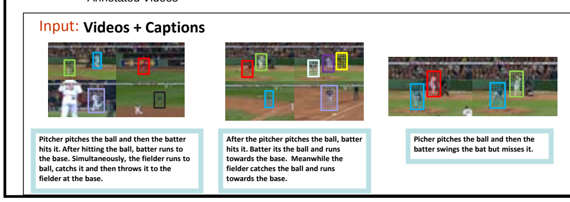
- A model that represents the set of storylines that can occur in a video corpus and the general causal relationships amongst actions in the video corpus is referred to as a "storyline model".
- Storyline models also indicate the agents likely to perform various actions and the visual appearance of actions.
- A storyline model can be regarded as a (stochastic) grammar, whose language (individual storylines) represents potential plausible "explanations" of new videos in a domain.





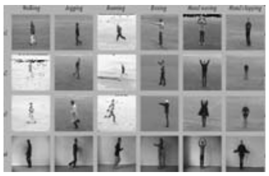
Learning Storyline Model

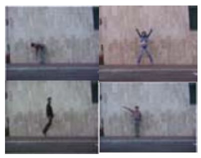
- Storyline model can be hand-specified by domain expert.
  - Cumbersome and sometimes Impossible.
- Learning Storyline Model involves
  - Learning Appearances of Actions
  - Learning Causalities between Actions
- Interested in learning storyline models from language and visual data.
  - Videos with Closed Captioning
  - Annotated Videos

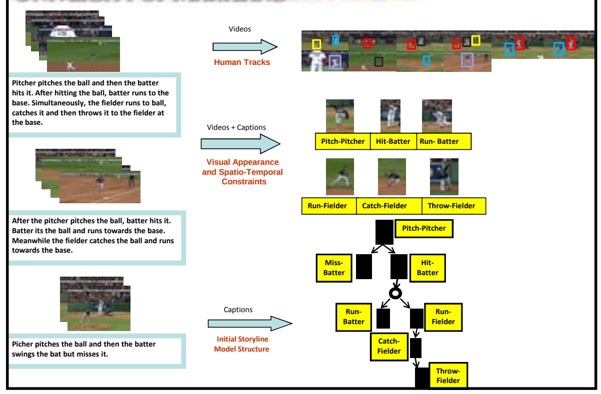


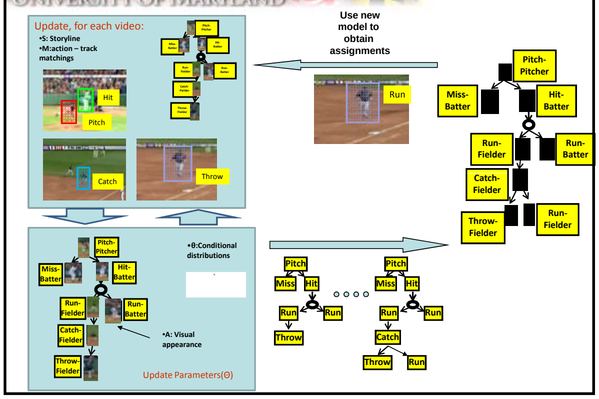
Learning Actions from Videos – Past Work

- Current approaches assume already segmented (space and time) actions in training datasets.
- Obtaining manual segmentations of training videos is a time consuming process.
- Weakly supervised learning involves using co-occurrence of visual features and verbs to learn appearance models
  - **Ignore the causal structure of videos**
  - **Markov Models used are computationally expensive**

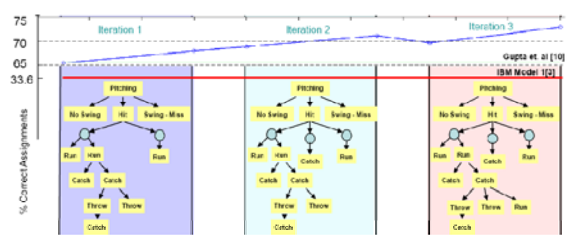








Learning Performance



Inferences with Storyline Model

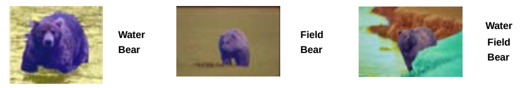
- Inference for a new video involves
  - Predicting storyline
  - Labeling human actions in the videos
- We formulate an integer-programming based approach which selects the storyline and labels actions simultaneously.

Results – Text + Storylines

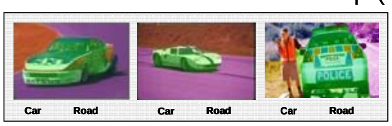
Results – Text + Storylines

Weakly Labeled Datasets

- Extract Nouns and Verbs from captions and co-occurrence to solve the correspondence problem.



Co-occurrence Relationship (Problems)

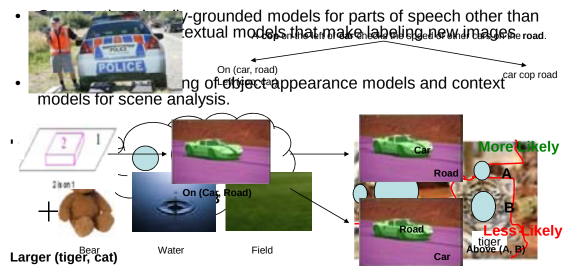


Hypothesis 1

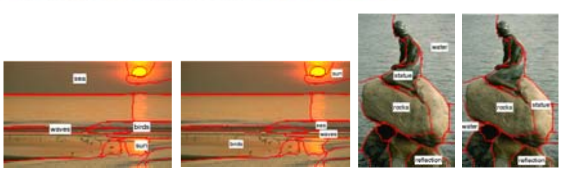
Hypothesis 2

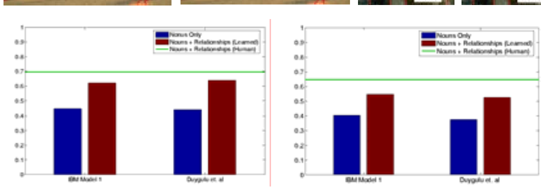
Beyond Nouns: Exploiting Prepositions and Comparative Adjectives

- Richer linguistic descriptions of images makes learning of object appearance models from weakly labeled images more reliable.
- …y-grounded models for parts of speech other than …extual models that make labeling new images…
- …ng of object appearance models and context models for scene analysis.



Gupta and Davis, ECCV 2008

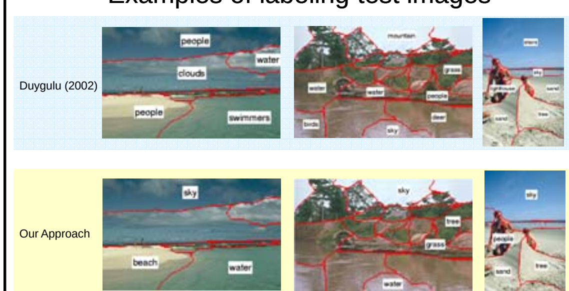




(a) Frequency Correct

(b) Semantic Range

Examples of labeling test images



UNIVERSITY OF MARYLAND

Conclusions

- Inferences based on causality and intentions are useful for data-mining and semantic understanding.
- Storyline model represent semantic structures which are used as a generative model for both linguistic descriptions and videos.
- Simultaneous learning of storyline-model and action appearance models lead to better performance as it harnesses structure in the videos and co-occurrences.
- Simultaneous inferences of storyline and actions in a video leads to better semantic understanding and better action labeling performance.

Automatic Annotation of YouTube-like Video Exploiting Online Communities

Automatic Annotation of YouTube-like Video

- 200,000 videos uploaded to YouTube daily
- New videos
  - No semantic information
  - Not searchable (except for file name) until viewers add tags and comments

Automatic Annotation of YouTube-like Video

- Low-level features are not enough for extracting semantic information
  - especially home-made videos
- Object detectors/classifiers
  - Inefficient (too much search, too many scales ,etc.)
  - Cannot have a detector for *everything*

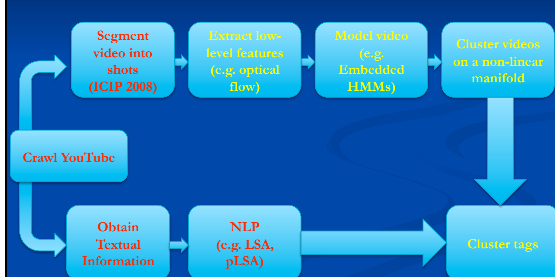
Automatic Annotation of YouTube-like Video

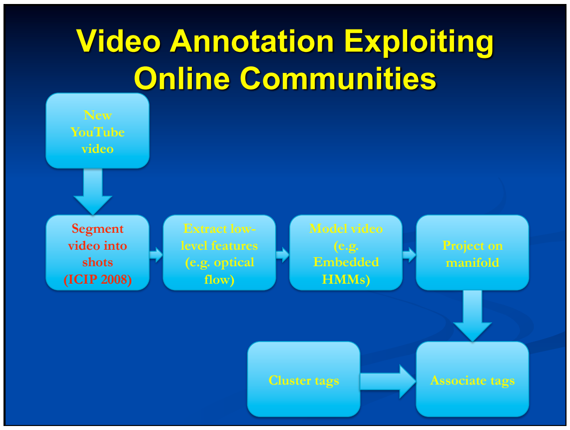
- Why is it so important national security?
  - The Finland shooter uploaded a video to YouTube the night before the shooting
- Automatic video tagging system

Automatic Annotation of YouTube-like Video

- Objective
  - Automatic video tagging system for YouTube-like videos
- Research does not exploit enormous amounts of data on online communities and social networks
  - Visual (YouTube, Yahoo Videos, Flicker)
  - Textual (User comments, user tags, etc.)

Video Annotation Exploiting Online Communities




Video Annotation Exploiting Online Communities
New YouTube video
Segment video into shots (ICIP 2008)
Extract low-level features (e.g. optical flow)
Model video (e.g. Embedded HMMs)
Project on manifold
Cluster tags
Associate tags

Fundamental Research

- Support for Outside Activities
  - MadCat, Bobcat
  - Library of Congress
  - Document Similarity of "record data" with pictures
- Page Segmentation and Line Detection
- Image Enhancement and Clutter Removal
- Document Partitioning and Reflow
- Revising of Document Image Classification
  - Genre?
  - Indexing and Retrieval

Potential Research Tasks

- Document Evaluation Repository and Server
  - Access to datasets and annotations
  - Collaborative annotation efforts
  - Public Release of DocLib
  - Support for Evaluations including development and test sets
  - Historic archives of evaluation results for comparison

Video

- Video Capture/Format Classification
  - Speech/Lecture
  - Event
  - News Cast (sports, weather, etc)
- Video Processing on Clusters/GPUs
- Transition of existing capabilities for
  - Segmentation
  - Text Detection