Indexing and Retrieval of Handwritten Document Images

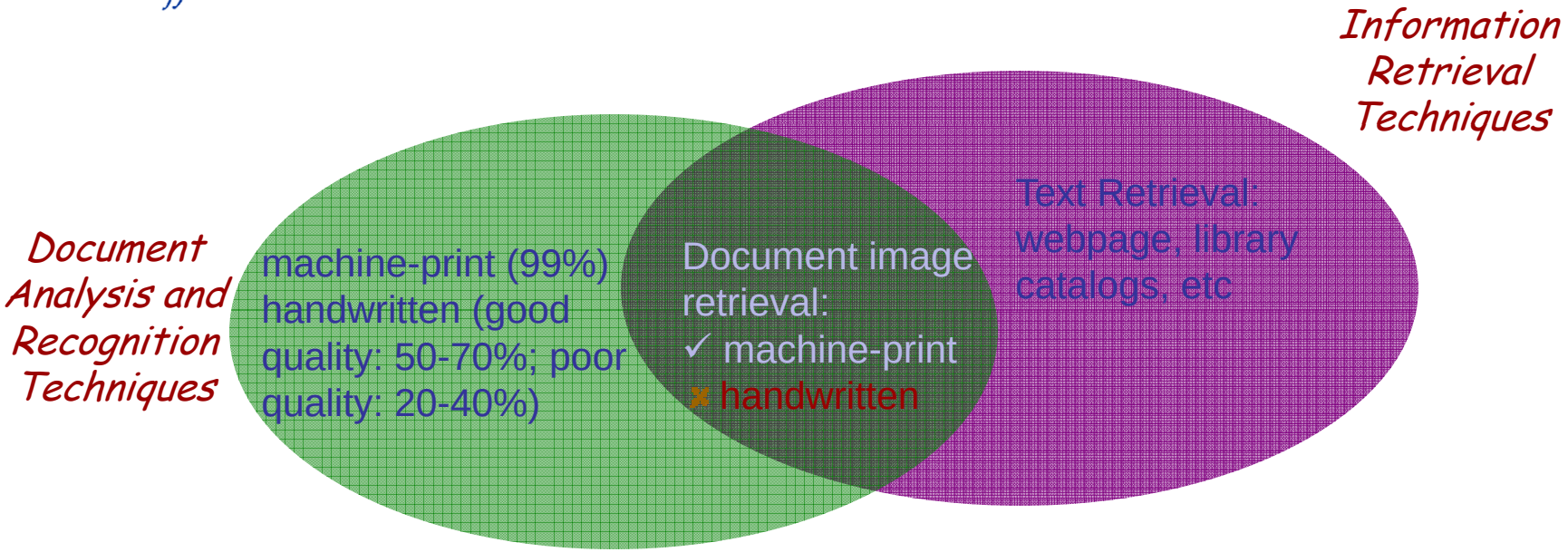
Huaigu Cao

hcao3@cubs.buffalo.edu

Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Problem Statement



- Search engine for handwritten document images performs poorly because of the low recognition rate
- Objective: Improve IR performance on handwritten data, especially that of poor quality

Document Retrieval – An Overview

- The similarity between a document and a query can be roughly measured by the number of occurrences of the query terms in the document, for example:

$$\{d_j\} = \{d_1 = \text{"pt has a trauma"}, d_2 = \text{"pt has breath difficulty"}\}$$

$$\{t_i\} = \{t_1 = \text{"pt"}, t_2 = \text{"has"}, t_3 = \text{"a"}, t_4 = \text{"trauma"}, t_5 = \text{"breath"}, t_6 = \text{"difficulty"}\}$$

$$freq_{i,j} = \begin{bmatrix} 1,1,1,1,0,0 \\ 1,1,0,0,1,1 \end{bmatrix} \qquad q = \text{"breath difficulty"}, qtf_{i,q} = [0,0,0,0,1,1]$$

- The frequency $freq_{i,j}$ is the most crucial for a document retrieval system

Estimating $freq_{i,j}$ from Noisy Handwriting Recognition Results

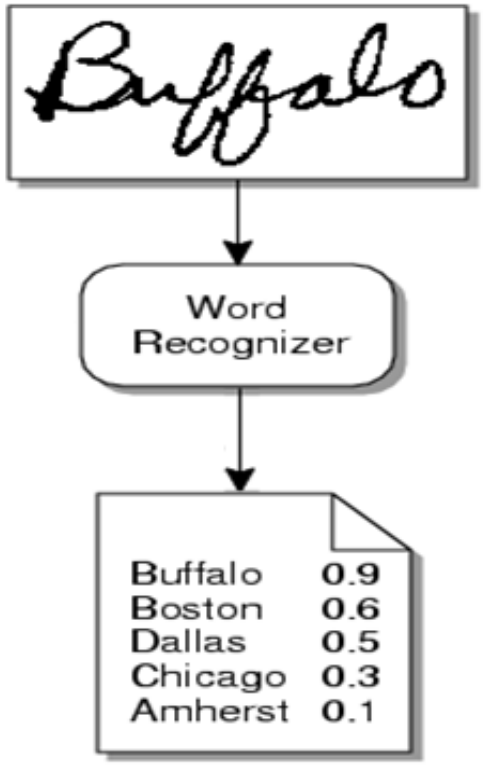
- Index built on top-*n* Word Recognition Results [Lee, ICDAR05]
- Assigning different weights to candidates

$$\hat{freq}_{i,j} = \sum_k \Pr(t_i \mid w_k) \quad \text{[Rath04]}$$

or $$\hat{freq}_{i,j} = \sum_k \frac{Const}{\mathrm{RANK}(t_i, w_k)} \quad \text{[Howe05]}$$

$w_1, w_2 \ldots$: word images in document d_j

$\Pr(t_i \mid w_k)$: word recognition posterior probability



- Perfect word segmentation?

Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Improved Estimation of $freq_{i,j}$ Incorporating Word Segmentation and Language Models [Our Approach]

$$\hat{freq}_{i,j} = \sum_{\vec{w}} \Pr(\vec{w} \mid \vec{o}) \cdot \sum_{\vec{\tau}} \Pr(\vec{\tau} \mid \vec{w}) \cdot freq_{i,\vec{\tau}}$$

$\vec{o} = [o_1 o_2 ... o_T]$: observation series

$\vec{w} = [w_1 w_2 ... w_L]$: word images

$\vec{\tau} = [\tau_1 \tau_2 ... \tau_L]$: terms (OCR results)

$\Pr(\vec{w} \mid \vec{o})$: word sequence segmentation probability

$\Pr(\vec{\tau} \mid \vec{w})$: word sequence recognition probability

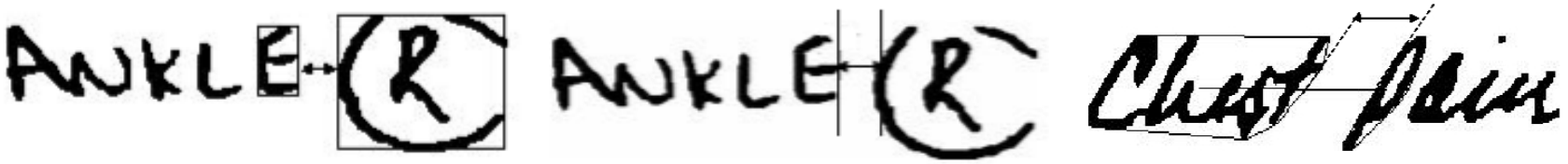
$freq_{i,\vec{\tau}}$: number of t_i in $\vec{\tau}$

Solved by *dynamic programming*

Word Gap Labeling

Features for each gap between two CC's

- fv_1: Euclidian distance between bounding boxes
- fv_2: Shortest white run length between two CC's
- fv_3: Distance between the convex hulls



By Bayes' rule

$$\Pr(Valid \mid \vec{fv}) = \frac{\Pr(Valid)\, p(\vec{fv} \mid Valid)}{\Pr(Valid)\, p(\vec{fv} \mid Valid) + \Pr(Non-valid)\, p(\vec{fv} \mid Non-valid)}$$

The likelihoods are estimated non-parametrically using *Parzen window*

Word Spotting Using Word Gap Labeling

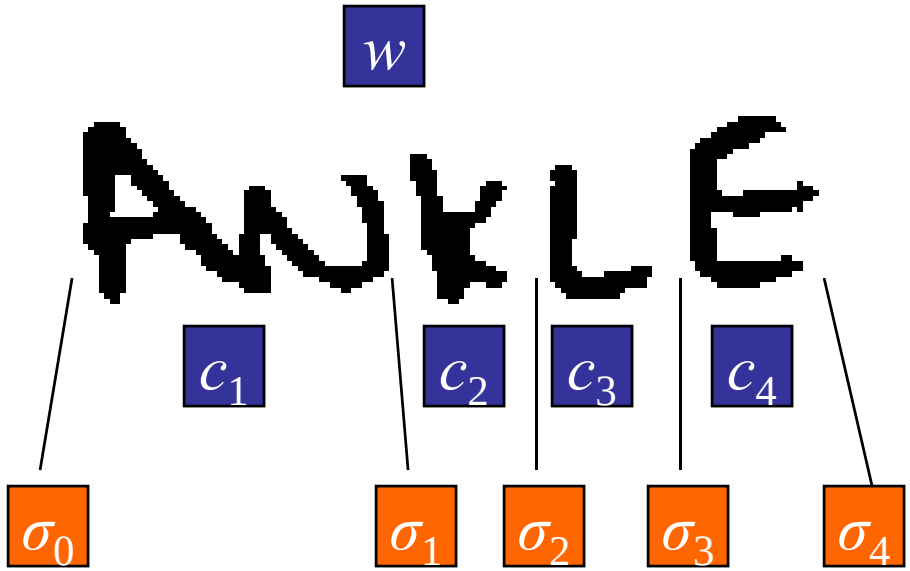
Word image (w) is representedby connected components :

$$w \rightarrow c_i, c_{i+1}, \ldots, c_j$$

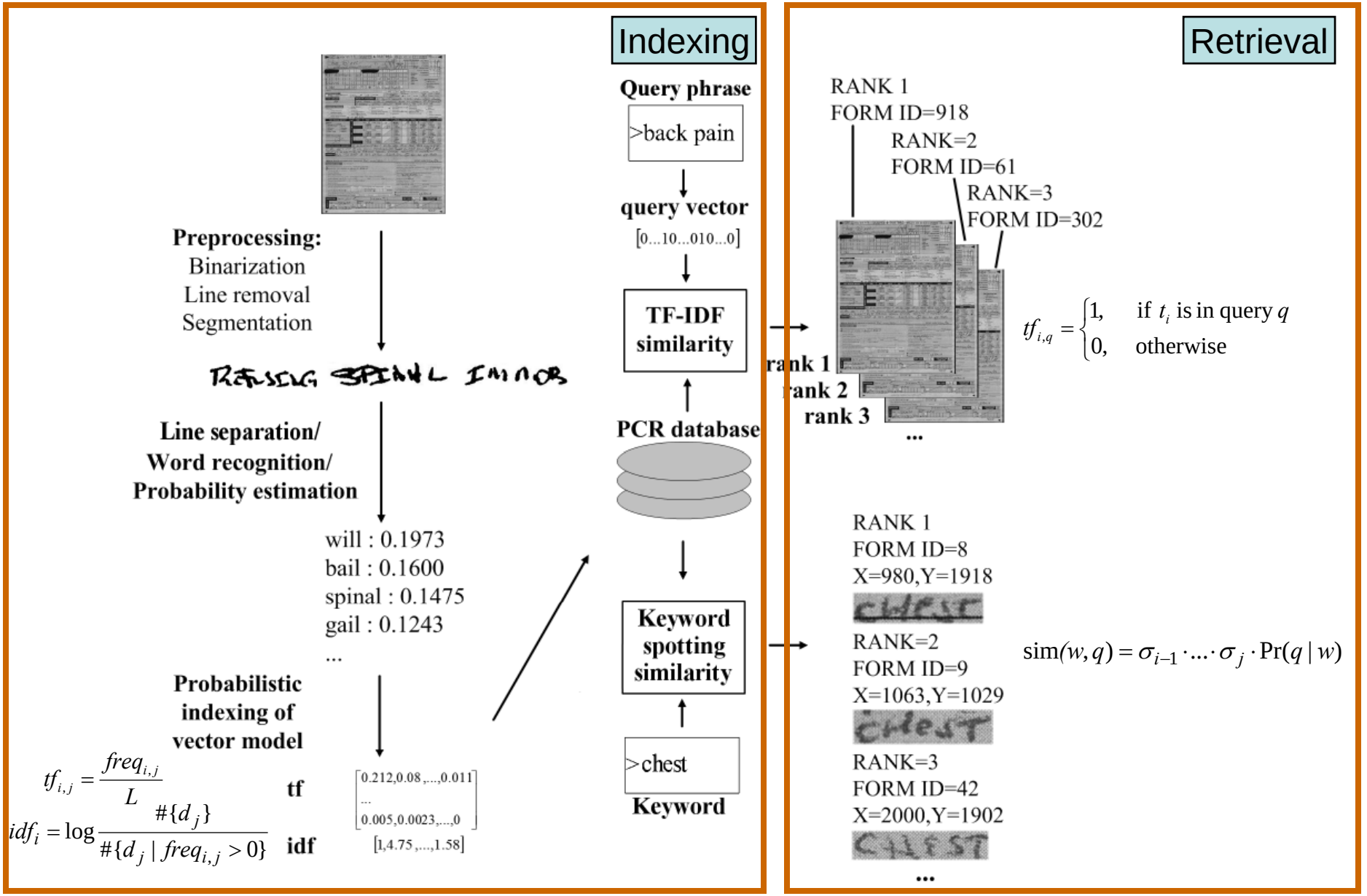
Word gap probability (σ_k) :
The probability of the gap between c_k and c_{k+1} being a word gap

Word - query similarity :

$$\text{sim}(w, q) = \sigma_{i-1} \cdot \ldots \cdot \sigma_j \cdot \Pr(q \mid w)$$



Search Engine Diagram



$$tf_{i,j} = \frac{freq_{i,j}}{L}$$

$$idf_i = \log \frac{\#\{d_j\}}{\#\{d_j \mid freq_{i,j} > 0\}}$$

$$tf_{i,q} = \begin{cases} 1, & \text{if } t_i \text{ is in query } q \\ 0, & \text{otherwise} \end{cases}$$

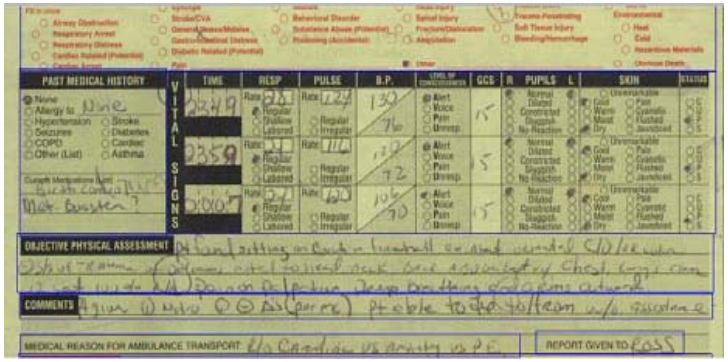
$$\text{sim}(w,q) = \sigma_{i-1} \cdot \ldots \cdot \sigma_j \cdot \Pr(q \mid w)$$

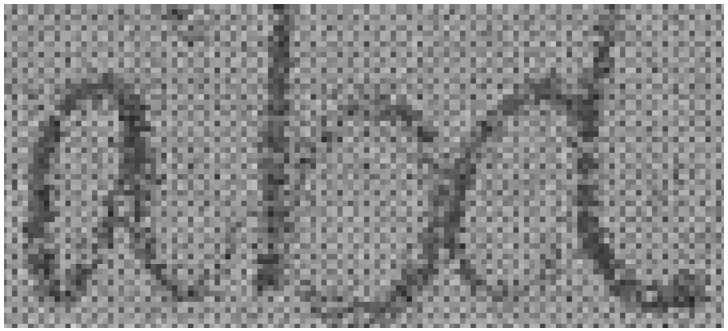
Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Data Set

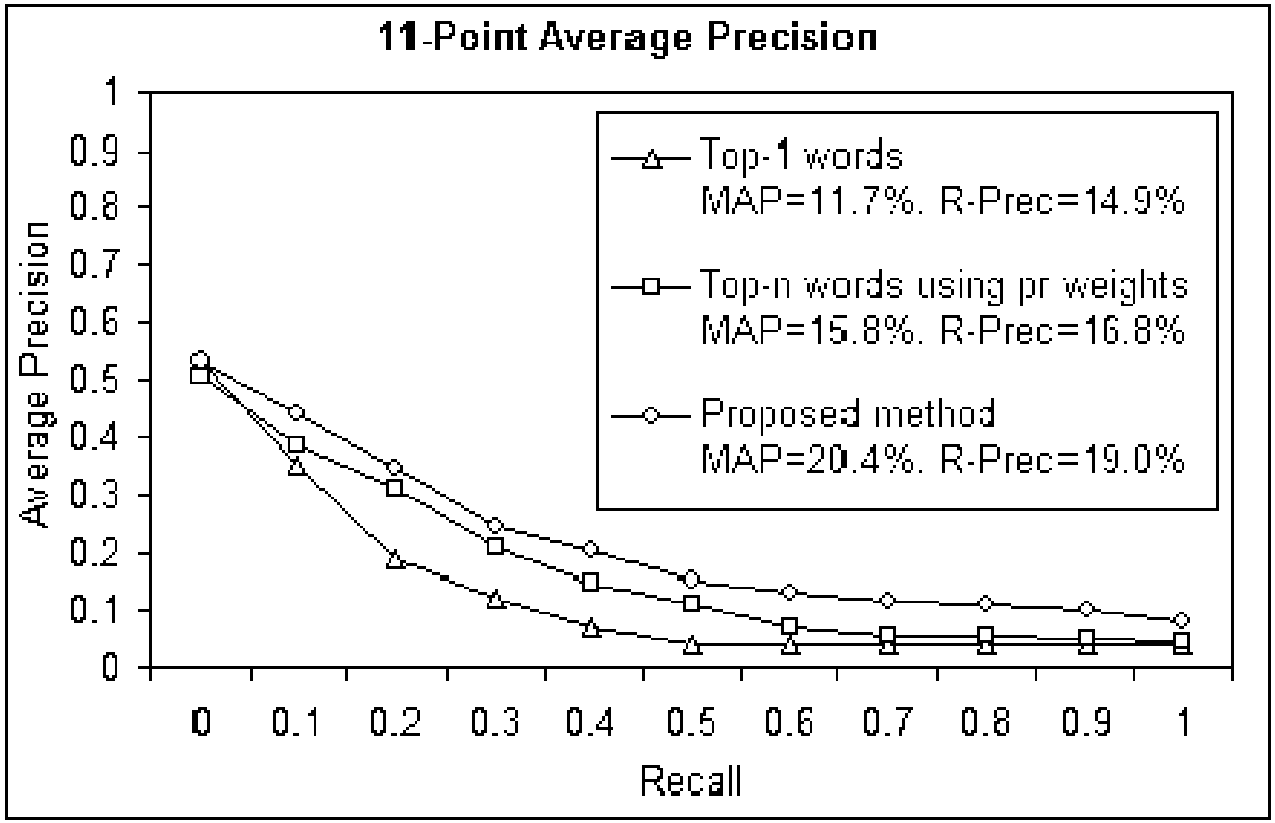
- Handwritten document image data: Pre-hospital Care Report (PCR) forms (written on carbonized paper, word recognition accuracy is typically <20%)





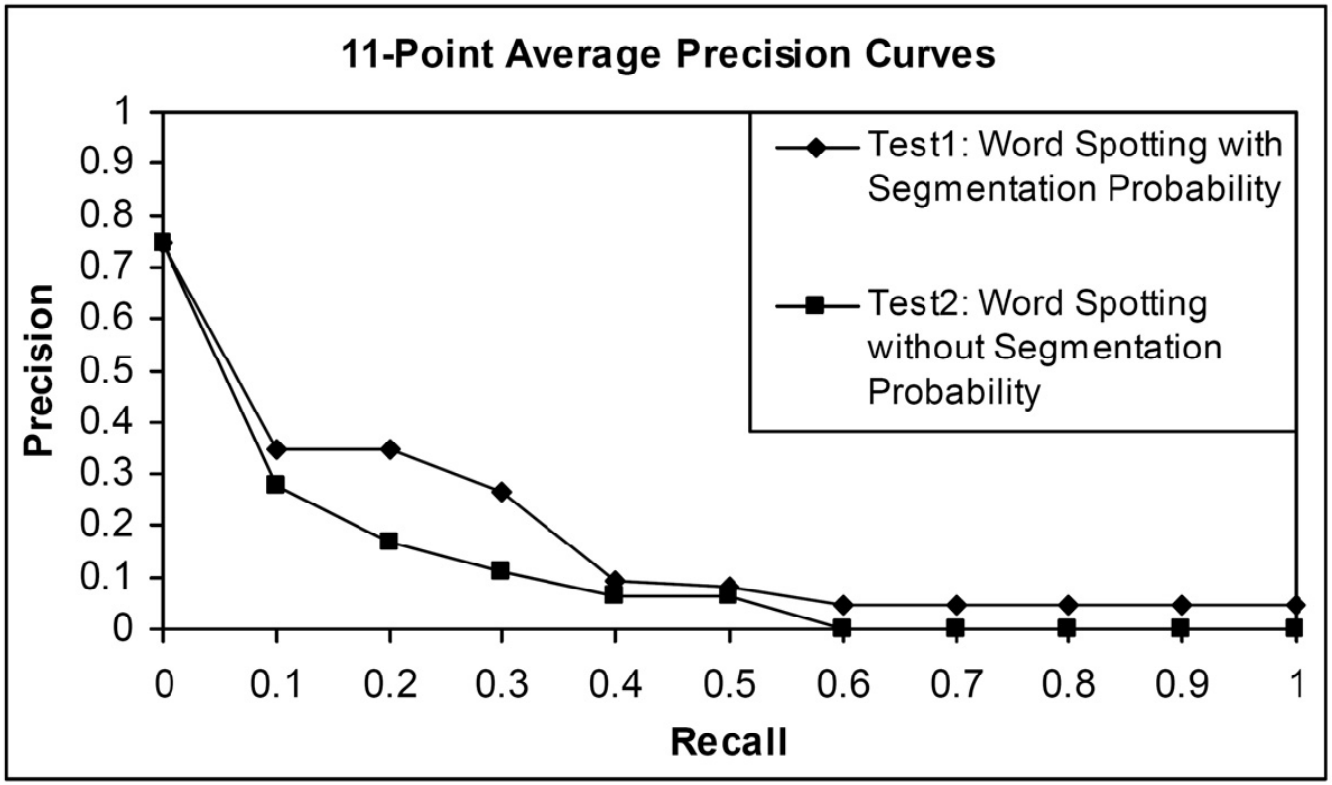
Document Retrieval Performance: Recall – Precision

342 Documents; 28 Queries



Word Spotting Performance: Recall – Precision

342 Documents; 33 Queries



Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Future Work

- Investigation of other high-level applications using the ranked OCR output
  - Machine translation for the handwritten documents
  - Handwritten document categorization