

Document Image Retrieval using Signatures as Queries

Sargur Srihari, Shravya Shetty, Harish Srinivasan,
Siyuan Chen, Chen Huang

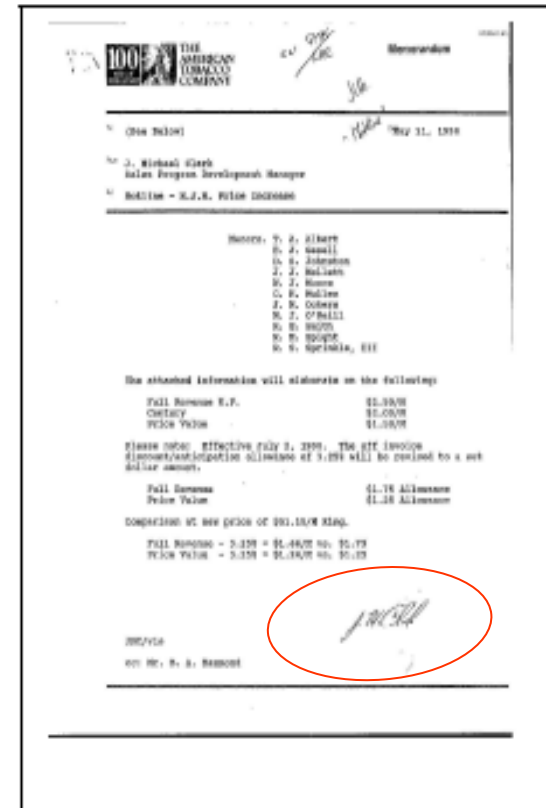
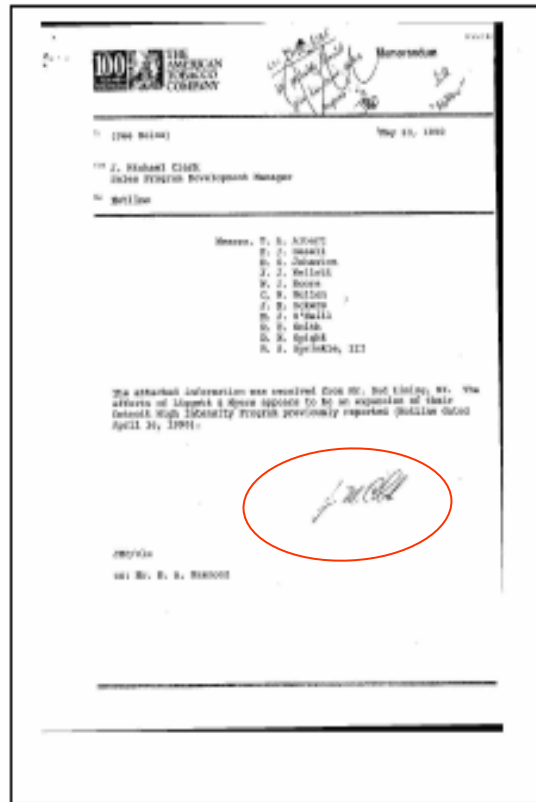
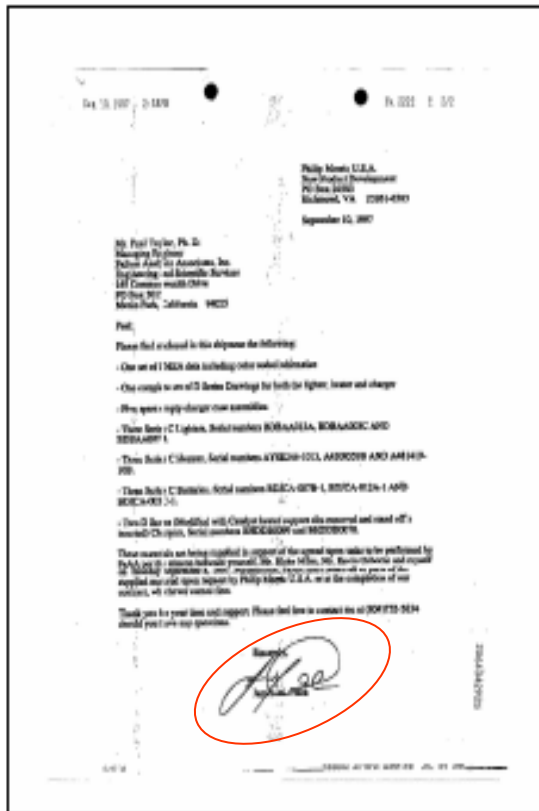
CEDAR, Department of Computer Science and Engineering
University at Buffalo, State University of New York

Gady Agam, Ophir Frieder
Illinois Institute of Technology

USA

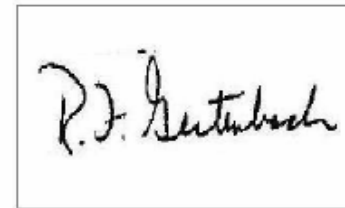
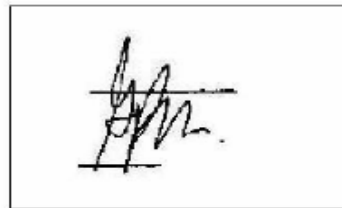
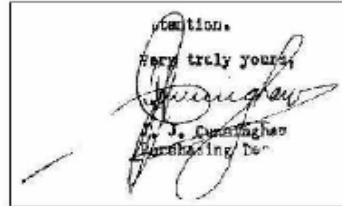
Signature-based Document Retrieval

Search business document archive based on signature image queries



Sample Queries: Cropped Signatures

Samples for different writers



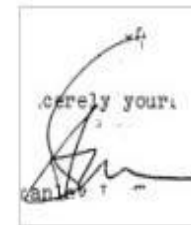
Samples for same writer



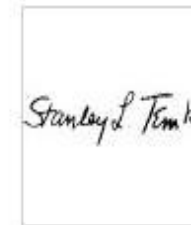
a



b



c



d



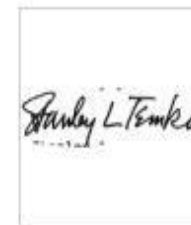
e



f



g



h

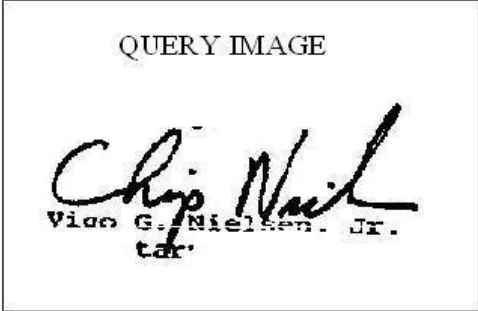
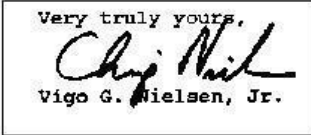
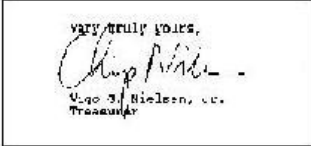
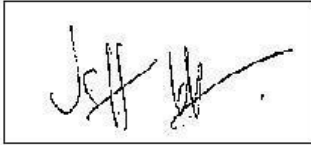
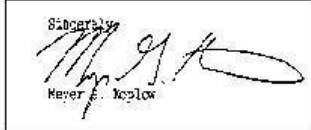



i

Signature Retrieval

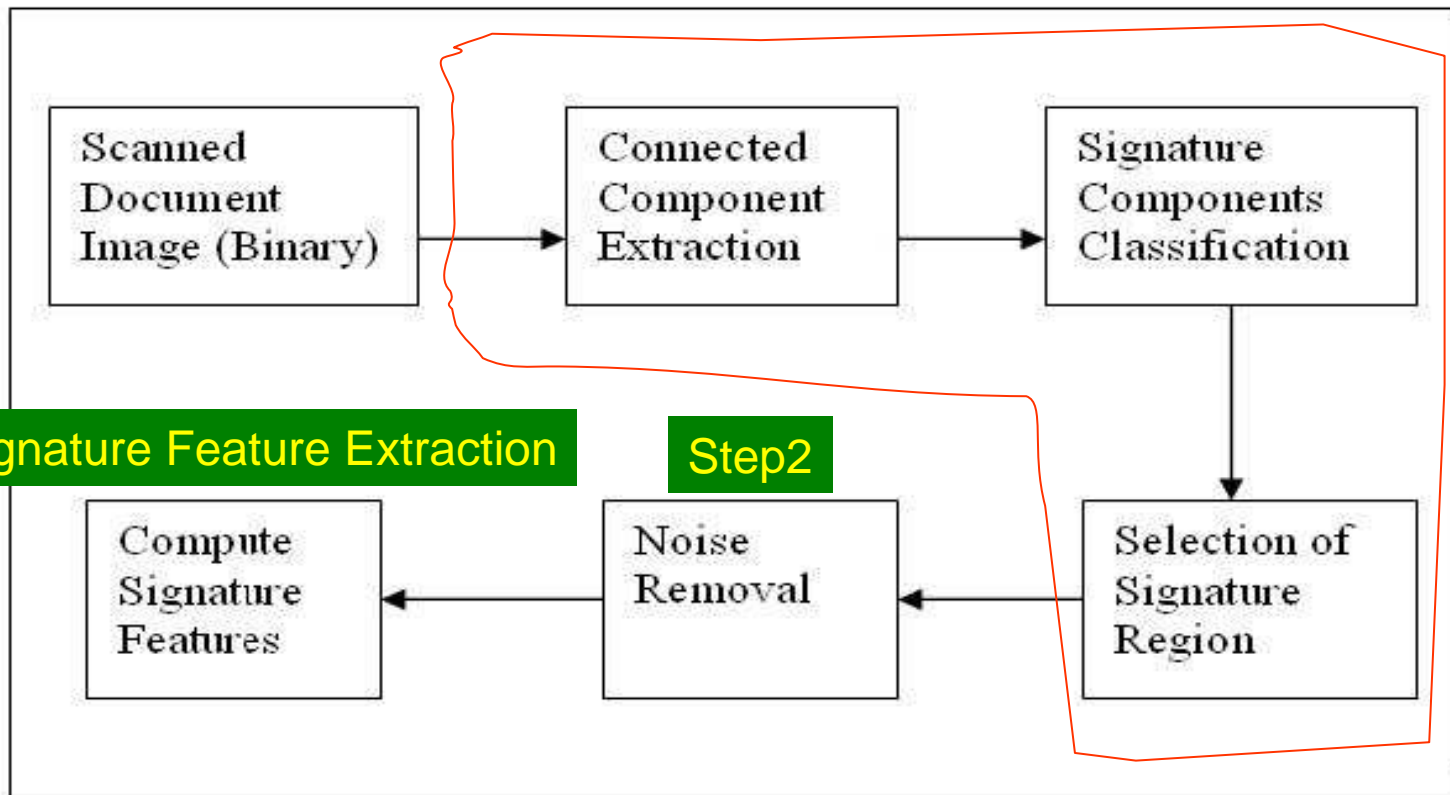
Retrieval

- Rank documents based on similarity to query signature

	RETRIEVAL		
	<u>Signature Snippets</u>	<u>Dissimilarity Measure</u>	<u>Document ID</u>
		0.26	N_3
		0.29	N_5
		0.39	I_5
		0.43	K_2
		0.49	L_1

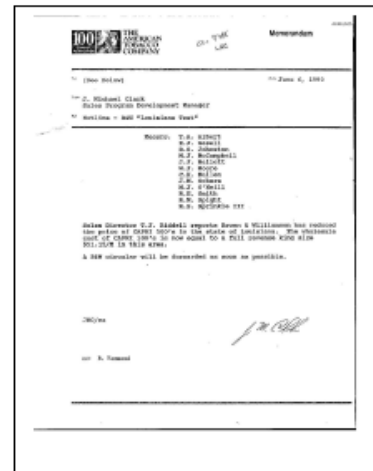
Indexing Documents

Step1: Signature Block Extraction

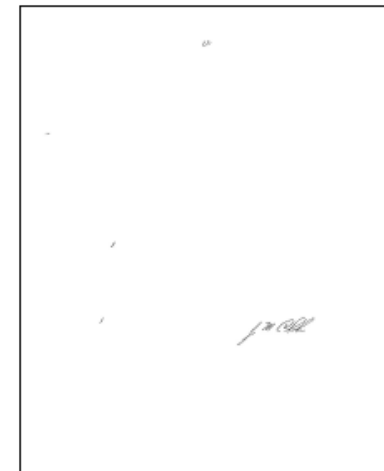


Step 1: Signature Block Extraction

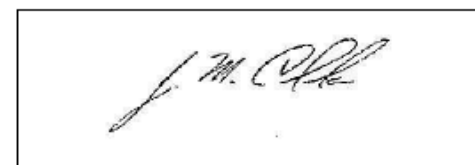
- Connected component analysis
- Component Feature Extraction
 - to identify logos, noise – Horizontal Run, Horizontal & Vertical Profile, Density, Size
 - to distinguish handwriting from print
 - Slope, Stroke Orientation (Gabor Filter), Density, Size
 - to distinguish handwritten text from signatures
 - Maxima and Minima count, Height variation, Slope, Size
- Classification using SVM
- Signature Block Identification
 - Large signature components are merged with other neighboring possible signature components
 - Overlapping blocks are merged



(a) Original Document



(b) Processed Document after Classification of Signature Components

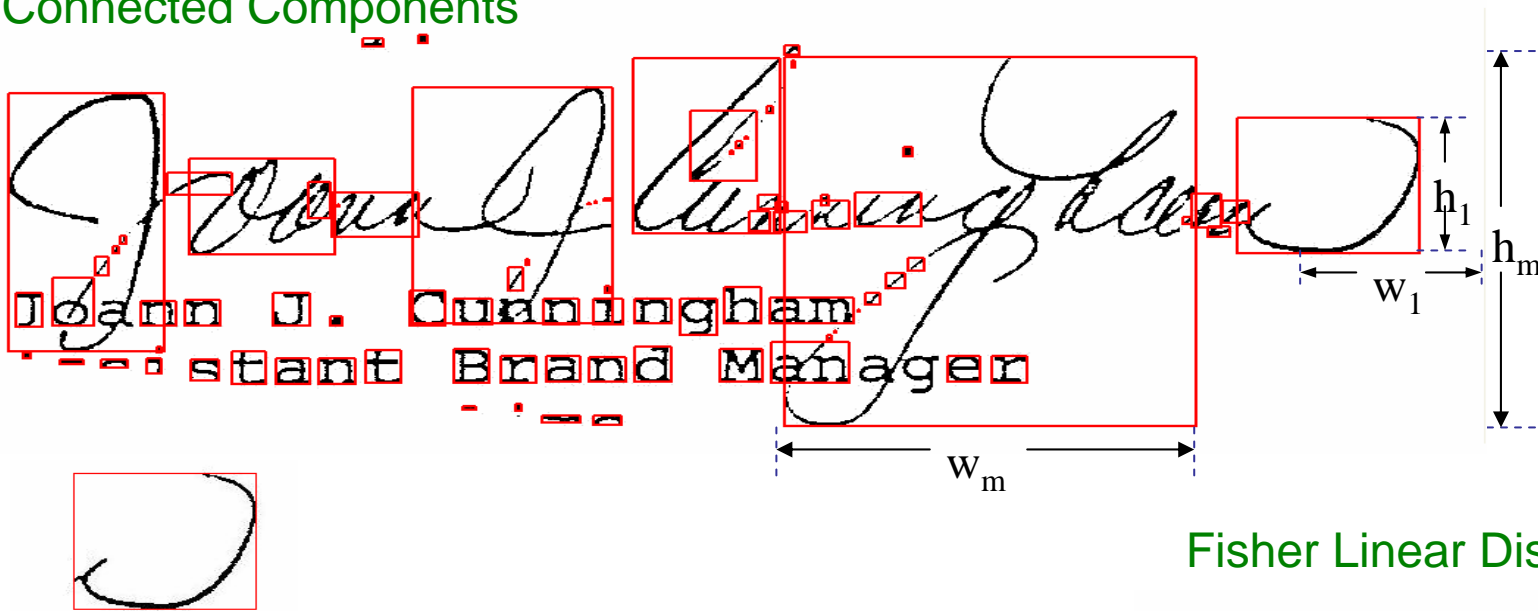


(c) Extracted Signature

In 92% (276/300) of cases resulting block contained most of the signature

Step 2: Noise Removal

Connected Components



Feature 1 =

relative contour size

$$= (h_1 + w_1) / (h_m + w_m) = 0.4034$$

Feature 2 =

aspect ratio =

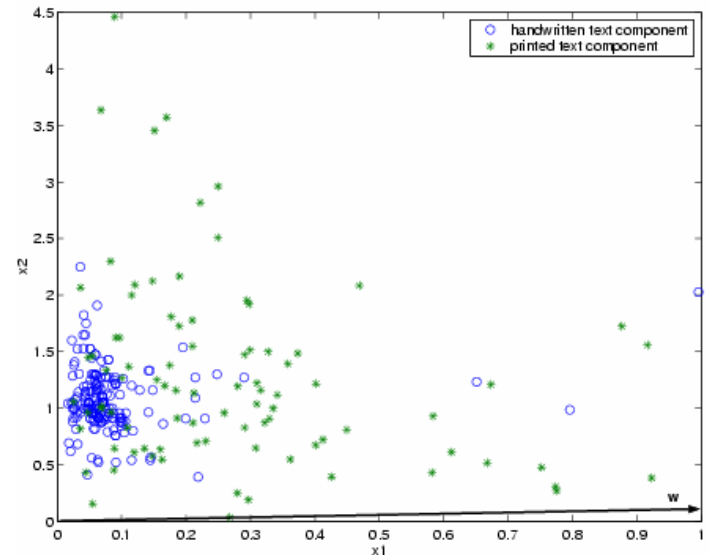
$$h_1 / w_1 = 0.9355$$

After projection

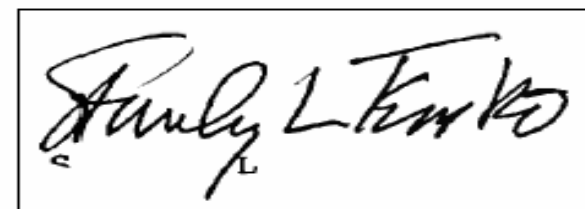
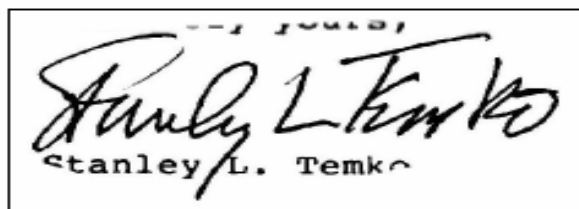
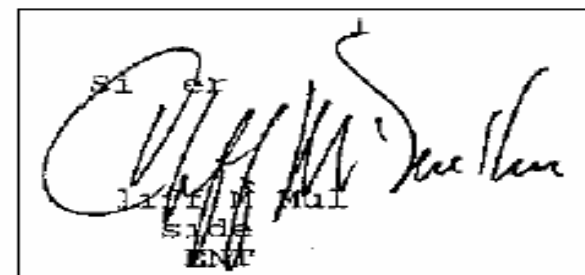
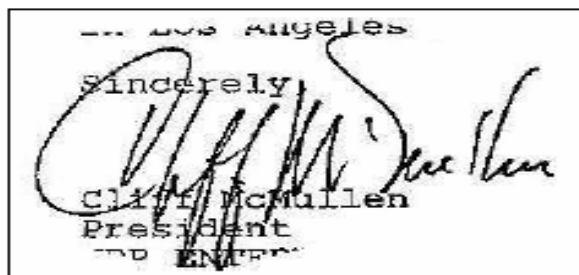
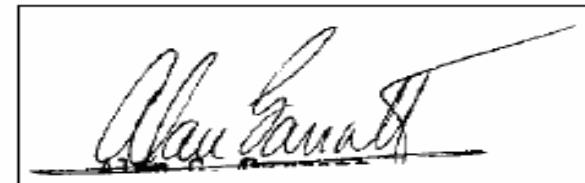
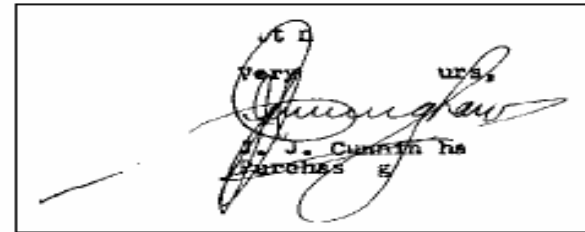
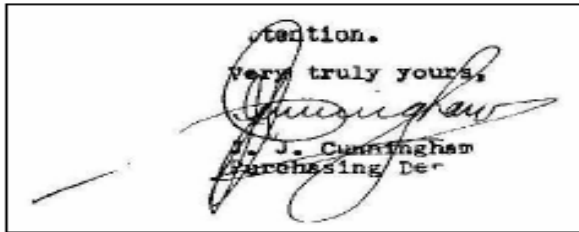
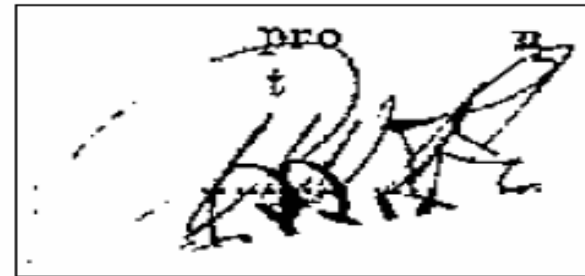
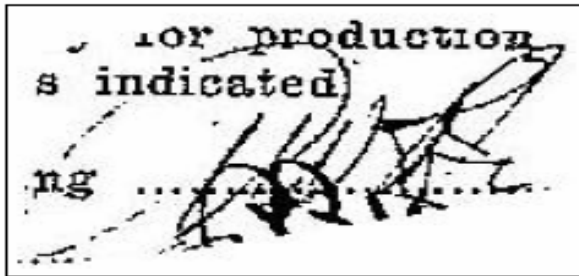
Samples are 1-D

A Bayes classifier is designed for the projected samples

Fisher Linear Discriminant

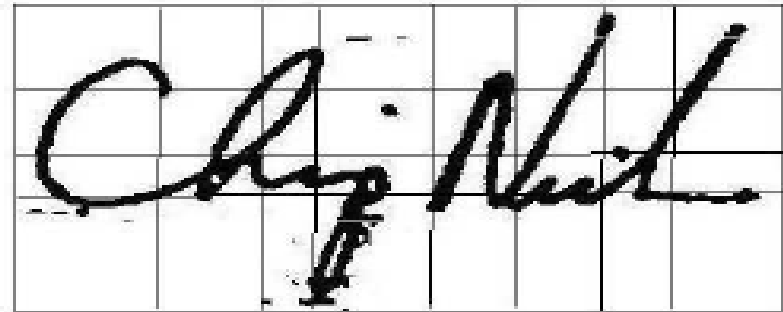


Sample images before and after noise removal



Step 3: Signature Feature Extraction

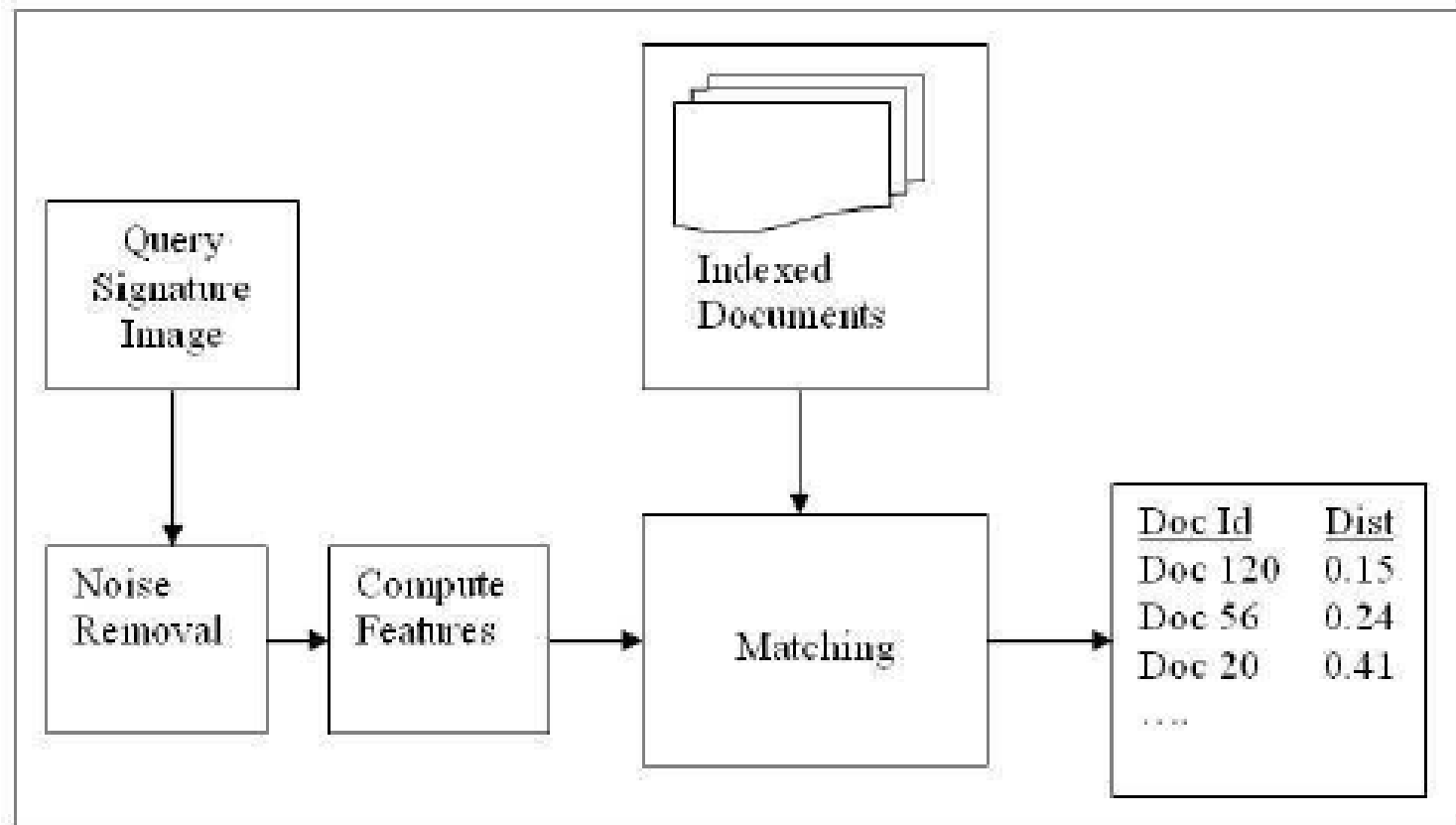
Signature Image under a 4 x 8 Division



1024 bit Feature Vector

```
00110100110101100001100011100001100001100001111111101111111
1000001110011110001100000000011001111001111011100111001
111001110010111101110111000011100000111001111001111000
111001111011110110000000000101110011111101111111
001000011111111111111111110001110001110111100011111
11111001111000000000000010011111111111111110111111
1111111111111111111111111011101111110010000011000000
00111000010000000001100000000001111100111111000000
000000000000110111110001001010000001101110000000011
100000110111110011011110000000000000000000000011100
011001100000100001100000100101000000011100110000011
01000000011111000010111110001100011110000011110100
0001111010000011000000000000000000000100000110001011
11110011111001110000011001000000000000000000000000000000
0000000000001000011010110111111100100000000000010
100000001100010000000000001010111000101100010100000
00000000000000111001010010000000000001000001100000
1100100000000110011011100100110000000000000000000000000
00000100101000010000000000000000001000000100000
```

Document Retrieval



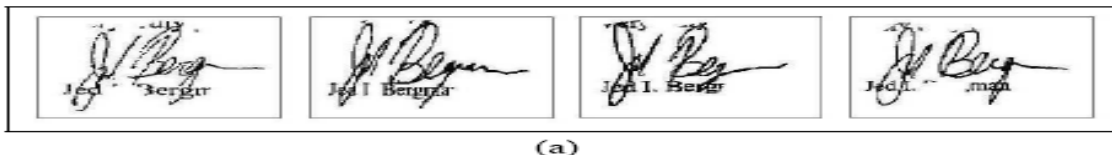
Matching algorithm

Distance between query and each indexed document in database is calculated using normalized correlation distance

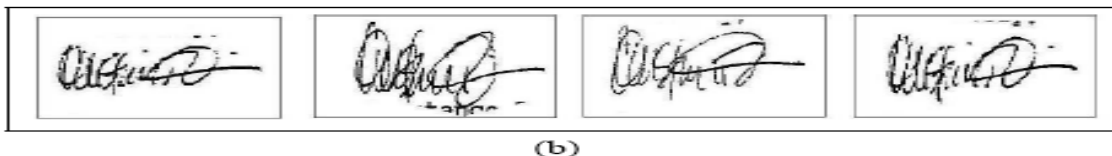
$$S(X, Y) = \frac{1}{2} + \frac{S_{11}S_{00} - S_{10}S_{01}}{2((S_{10} + S_{11})(S_{01} + S_{00})(S_{11} + S_{01})(S_{00} + S_{10}))^{1/2}}$$

Example of Distance Values

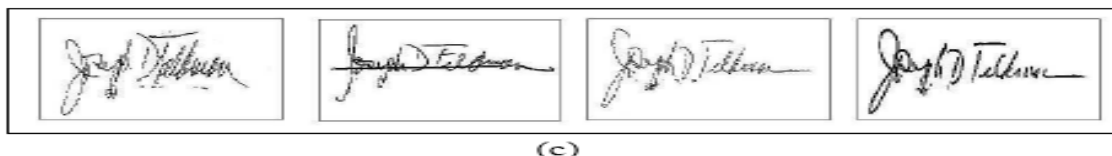
Writer 1



Writer 2



Writer 3



	Writer 1				Writer 2				Writer 3			
	1_a	1_b	1_c	1_d	2_a	2_b	2_c	2_d	3_a	3_b	3_c	3_d
1_a	0	0.22	0.26	0.25	0.36	0.37	0.4	0.32	0.33	0.35	0.32	0.35
1_b	0.22	0	0.25	0.27	0.34	0.36	0.38	0.31	0.3	0.34	0.3	0.36
1_c	0.26	0.25	0	0.25	0.31	0.38	0.36	0.34	0.34	0.35	0.35	0.39
1_d	0.25	0.27	0.25	0	0.35	0.36	0.36	0.35	0.35	0.37	0.36	0.4
2_a	0.36	0.34	0.31	0.35	0	0.27	0.24	0.28	0.33	0.33	0.35	0.4
2_b	0.37	0.36	0.38	0.36	0.27	0	0.25	0.27	0.36	0.38	0.34	0.38
2_c	0.4	0.38	0.36	0.36	0.24	0.25	0	0.22	0.35	0.37	0.38	0.43
2_d	0.32	0.31	0.34	0.35	0.28	0.27	0.22	0	0.31	0.35	0.33	0.39
3_a	0.33	0.3	0.34	0.35	0.33	0.36	0.35	0.31	0	0.26	0.24	0.29
3_b	0.35	0.34	0.35	0.37	0.33	0.38	0.37	0.35	0.26	0	0.27	0.28
3_c	0.32	0.3	0.35	0.36	0.35	0.34	0.38	0.33	0.24	0.27	0	0.17
3_d	0.35	0.36	0.39	0.4	0.4	0.38	0.43	0.39	0.29	0.28	0.17	0

Values are smaller within writer

Dataset for Manually Extracted Signatures

Total No of Signatures	447
Number of Writers	40
Number of Signatures With Printed Text	137
Avg Number of Signature Types per Writer	2
Max Number of Signature Types per Writer	6

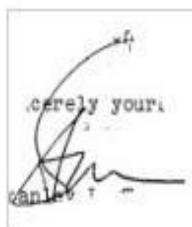
All Samples for a Writer



a



b



c



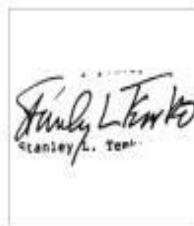
d



e



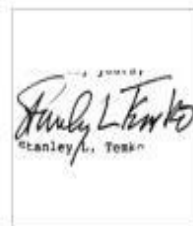
f



g



h

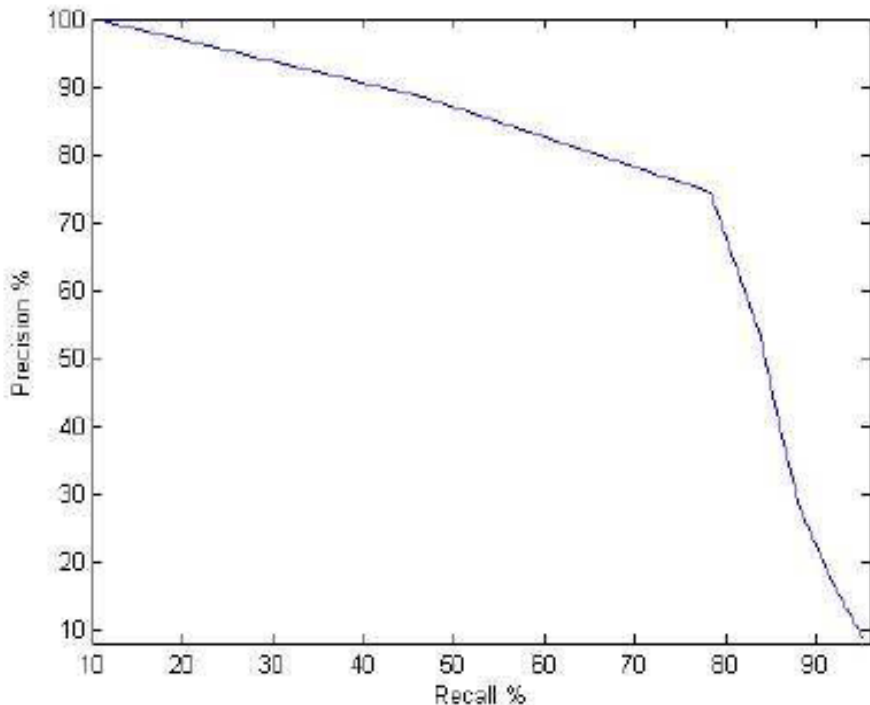


i

60 queries

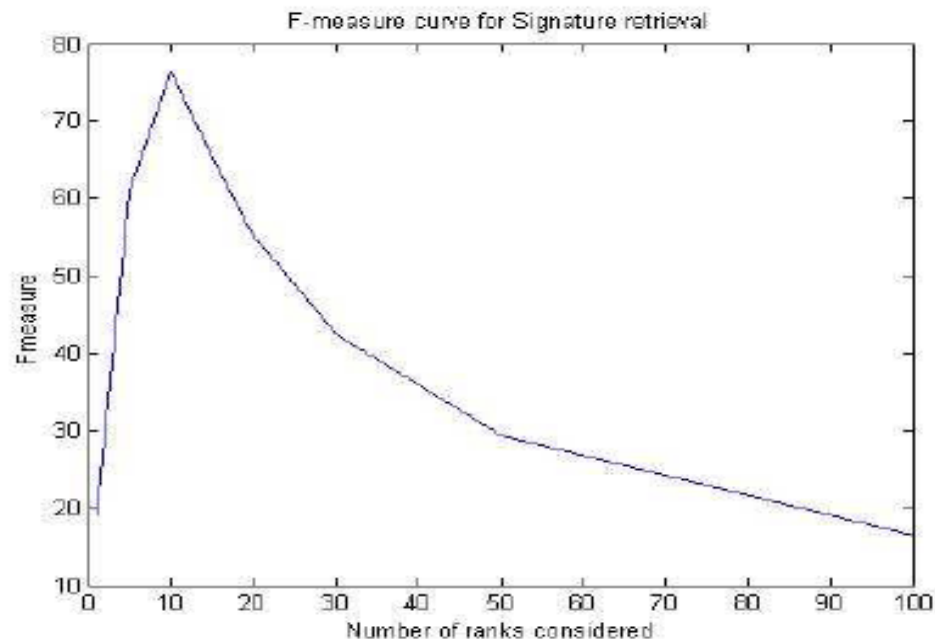
Performance for Manual Extraction

Precision-Recall



A precision of 74.5% was obtained at a recall of 78.28%.

F Measure: Harmonic Mean Of Precision and Recall



On considering the Top 10 ranks, a F-measure value of 76.3 was obtained.

Dataset for Automatic Extraction (Entire Document)

Automatic Signature Extraction was tested on a dataset of 300 documents, including documents with

- Printed Text
- Handwritten Text
- Logos
- Noise
- Scratches, Scribbles, Words Circled
- Lines and Black Borders
- Tables, Seals
- Poorly Scanned Documents

Tested on 80 queries

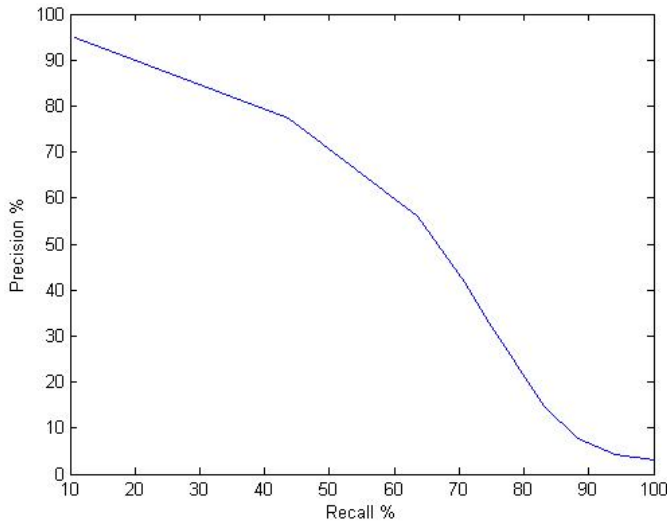
Performance Improvement using Query Expansion

Automatic Relevance Feedback

- A query expansion is done using the feedback (retrieval results) of the matching algorithm
- The signature image returned by the matching algorithm with the lowest score is added to the existing query to formulate an expanded query
- Retrieval is then performed using the expanded query

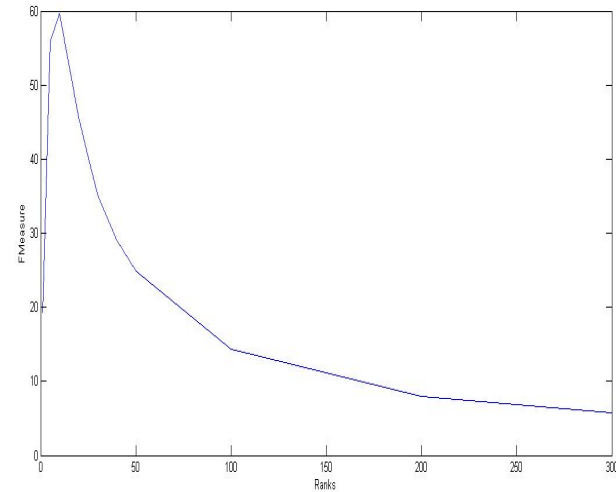
Performance for Automatic Extraction

Precision-Recall



Comparison of Precision-Recall Curves of signature retrieval results for 35 writers – Precision of 56% at a Recall of 63%

F Measure: Harmonic Mean Of Precision and Recall



On considering the Top 10 ranks, a F-measure value of 59.6 was obtained.

Conclusion and Future Work

- Method can be extended to several other document image retrieval applications
- Potential improvements:
 - increasing the feature set
 - using contextual information
 - handling multiple signature types
 - handling multiple signatures on same document