IDAP

International Digital Archives Project

Cursive Handwriting Recognition for Document Archiving

Trish Keaton

Rod Goodman

California Institute of Technology

Motivation

Numerous documents have been conserved in archives all over the world, however their accessibility is limited. Advancements in cursive OCR have the potential of transforming the primary routes of archive access -- from ***Microfilm access*** *to* ***Internet access****.*

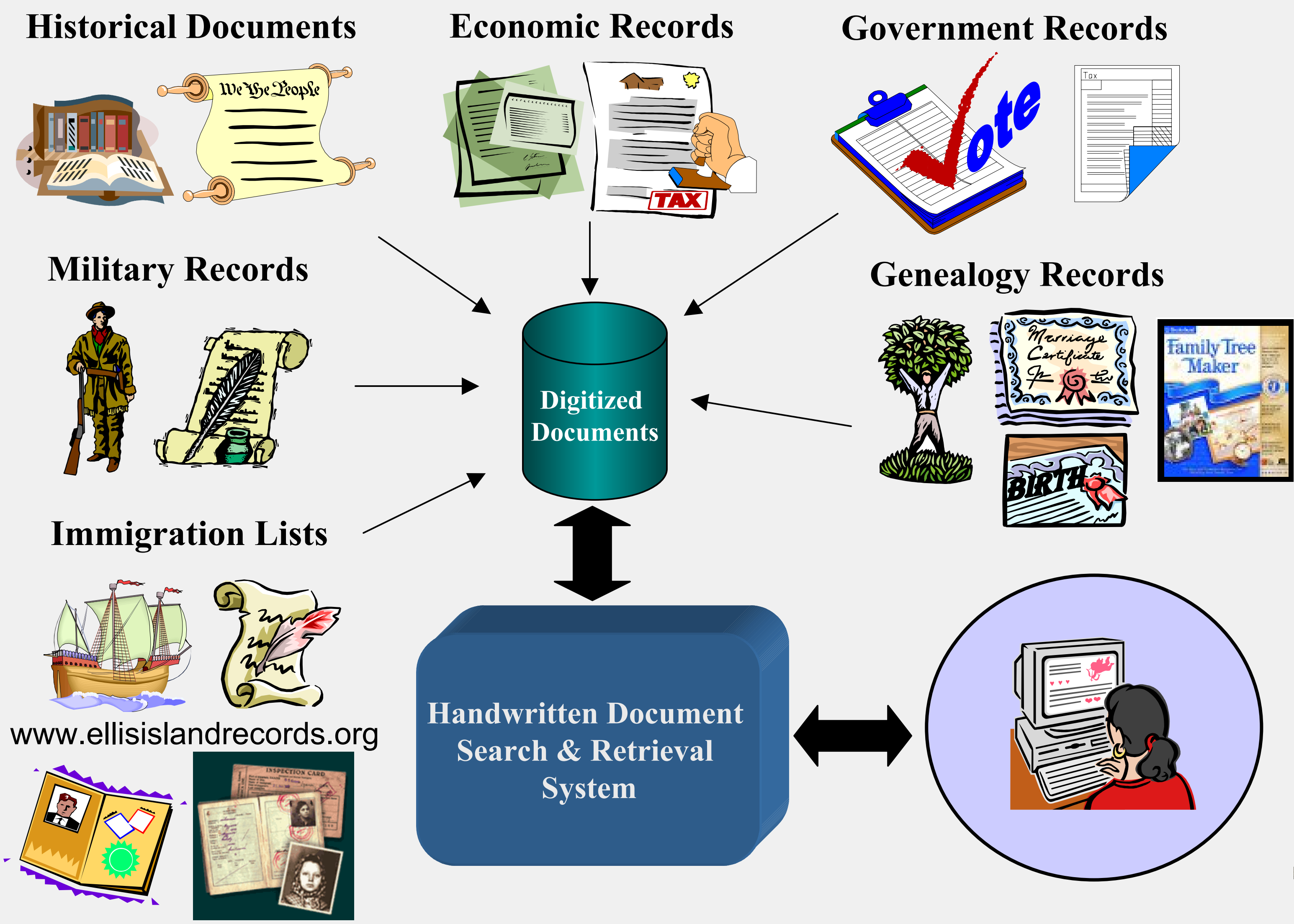
Goal:

Develop OCR algorithms that can automatically recognize cursive handwriting in archive documents with high accuracy, through the exploitation of :

- structured document field analysis
- multiple levels of contextual analysis (e.g., geographical, time period)
- recognition of writing style

IDAP

Vision - Internet Access to Archived Documents



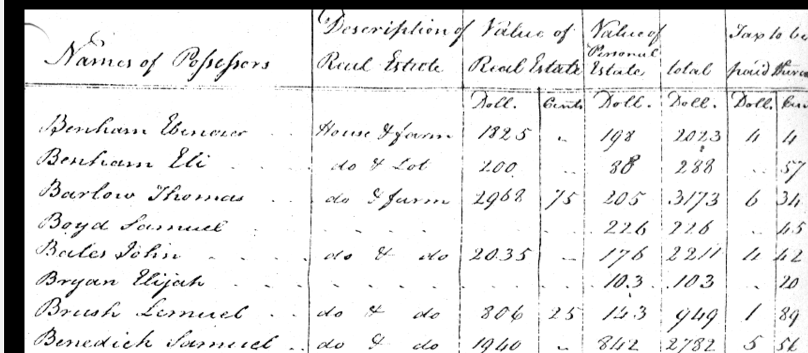
The Technical Challenge

IDAP

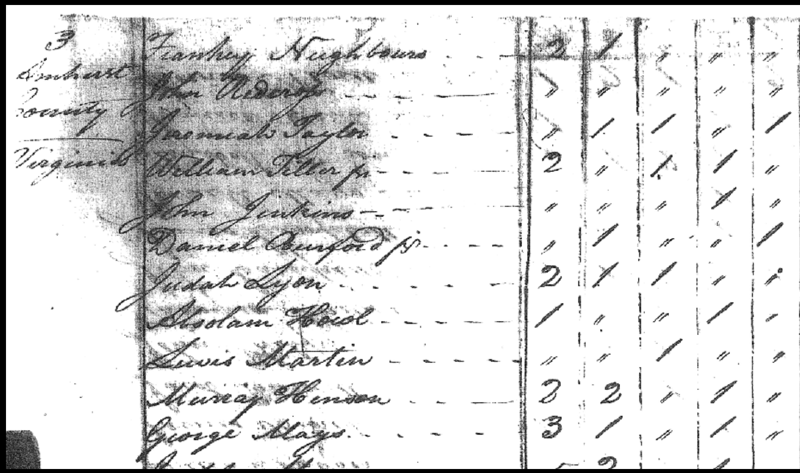
Design a system for the automatic indexing and retrieval of scanned documents written in cursive script by multiple authors.

Difficulties:

- High variability due to the writing styles of multiple authors
- Noise due to the document paper, and scanning artifacts.
- Document form lines, and stray marks or underlines.
- Overlapping words, and mixed styles (e.g., cursive and handprint).
- Lack of ground truth information or suitable data for training purposes.

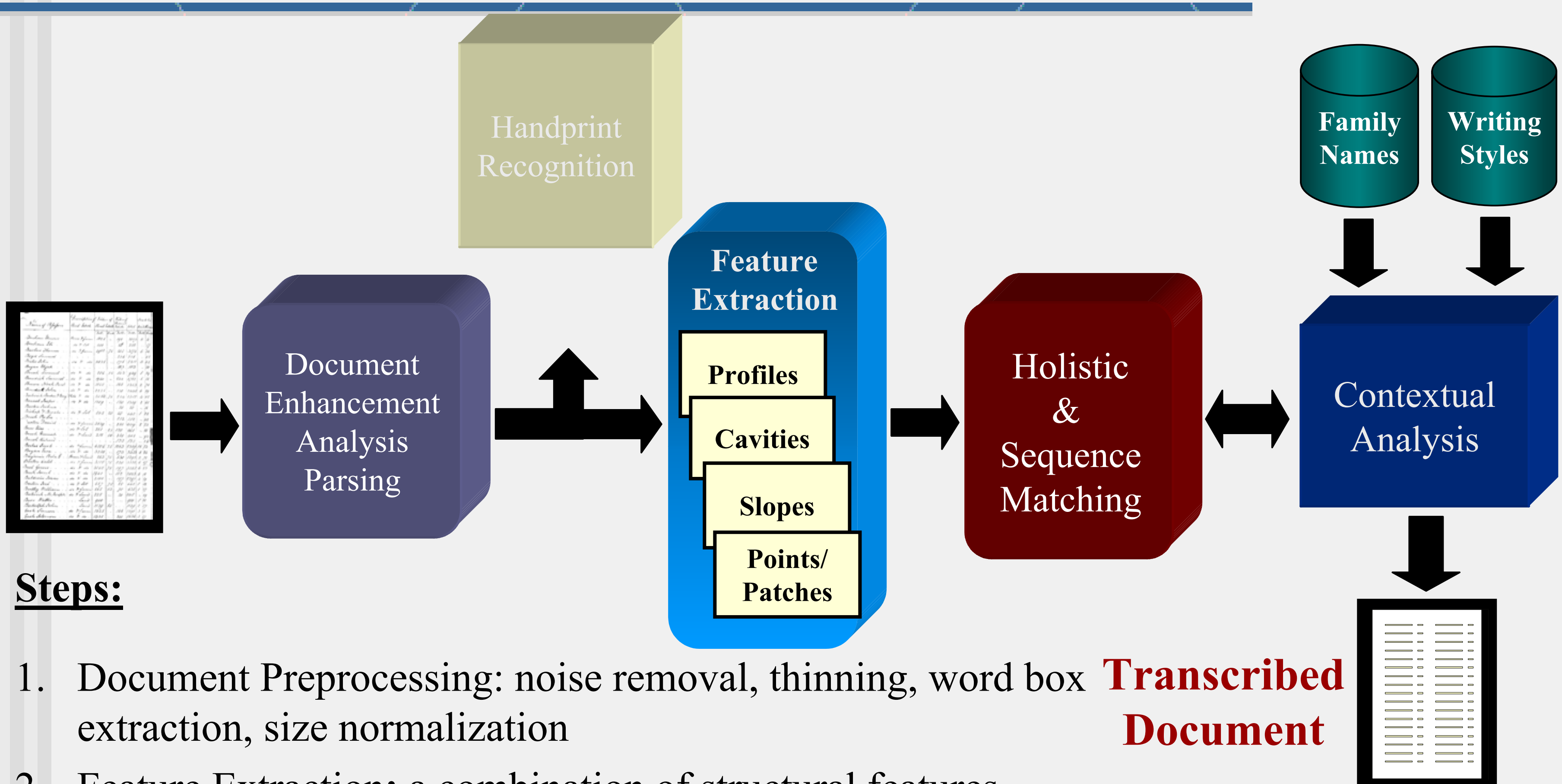


Names of Possessors	Description of Real Estate	Value of Real Estate		Value of Personal Estate	Total	Tax to be paid thereon	
		Doll.	Cents	Doll.	Doll.	Doll.	Cents
Benham Ebenezer	House & farm	1825	..	198	2023	4	4
Benham Eli	do & Lot	200	..	88	288	..	57
Barlow Thomas	do & farm	2968	75	205	3173	6	34
Boyd Samuel				226	226	..	45
Boles John	do & do	2035	..	176	2211	4	42
Bryan Elijah				103	103	..	20
Brush Lemuel	do & do	806	25	143	949	1	89
Benedick Samuel	do & do	1940	..	842	2782	5	56



Our Cursive Recognition Approach

IDAP



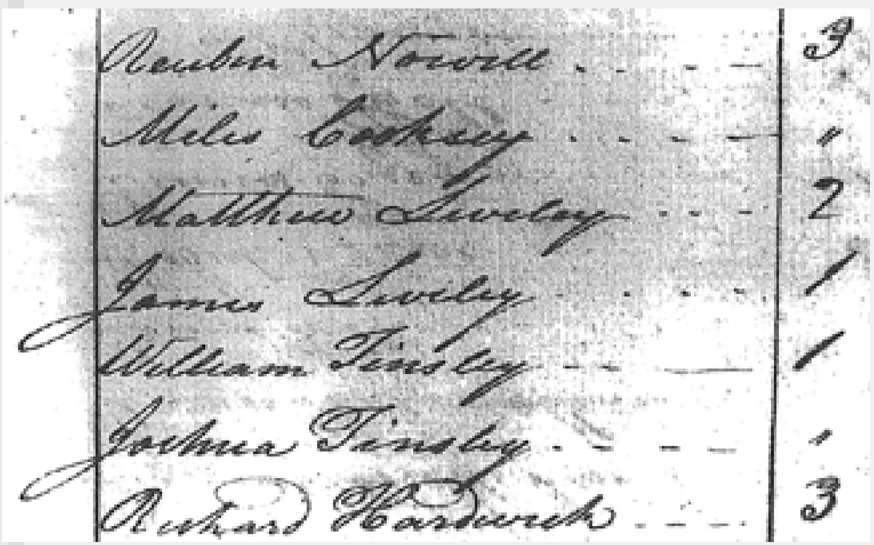
Steps:

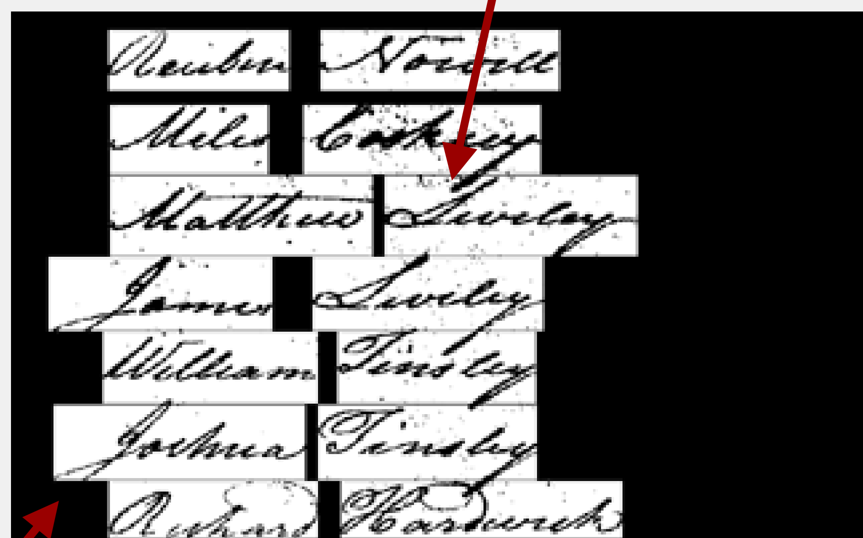
1. Document Preprocessing: noise removal, thinning, word box extraction, size normalization
2. Feature Extraction**:** a combination of structural features.
3. Holistic Matching**:** fuse decisions from multiple classifiers.
4. Sequence Matching: Hidden Markov Model (HMM) based.
5. Contextual Analysis**:** using multiple levels of context.

Document Parsing First Approach

- Split image into “text” / “non-text” regions using projection analysis only.
- Extract the names/words.

Non-Rejection of Clutter

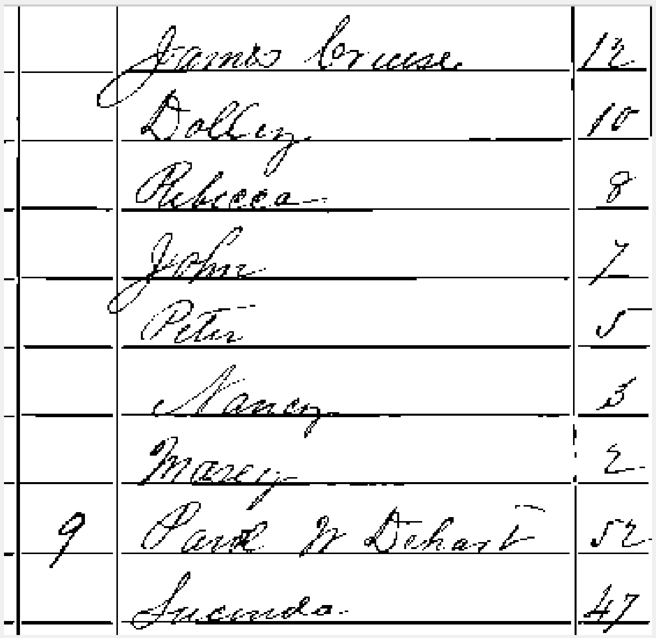


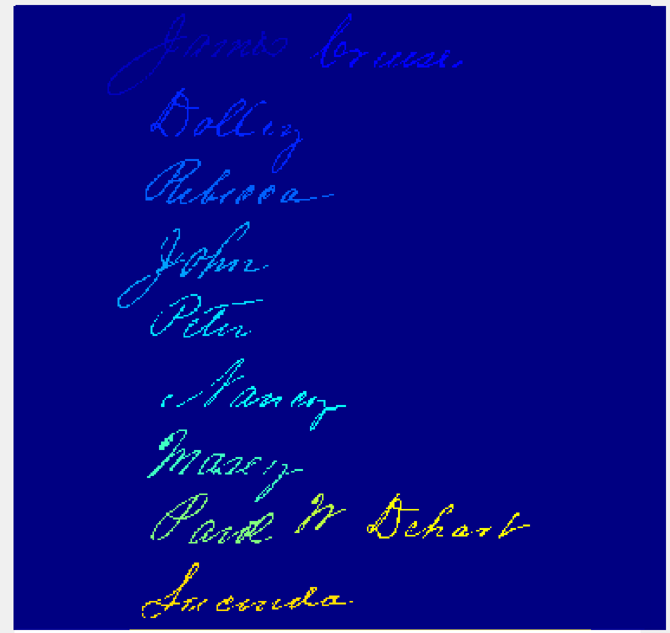


Clips Descenders

Document Parsing Current Approach

IDAP

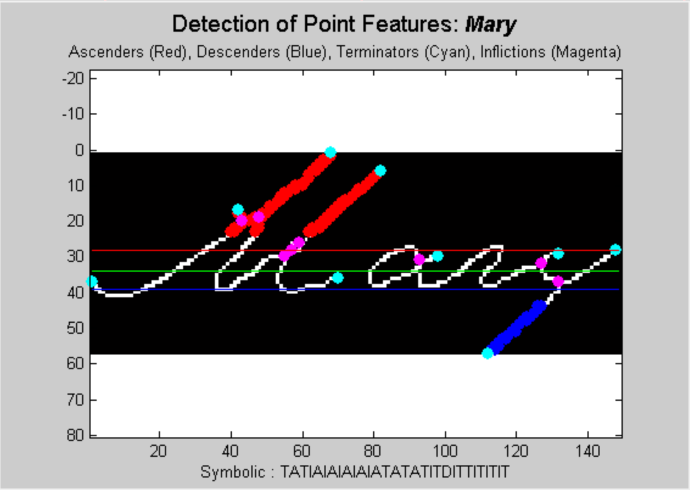




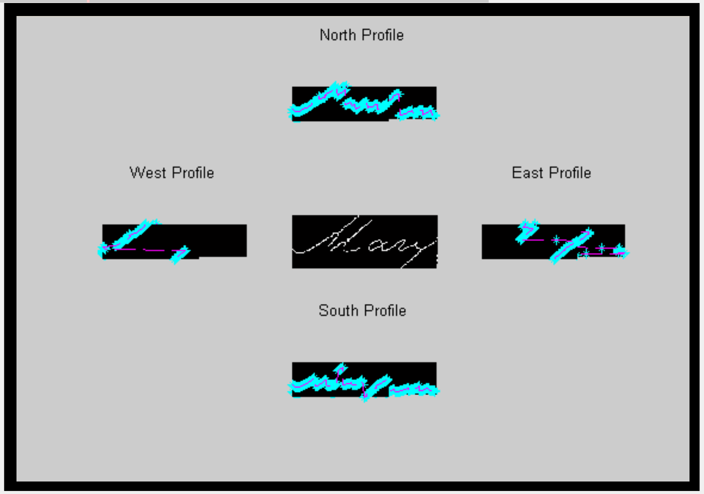
- Remove form lines using a Hough transform technique.
- Estimate the name field using projection analysis.
- Extract the connected components.
- Group components into words by analyzing their sequence of ascender/descender patterns, gaps, and pitch.

Feature Extraction- Feature Sets

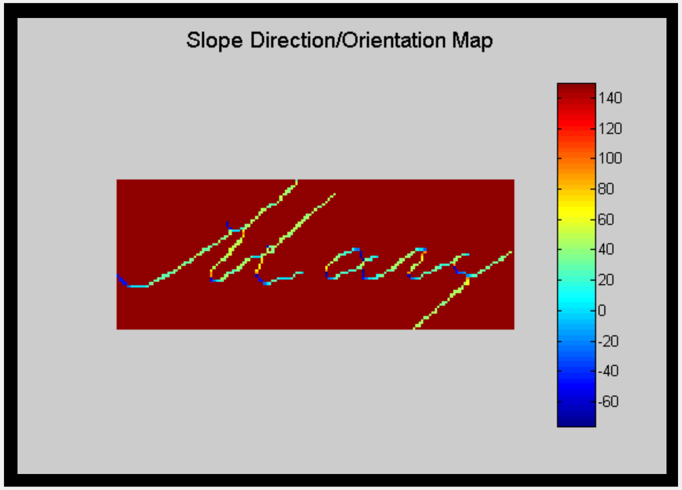
IDAP



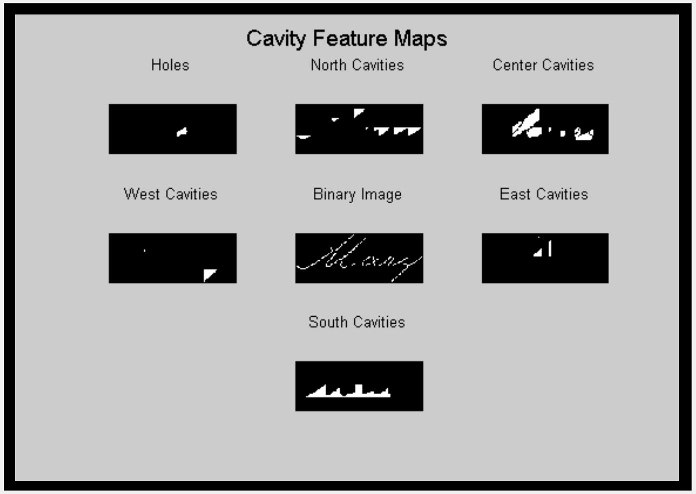
- Ascender/descender & junction points.



- DCT encoded directional projections.



- Slope orientation histograms.



- Coarsely encoded cavity feature maps.

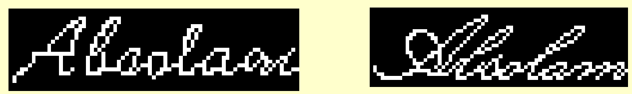
IDAP

Holistic Matching

Holistic matching is used to reduce the number of candidate lexicon matches.

Training Phase:

- Prototypical feature vector exemplars are stored for each word. Samples are collected or synthetically generated using font classes that resemble handwriting.



Brush font class   Kunstler font class

Testing Phase:

- Lexicon filtering based on word length, and presence of ascenders and descenders.

- Measure similarity between feature vectors using Chi-Square Statistic:

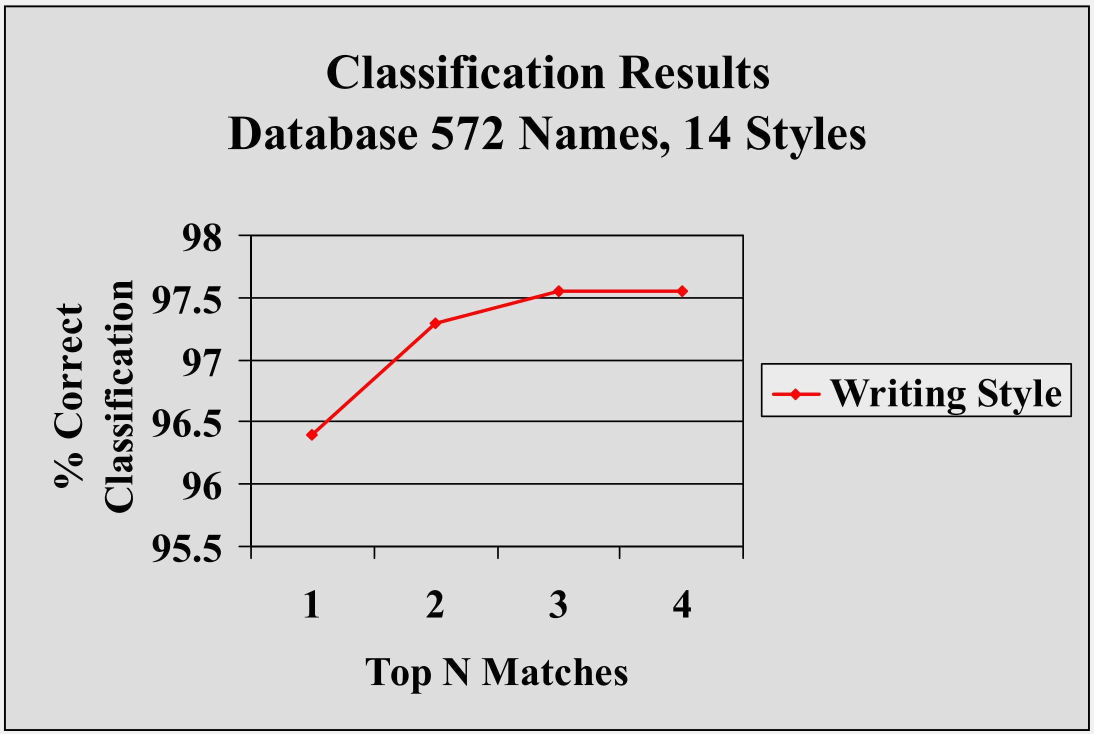
$$d\chi^2(X,Y) = \sum_{i=1}^{N} \frac{(X_i - Y_i)^2}{(X_i + Y_i)}$$

- Fusion of decisions from classifiers based on different feature sets.

IDAP

Writing Style Recognition Performance

Writing style recognition scored on a data base of 572 names written in 14 font styles. Matching based on slope orientation features.



Font Examples

Brush
Rage
English
Edwardian
French
Typoup
Lucida
Magneto
Vivaldi
Vladimir

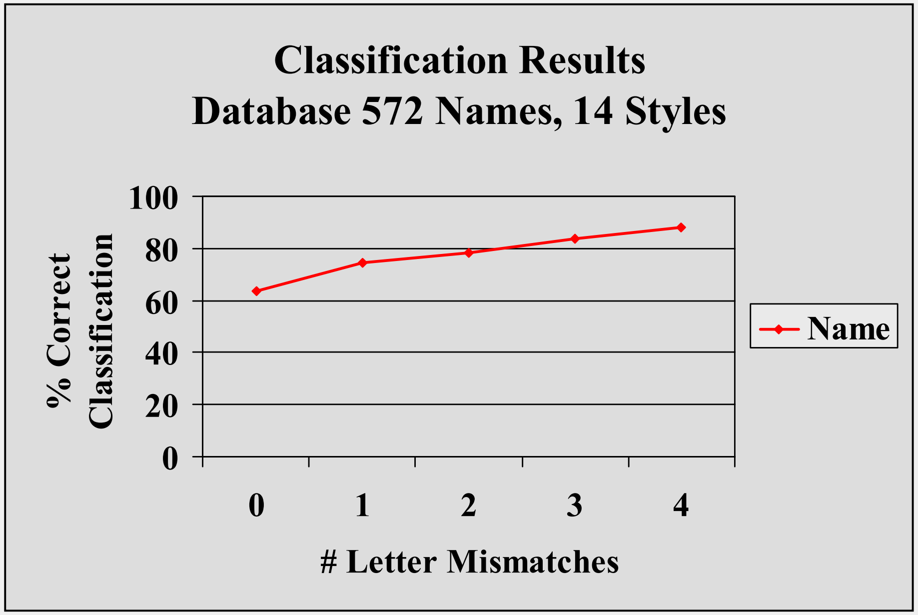
96.4% of writing styles were correctly identified.

Name Recognition Performance Experiment 1

IDAP

Name recognition scored on a data base of 572 names written in 14 font styles. Matching based on a weighted fusion of classifiers based on cavity, profiles and slope orientation features.

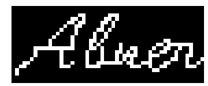
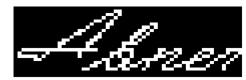
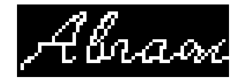
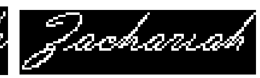
- **63.9% of names were correctly matched – if NO error allowed.**
  Many errors were due to spelling variations such as: Absolam & Absalam, Bobbett & Bobbitt
- **74.2% of names were correctly matched allowing an error of 1 letter.**
- **88.1% of names were correctly matched allowing an error of 4 letters.**



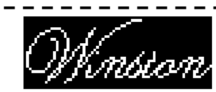
Name Identification Examples
Experiment 1

IDAP

Correctly Identified Names:

Name	Test Case	Ranked Matches (1 → 2)		Conf.
Abner				0.88274
Matthew				0.82387
Staples				0.77233
Zachariah				**0.85625**

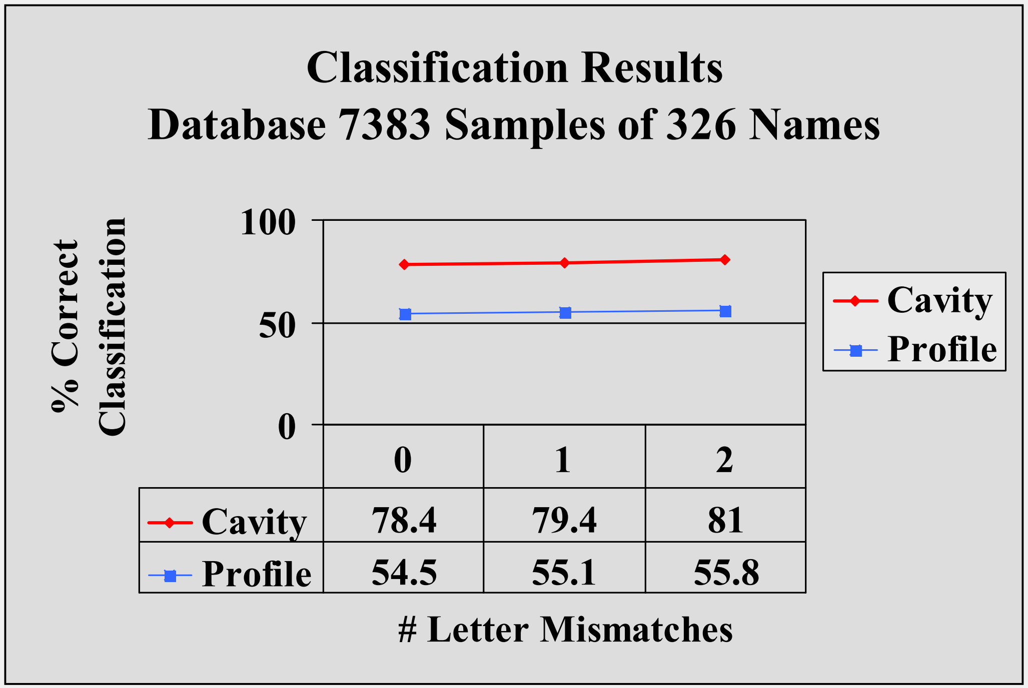
Incorrectly Identified Names:

Name	Test Case	Ranked Matches (1 → 2)		Conf.
Adam				0.86472
Bennett				0.83087
Russell				0.67834
Wingfield				**0.79134**

Name Recognition Performance Experiment 2

IDAP

Name recognition scored on a data base of 7383 image samples of 326 names extracted from the 1860 Virginia census. Matching based on cavity and profile features.



- **Many errors were due to one letter confusions : Adam – Adams, Ann – Anna, Fair – Fain, Francis – Frances, Hall – Hill, Ida – Ira, Tazwell – Tazwill, Wood – Woods.**
- **Name confusions accounting for most of the errors: James-Jane, James-Jones, Martha-Martin**

Name Identification Examples
Experiment 2

IDAP

Correctly Identified Names:

Name	Test Case	Top Match	Conf.
Catherine			0.97724
Henry			0.79388
Madison			0.7129
Thomas			0.90142

Incorrectly Identified Names:

Name	Test Case	Top Match	Conf.
Albert - Dehart			0.74505
Jane - James			0.60272
Tazwill - Tazwell			0.58674
William - Willson			0.80992

Name Recognition Performance Experiment 3

IDAP

Name field recognition scored on a total of 1013 names processed from 55 sample documents of the 1860 Virginia census. Classification was based on the holistic matching of cavity features against a training set of 323 unique names (10 samples/name) .

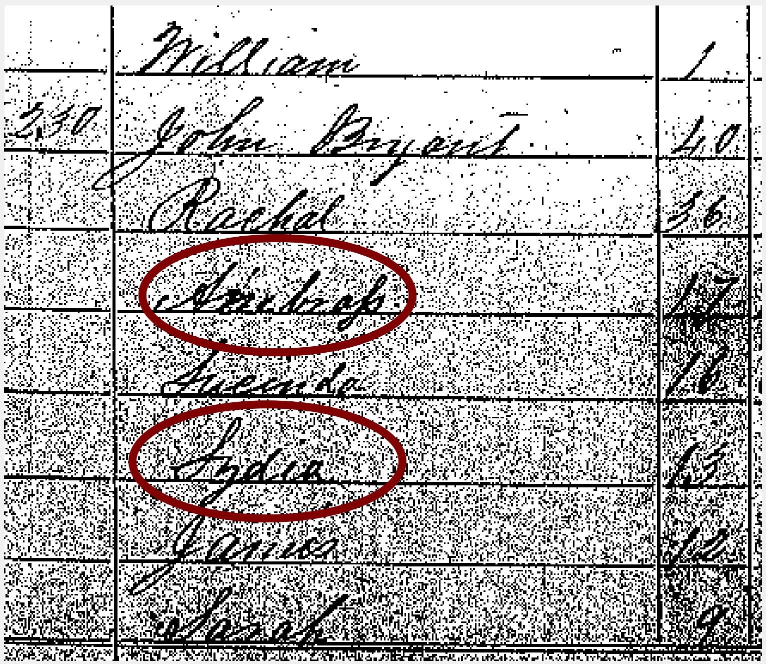
% Names Correctly Classified	**50.05%**
% Names Incorrectly Classified	**2.12%**
% Names Rejected	**47.83%**

- **31.39% of unrecognized names were degraded by speckle noise resulting from the document scanning/digitization process.**
- **26.94% of unrecognized names did not appear in the training set.**
- **14.31% of unrecognized names were a result of word parsing errors.**
- **6.81% of unrecognized names were a result of name field parsing errors.**
- **6.71% of unrecognized names were a result of line & noise removal errors.**
- **6.74% of rejected names were correctly classified, but rejected due to a low confidence (< 60%).**

Name Field Recognition Example (1)

IDAP

Original Document with Speckle Noise



Errors due to names not appearing in training set.

Low confidence due to noise degradation.

Recognition Results

William
0.7724

John
0.7207

Bryant
0.92665

Rachal
0.93356

Josiah
0.40791

Lucinda
0.74296

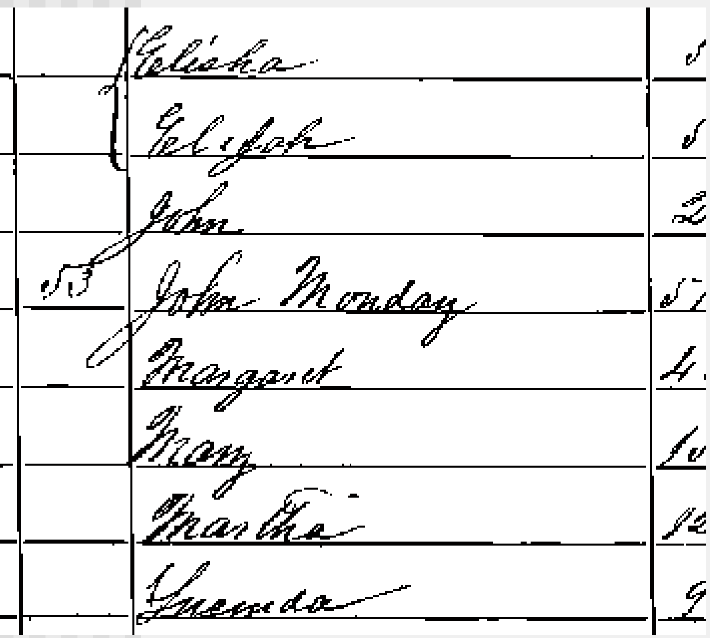
Agnes
0.23546

James
0.75157

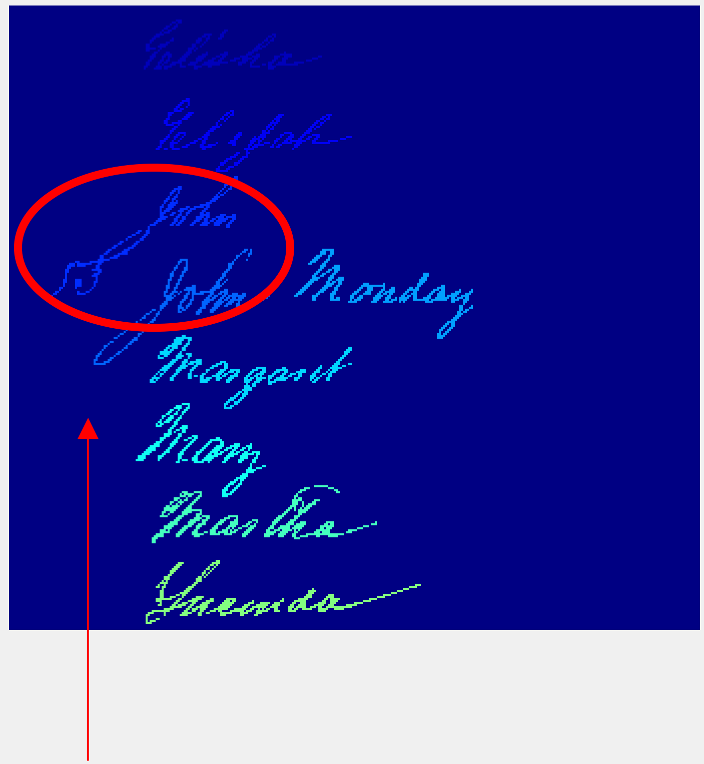
Sarah
0.45188

Name Field Recognition Example (2)

Original Document



Parsing Results

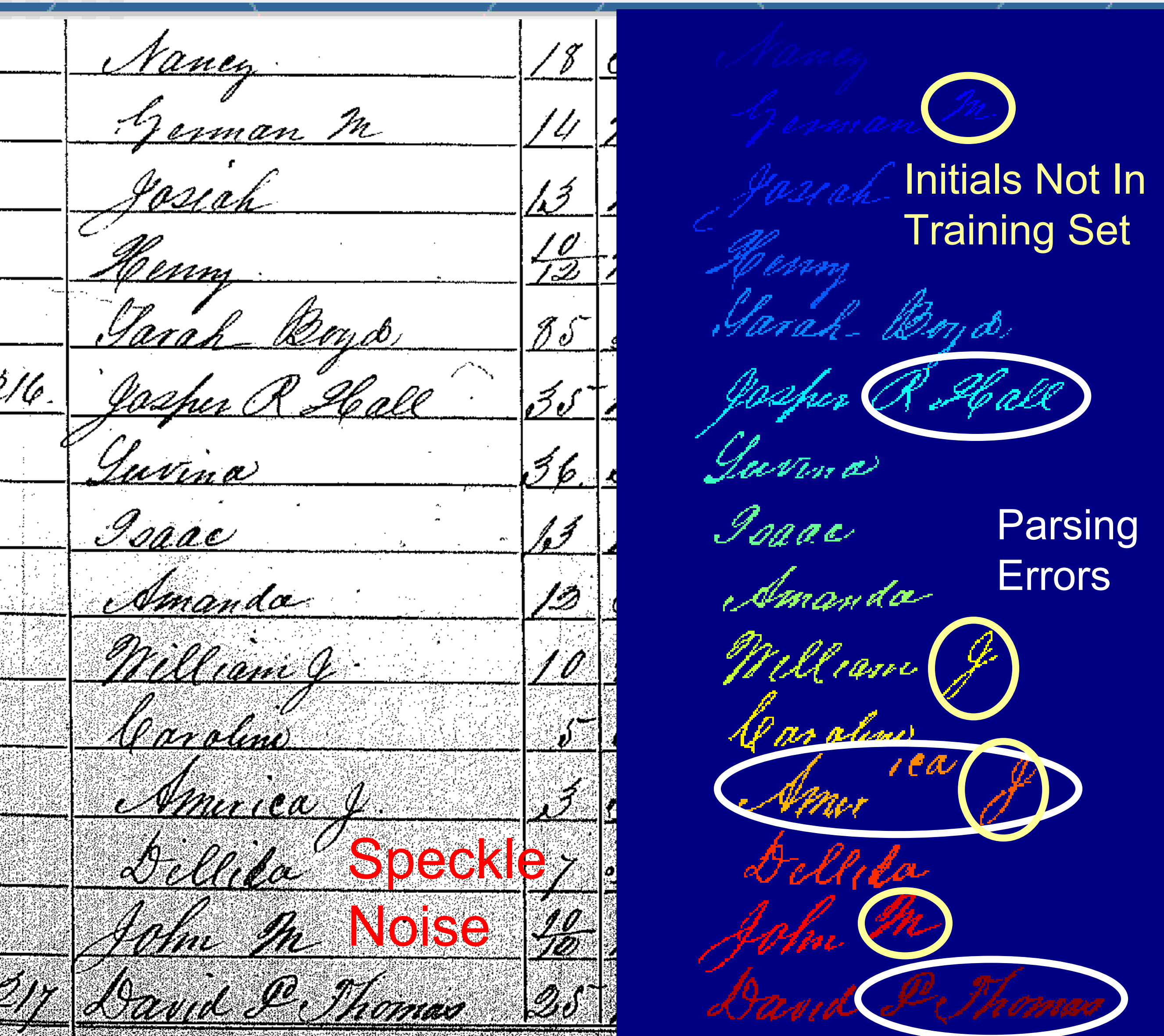


Recognition Results

Low confidence due to name field parsing error.

Name Field Recognition Example (3)

IDAP

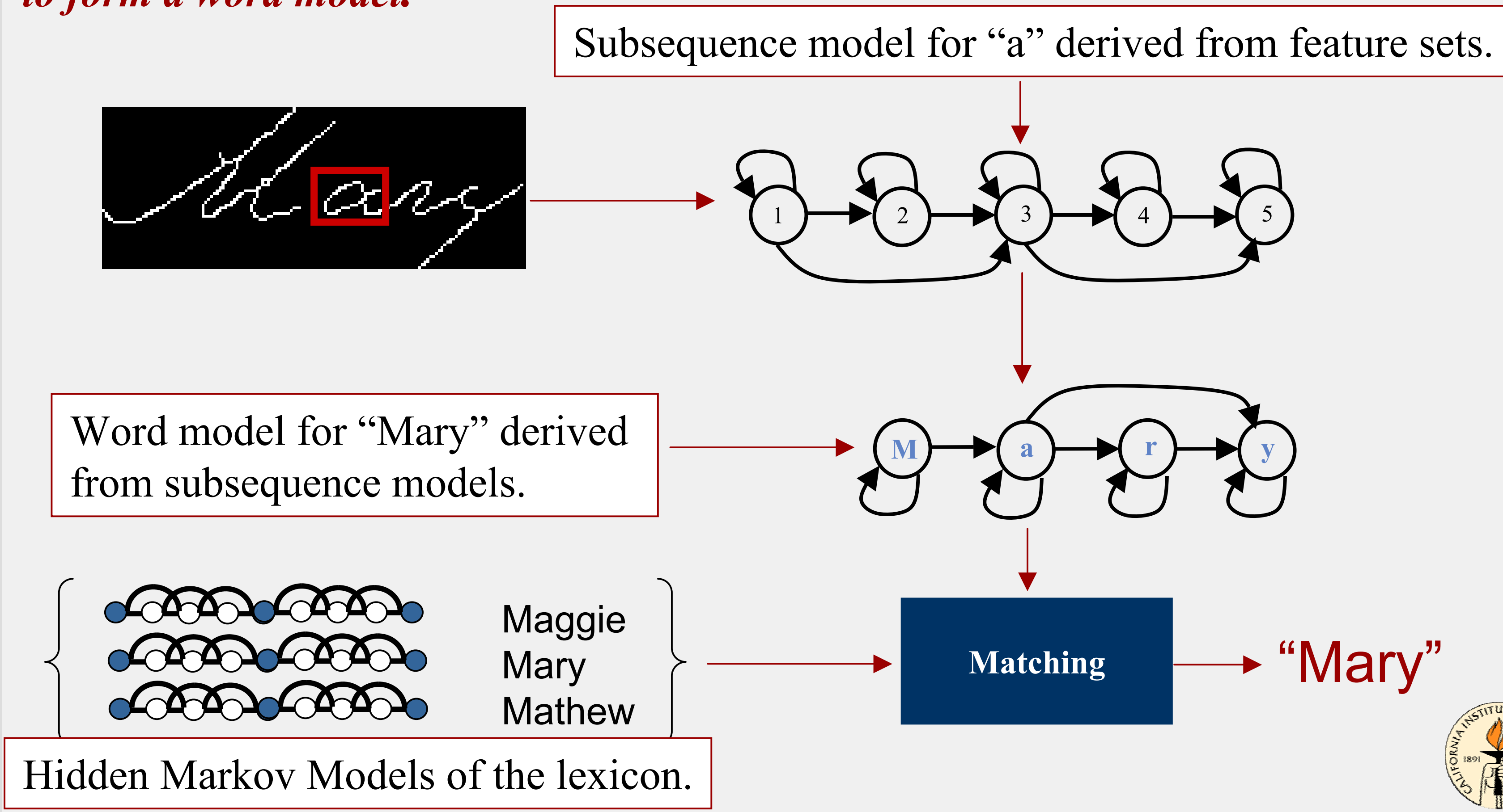


Nancy 0.98279		
German 0.8802	John 0.13419	
Josiah 0.68491		
Henry 0.73367		
Sarah 0.49852	Boyd 0.77243	
Joseph 0.4406	Joseph 0.31942	
Luvina 0.64195		
Isaac 0.97143		
Amanda 0.82039		
William 0.88394	Charles 0.12068	
Caroline 0.84526		
Ann 0.7615	James 0.25477	Jane 0.20168
Belcher 0.21408		
John 0.9061	John 0.91679	
Daniel 0.67248	William 0.39678	

Sequence Matching

IDAP

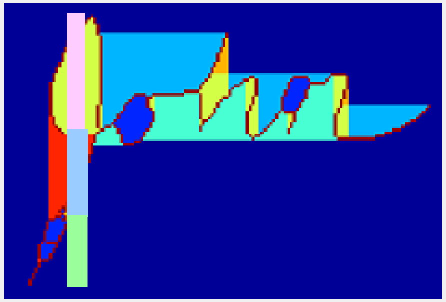
- *Words are represented as a sequence of feature symbols (which could represent a single letter or multiple letter subsequence).*
- *A Hidden Markov Model (HMM) is trained for each subsequence, and concatenated to form a word model.*

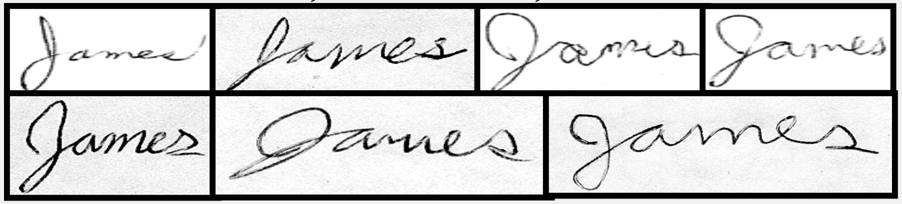


IDAP

Stroke vs. Cavity CHMM Modeling

- A 5 pixel width sliding window split it into 3 regions is used.
- We compute the stroke or cavity density within each region to create the feature vector.
- Train a left-to-right continuous 18-state and 22-state HMM on the stroke & cavity features vectors.
- Name recognition results scored on a difficult database of 13 names all beginning with the letter "J" written by 7 authors, and containing common confusions: James – Jane, Jeff – Jill, Josh - John





Holistic Matching:	**20% of names were correctly matched.**
Stroke CHMM:	**38% of names were correctly matched.**
Cavity CHMM:	**50% of names were correctly matched.**

Future Research Directions

- Model relationships between features to design detectors that can spot names and parts of names without the need for highly accurate word segmentation.
- Experiment with different sequence matching algorithms (e.g., HMMs, graphical models) that will be employed at the sub-word level to better cope with a lack of representative training examples.