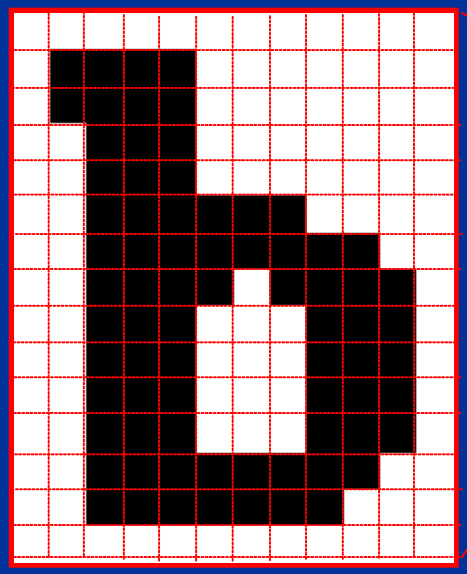
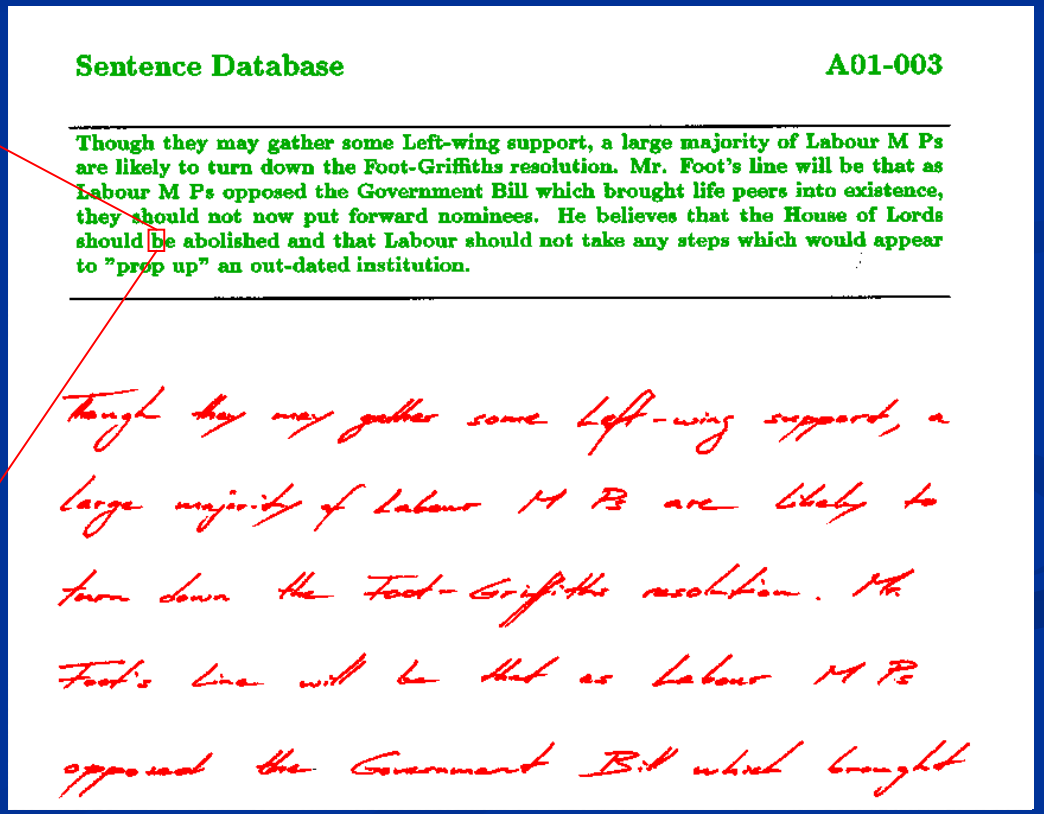
Connected Component Level Method Identification in Automatic Titleboard Indexing

Samuel James Pinson
Mark Pinson
Dr. William Barrett
Computer Science Department
Brigham Young University

Connected Component Level Method Identification

- Distinguishing between machine print and handwriting





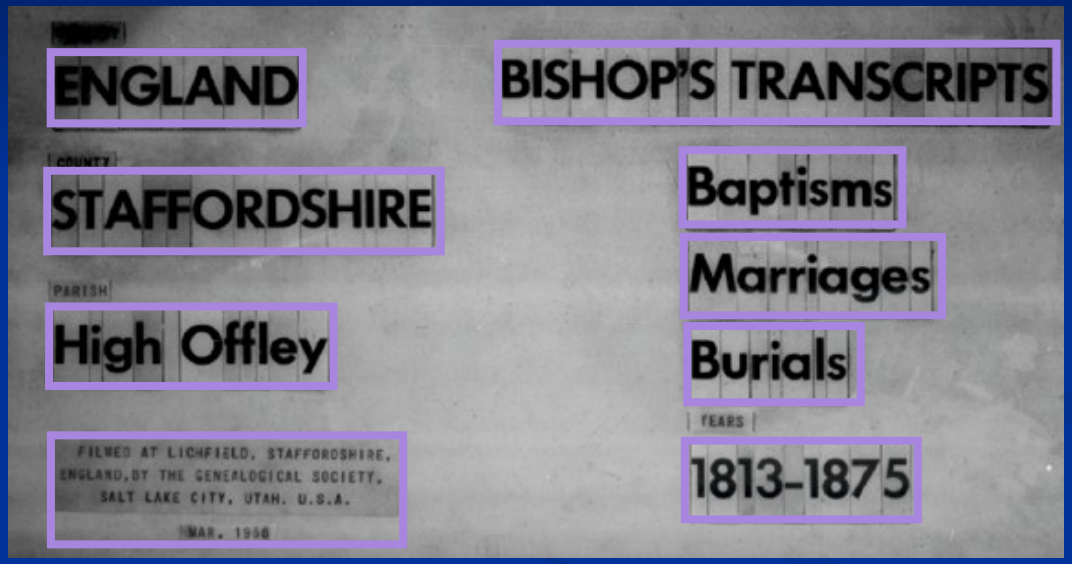
Sentence Database A01-003

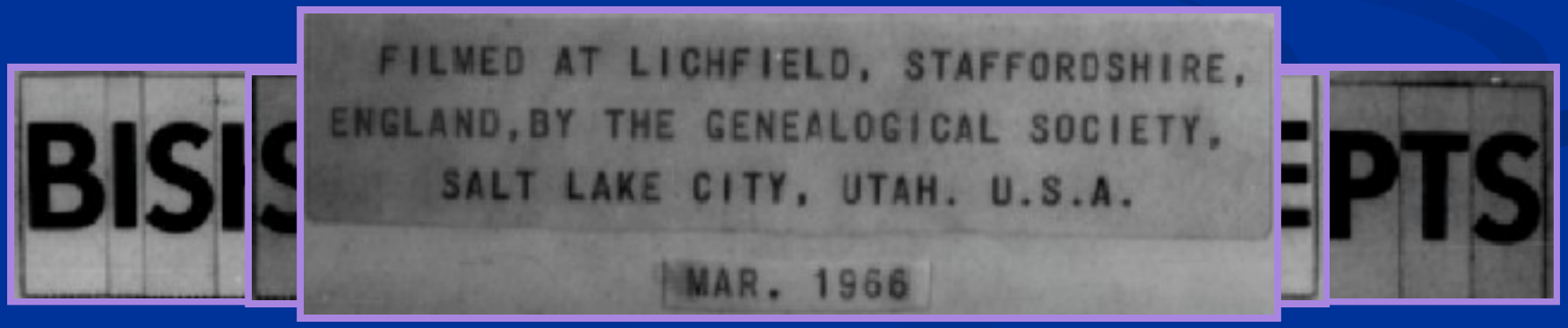
Though they may gather some Left-wing support, a large majority of Labour M Ps are likely to turn down the Foot-Griffiths resolution. Mr. Foot's line will be that as Labour M Ps opposed the Government Bill which brought life peers into existence, they should not now put forward nominees. He believes that the House of Lords should be abolished and that Labour should not take any steps which would appear to "prop up" an out-dated institution.

Though they may gather some Left-wing support, a large majority of Labour M Ps are likely to turn down the Foot-Griffiths resolution. Mr Foot's line will be that as Labour M Ps opposed the Government Bill which brought

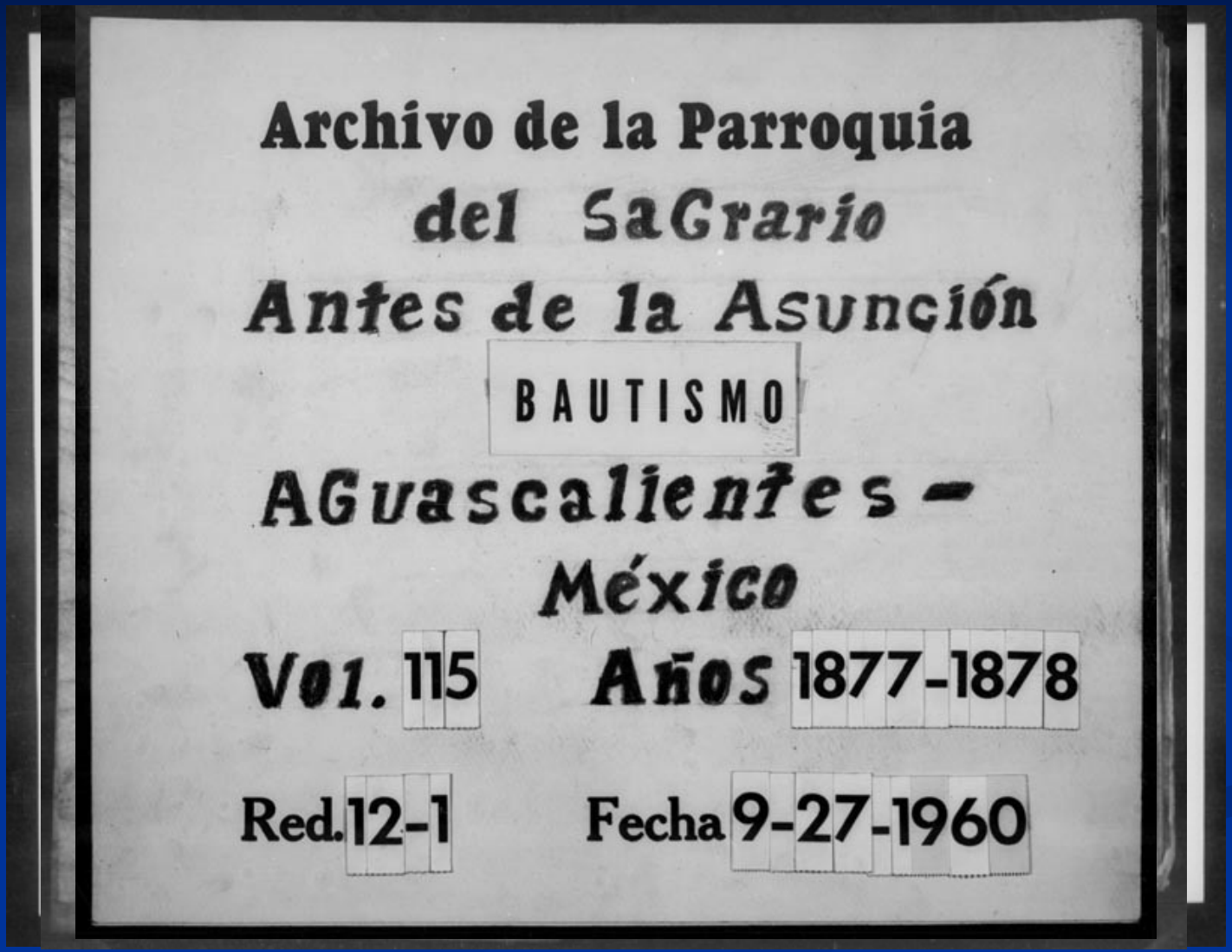
Microfilm Titleboards

- Type
- Location
- Time
- Acquisition





Microfilm Titleboards

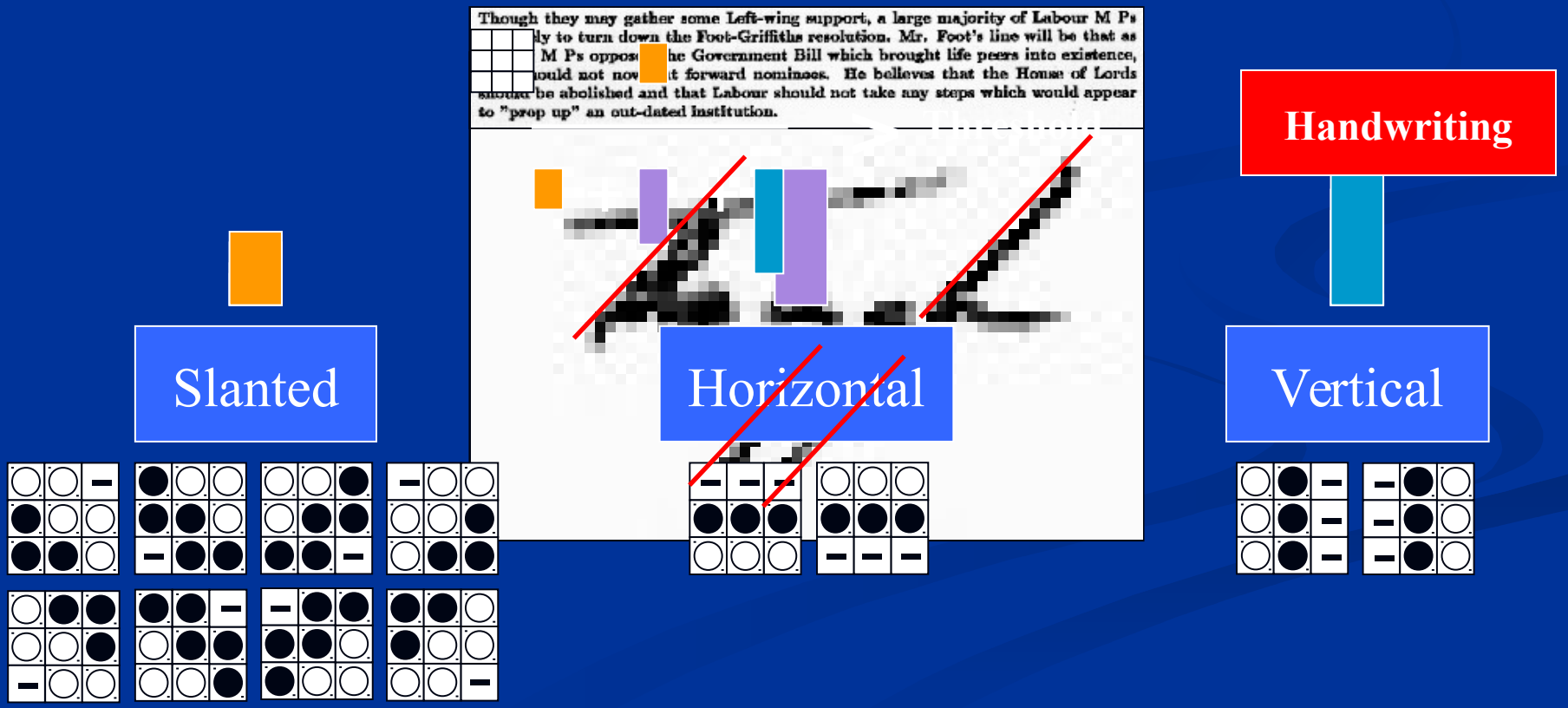


Problem Statement

- To make genealogical microfilm more accessible by automatically building a searchable index over titleboards
  - Preprocess titleboards
  - Distinguish between machine print and handwriting (Method identification)
  - Recognize machine print and handwriting (OCR)
  - Build searchable index

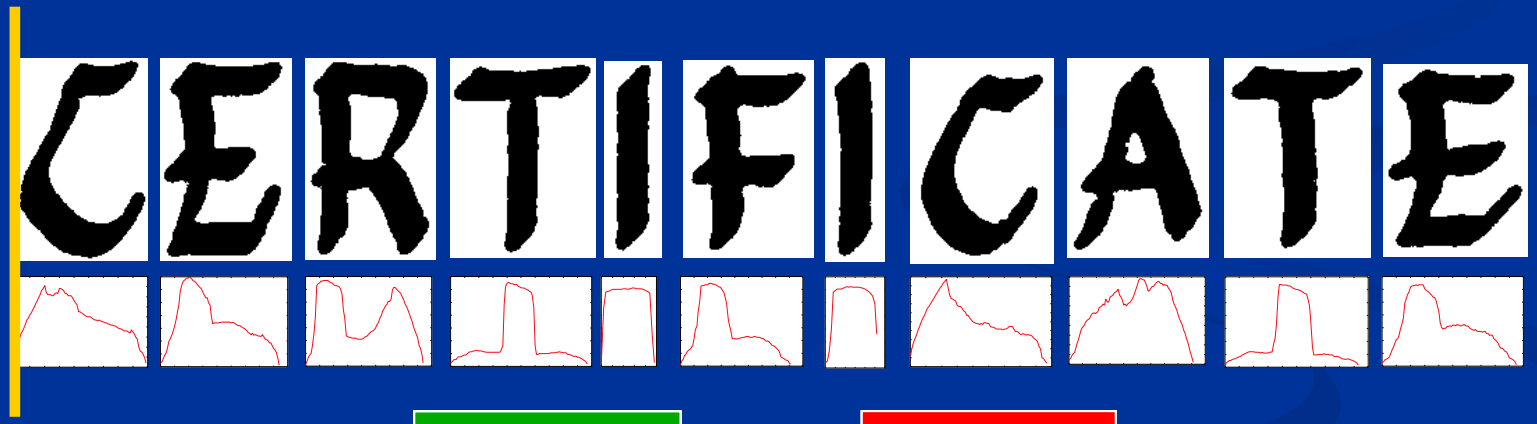
Method Identification

- Discriminator Between Handwritten and Machine-Printed Characters [Umeda et al. 1990]
  - U.S. Patent 4,910,787



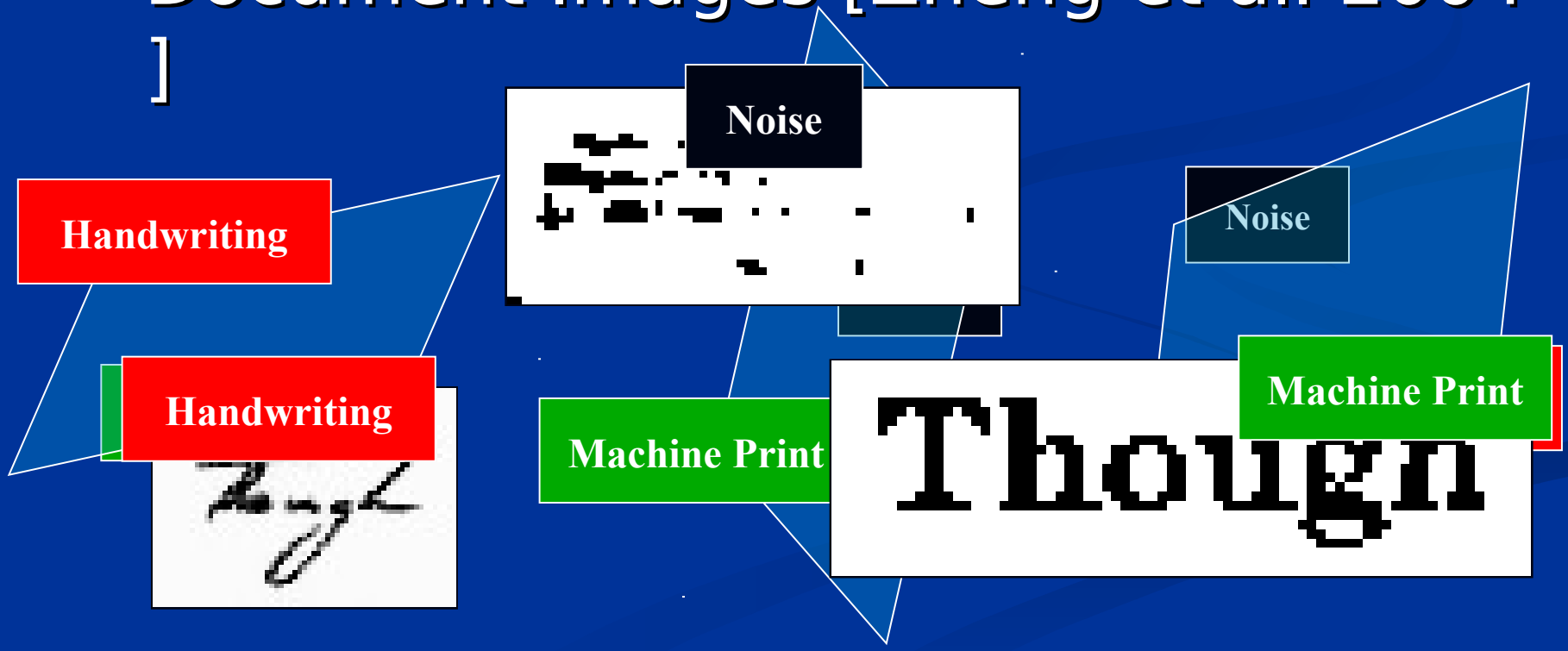
Method Identification

- Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models [Guo et al. 2001]



Method Identification

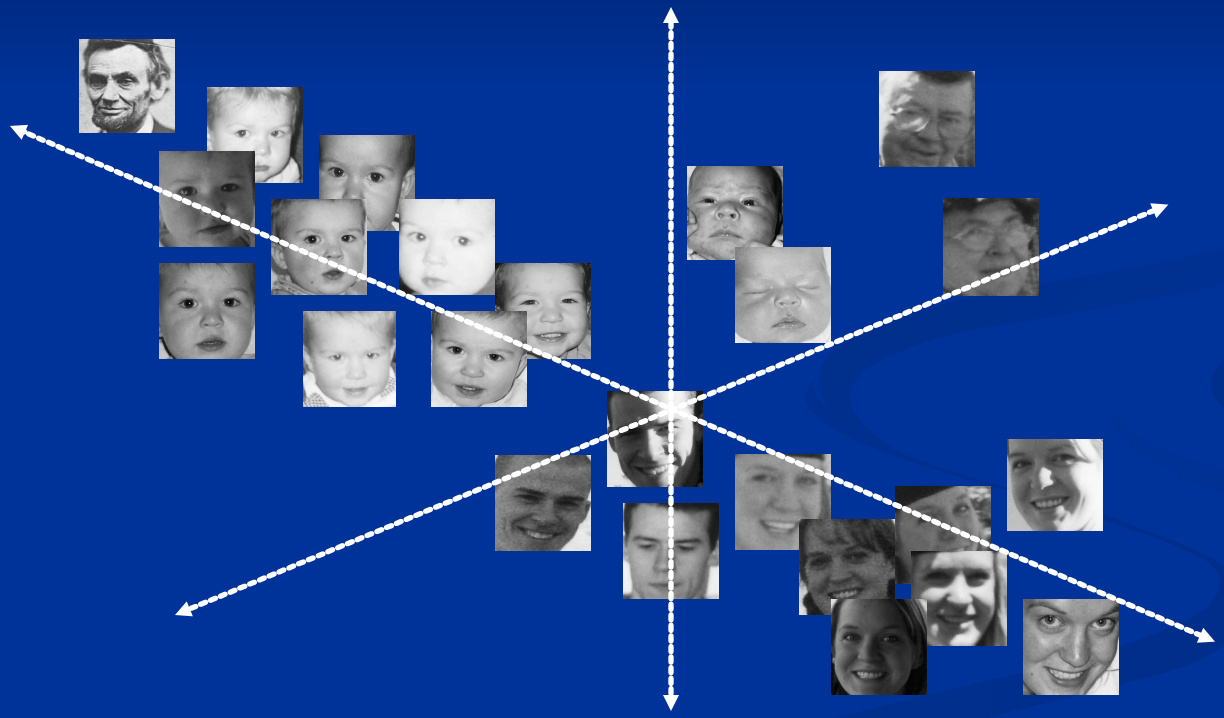
- Machine Printed Text and Handwriting Identification in Noisy Document Images [Zheng et al. 2004]



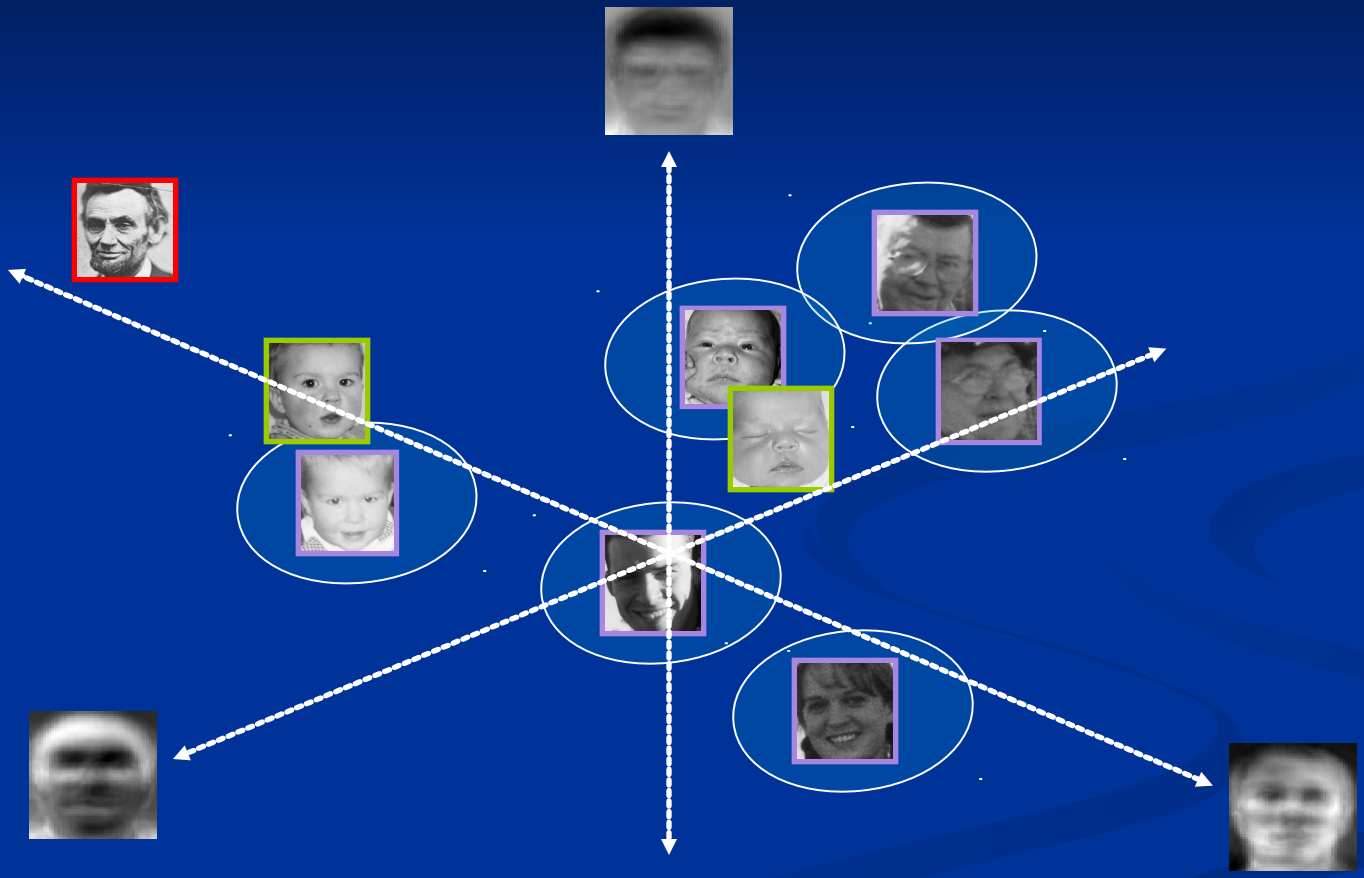
Method Identification

- Eigenfaces [Turk and Pentland 1991]

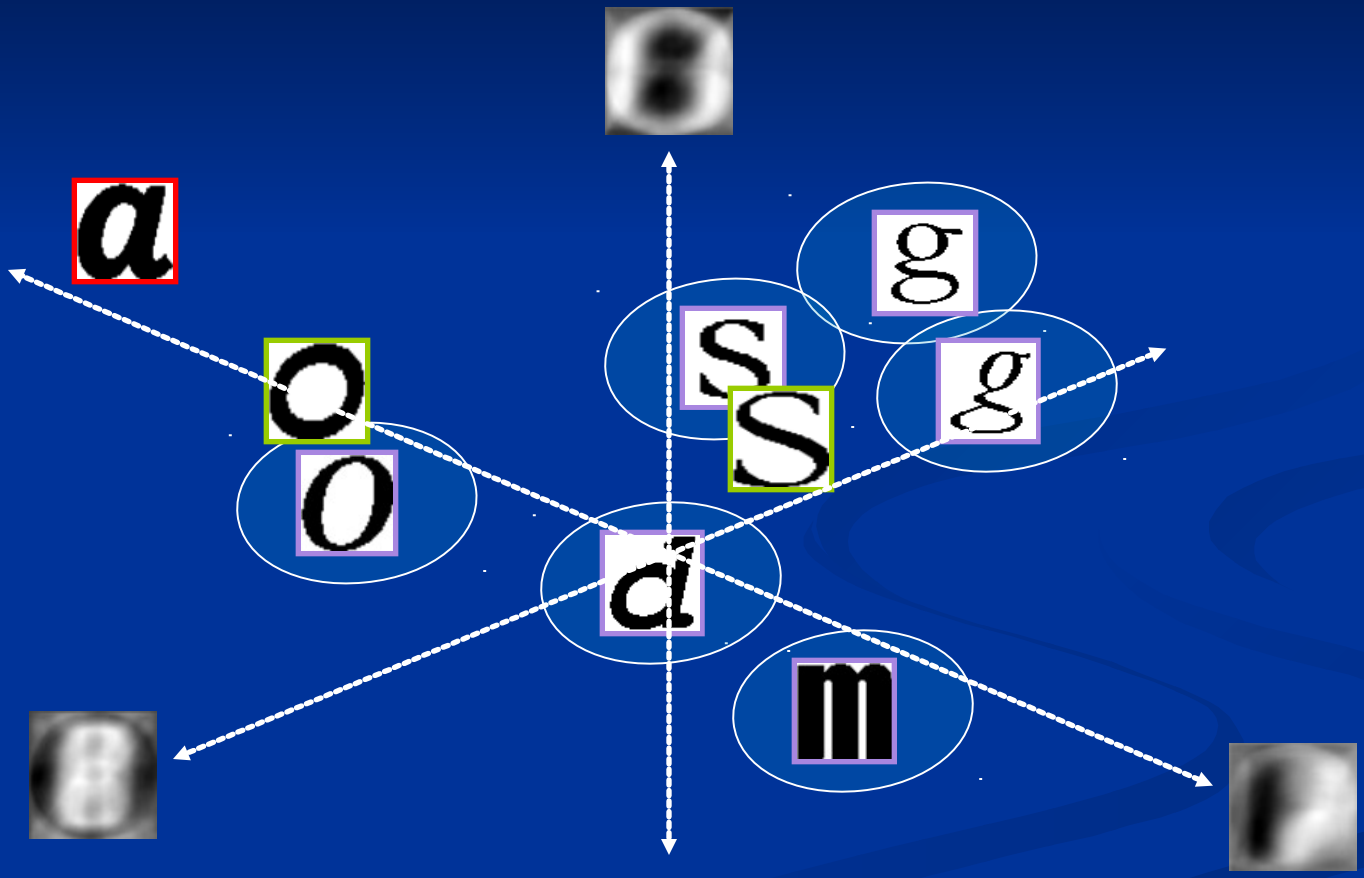
Method Identification



Method Identification

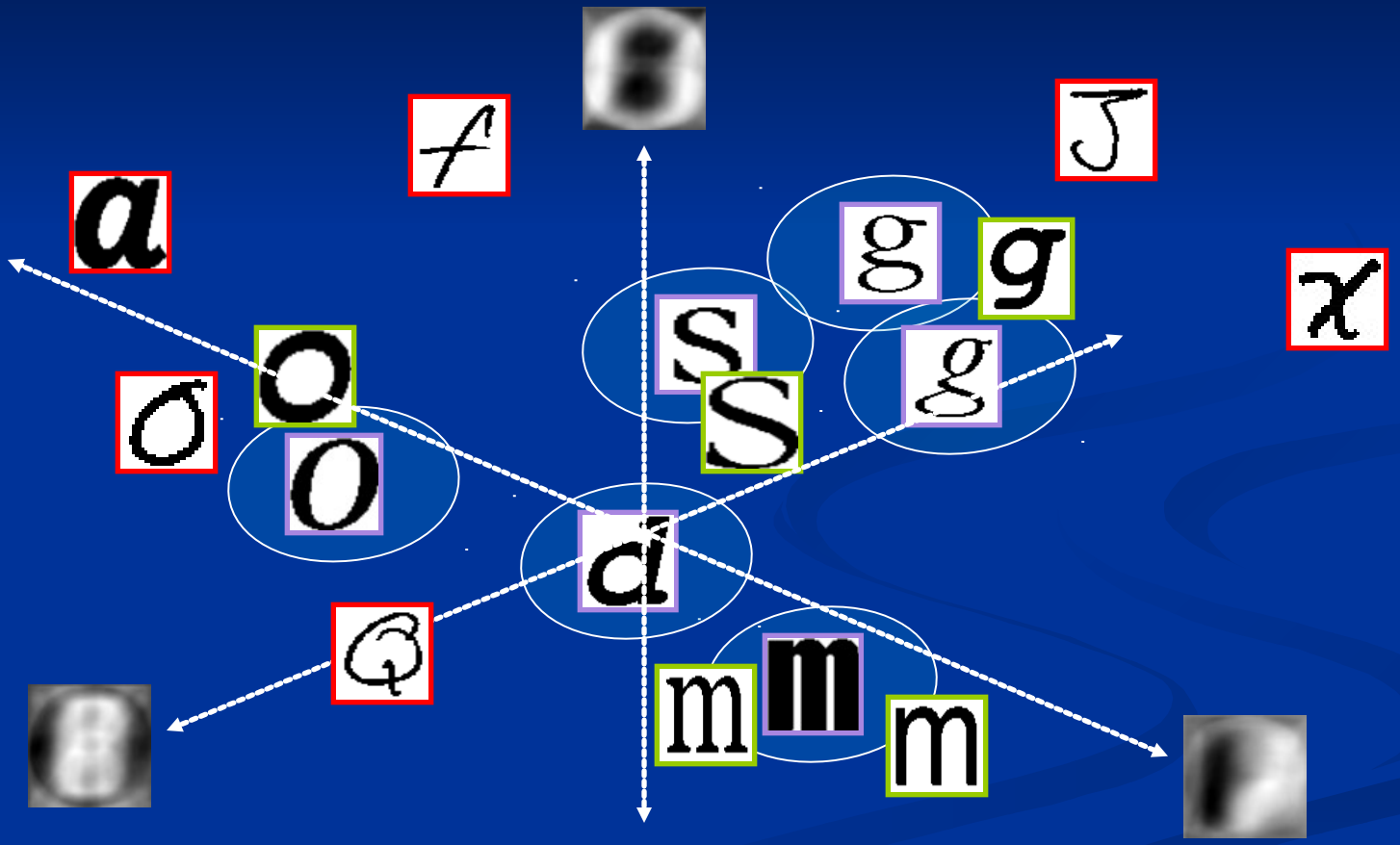


Method Identification



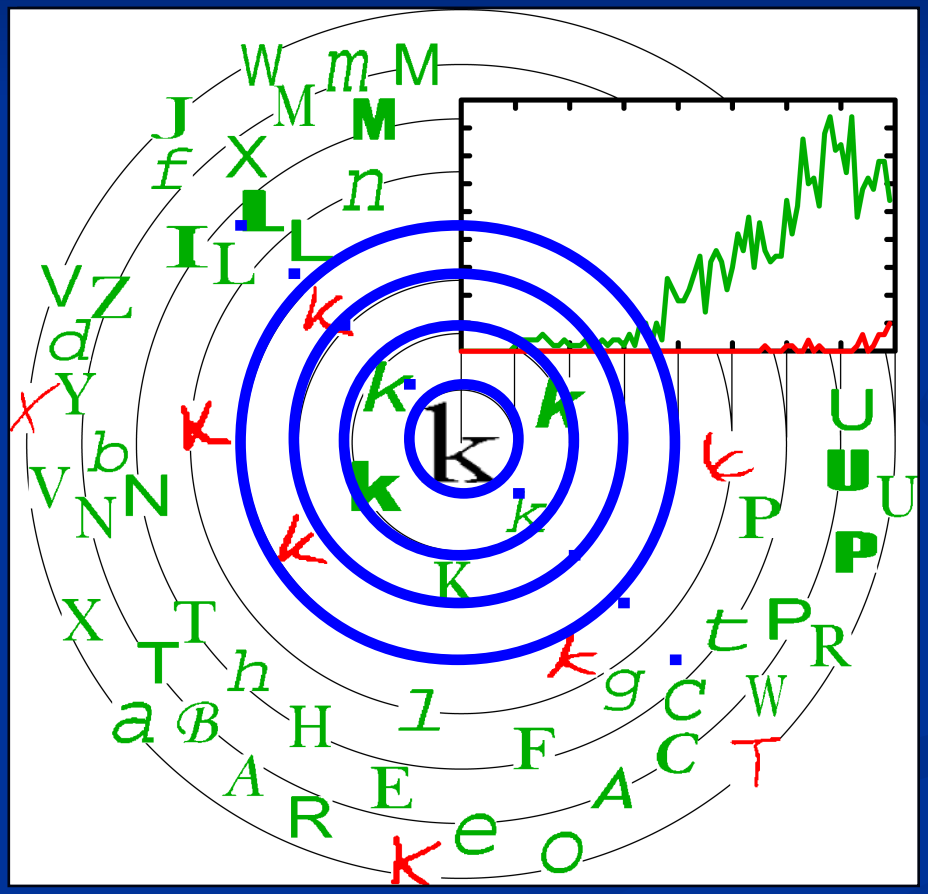
- The Use of Eigenpictures for Optical Character Recognition [Muller and Herbst 1998]

Method Identification



Method Identification

- Determining a local distance threshold via radial density



Global Target Precision

98%

Local Precision

92%

Index Construction

Archivo de la Parroquia

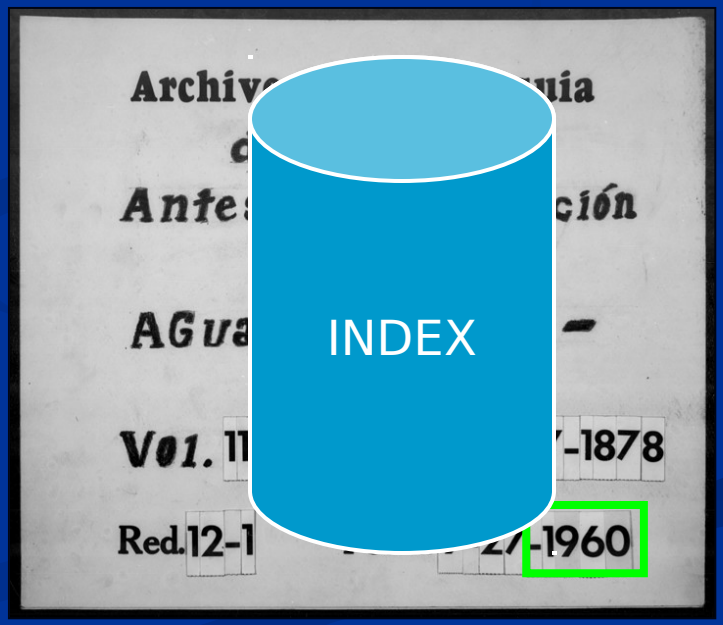
BAUTISMO

115 1877 – 1878

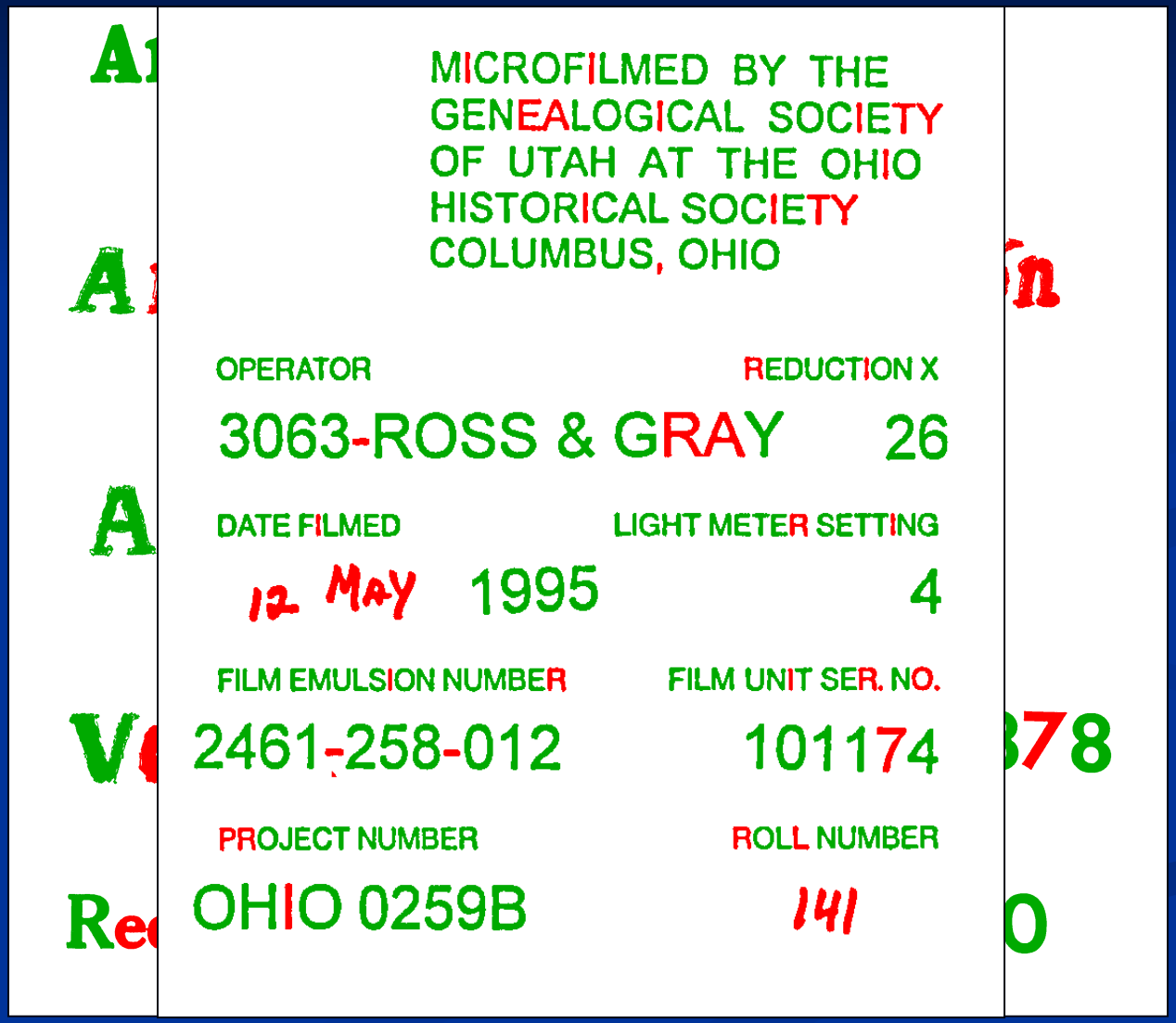
Red. 12-1 Fecha 9-27-1960

Querying

<bautismo> AND <1960>



Results


MICROFILMED BY THE
GENEALOGICAL SOCIETY
OF UTAH AT THE OHIO
HISTORICAL SOCIETY
COLUMBUS, OHIO
OPERATOR
REDUCTION X
3063-ROSS & GRAY
26
DATE FILMED
LIGHT METER SETTING
12 MAY 1995
4
FILM EMULSION NUMBER
FILM UNIT SER. NO.
2461-258-012
101174
PROJECT NUMBER
ROLL NUMBER
OHIO 0259B
141

Results

	Predicted Handwriting	Predicted Machine Print
Actually Handwriting	**88.9%**	**11.1%**
Actually Machine Print	**16.6%**	**83.4%**

Future Work

- Robust preprocessing and segmentation
- Incorporate lexical, font, and style context.
  - Metadata about indexed terms: script, language, meaning
- Specialize the set of representative machine print connected components

Conclusions

- Connected component level method identification
- Progress towards automatic titleboard indexing