ICPR 2020
Textual-Content-Based Classification of Bundles of Untranscribed Manuscript Images

Jose Ramón Prieto, Vicente Bosch, Enrique Vidal, Carlos Alonso, M. Carmen Orcero, Lourdes Marquez

Universitat Politècnica de València, Insituto Andaluz del Patrimonio Historico

Instituto Andaluz del Patrimonio Histórico
CONSEJERÍA DE CULTURA Y PATRIMONIO HISTÓRICO

Outline

Motivation

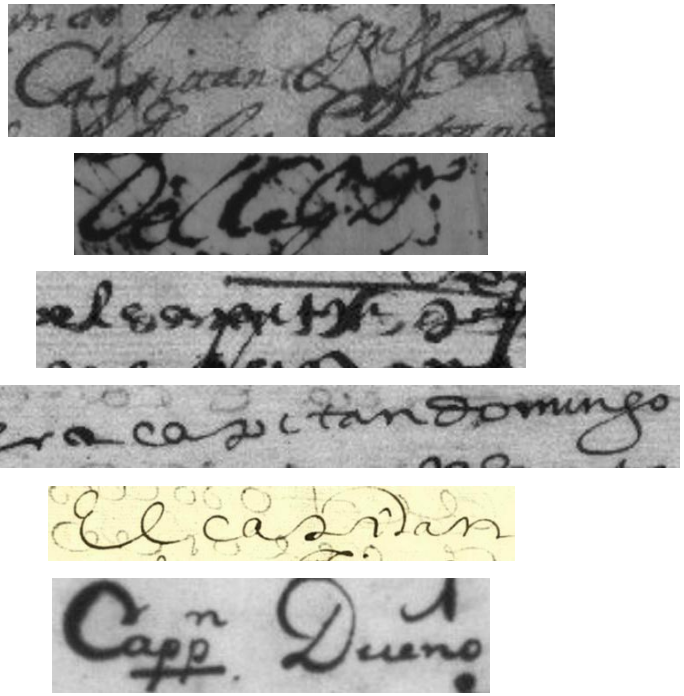
- Document Classification
  - Classify bundles of handwritted images
  - Very difficult to transcribe

Carabela Corpus Sample Pages

Motivation

- Document Classification
  - Classify bundles of handwritted images
  - Very difficult to transcribe

Carabela Corpus Sample Pages



Examples of important difficulties exhibited by Carabela

Background - PrIx

- The Probabilistic Indexing (PrIx) framework was proposed to deal with the intrinsic word-level uncertainty of handwritten text

Background - PrIx

- The Probabilistic Indexing (PrIx) framework was proposed to deal with the intrinsic word-level uncertainty of handwritten text

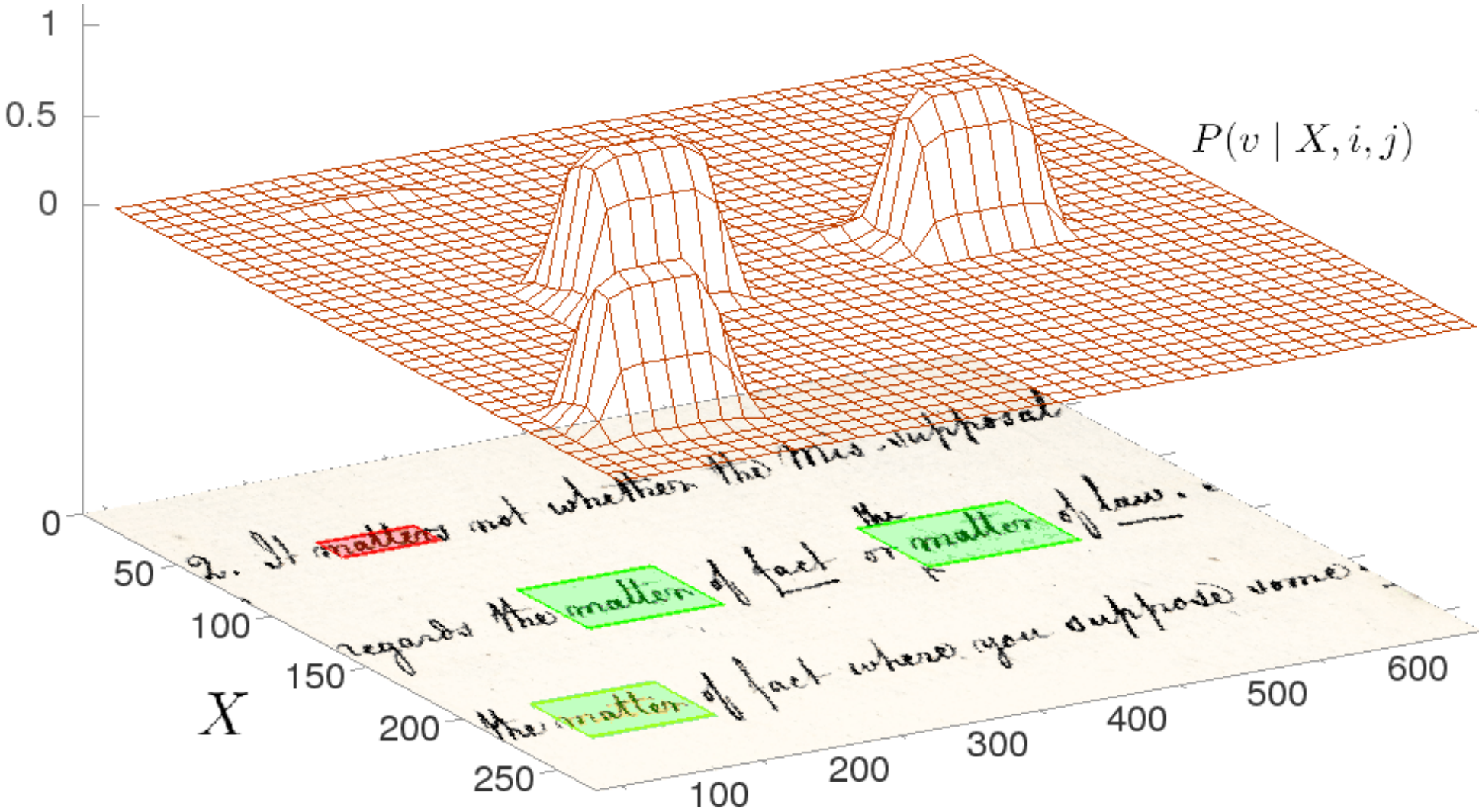
- Text elements are referred to as *"pseudo-words spots"*

Background - PrIx

- The Probabilistic Indexing (PrIx) framework was proposed to deal with the intrinsic word-level uncertainty of handwritten text
- Text elements are referred to as *"pseudo-words spots"*
- Relevance probability (RP), $P(R|X, v)$ of each image X for each pseudo-word v

$$P(R \mid X, v) = \sum_{i,j} P(R, i, j \mid X, v) \approx \max_{i,j} P(v \mid X, i, j) \approx \max_{b \in X} P(v \mid X, b)$$

Background - PrIx



Background - Representation

- The Bag of Words model is assumed
- Information Gain to select most relevant words, based on the following frequencies:
  - $f(t_v)$: the number of documents in D which contain v
  - $f(c, t_v)$: the number of documents of class c which contain v
- Tf·Idf which is based on the following frequencies
  - $f(t_v)$ as in Information Gain
  - $f(D)$: the total number of words in D
  - $f(v, D)$: the number of times v appears in D

Estimating Word and Document Frequencies from PrIx

Word and document frequecies needed to compute IG and Tf·Idf are estimated from image Relevance Probabilities as follows:

$$f(D) \equiv n(X)$$

$$f(t_v) \equiv m(v, \mathcal{X})$$

$$f(c, t_v) \equiv m(v, \mathcal{X}_c)$$

$$f(v, D) \equiv n(v, X)$$

$$E[n(X)] = \sum_{x \in X} \sum_{v} P(R \mid x, v)$$

$$E[n(v, X)] = \sum_{x \in X} P(R \mid x, v)$$

$$E[m(v, \mathcal{X})] = \sum_{X \in \mathcal{X}} \max_{x \in X} P(R \mid x, v)$$

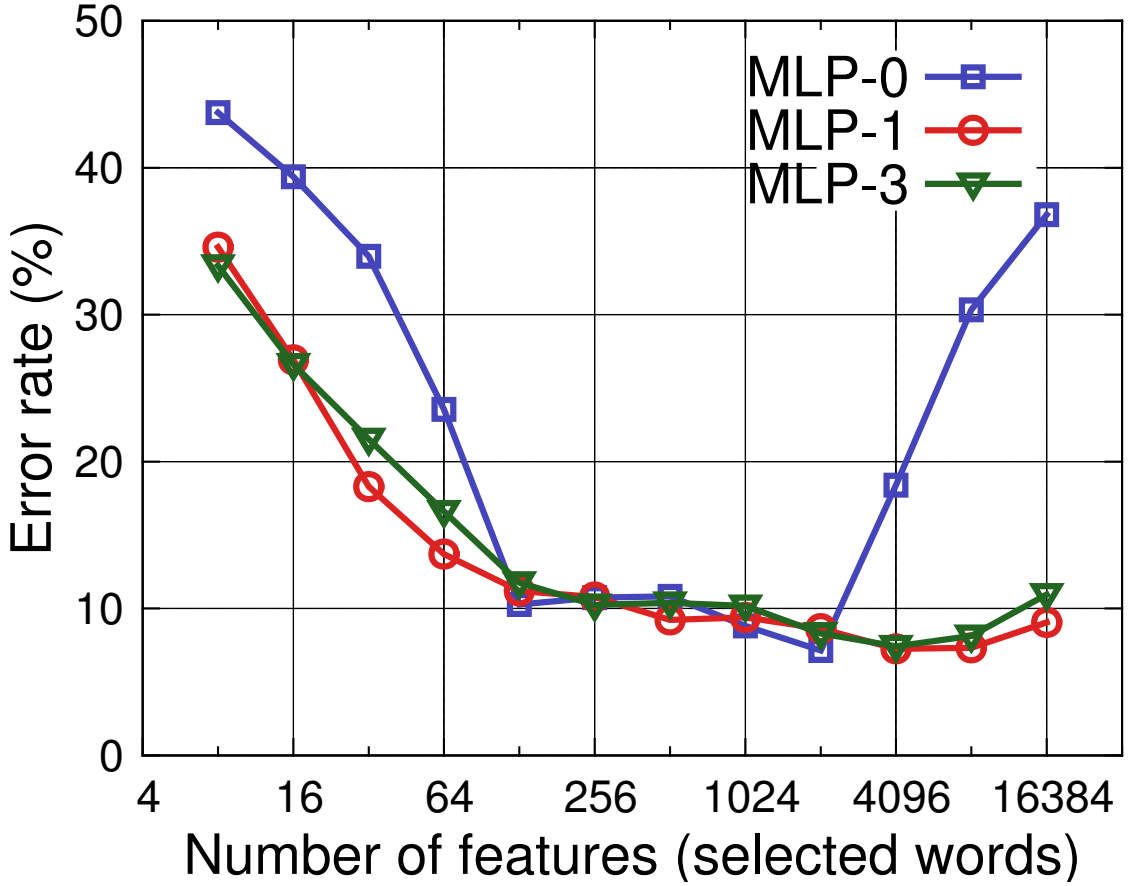
Document Image Classification

Optimal prediction of the class of an image document X is achieved by the maximum class posterior.

- Document Image Representation: Bag of Words model based on words and documents frequencies esimated from PrIx
- Multinomial Naive Bayes: A linear classifier equivalent to a (plain) perceptron
- Multilayer Perceptrons trained with cross entropy loss
  - MLP-0: 0-hidden-layers MLP
  - MLP-1: a proper MLP including one hidden layer with 64 ReLU neurons and batch normalization
  - MLP-3: 3 hidden layers including 16, 32 and 64 ReLU neurons and batch normalization

Results - MLPs

Classification error rate for three MLP classifiers



Best classification error rate (7.1%) obtained by the plain perceptron (MLP-0), for a relatively large vocabulary of the 2048 words with largest Information Gain.

Results - MLP & MNB

Classification error rates using C-MNB, compared with MLP-0, for increasing sizes of the vocabulary selected using Information Gain.

	Number of Features							
Error (%)	16	64	256	1024	2048	4096	9192	16384
C-MNB	42.6	44.5	26.5	20.0	16.1	14.8	15.5	20.6
MLP-0	39.4	23.6	10.7	8.8	7.1	18.4	30.3	36.8

All MLP based methods provide, in general, better results than the C-MNB method.

Conclusions

- Presented an approach that is able to perform textual-content-based document classification directly on documents of untranscribed handwritten text images.
- Overcomed the need to explicitly transcribe manuscripts, which is generally unfeasible for large collections

Future Works

- Better representation using geometric information of pseudo-words
- Improvise the term selection method to get accurate results with smaller vocabularies.