Transkribus User Conference

Keyword Spotting in Large Scale Documents

Alejandro H. Toselli and Enrique Vidal

[ahector,evidal]@prhlt.upv.es

Pattern Recognition and Human Language Technology

Reseach Center

READ

Universitat Politècnica de València

Spain

November, 2017

Index

Index

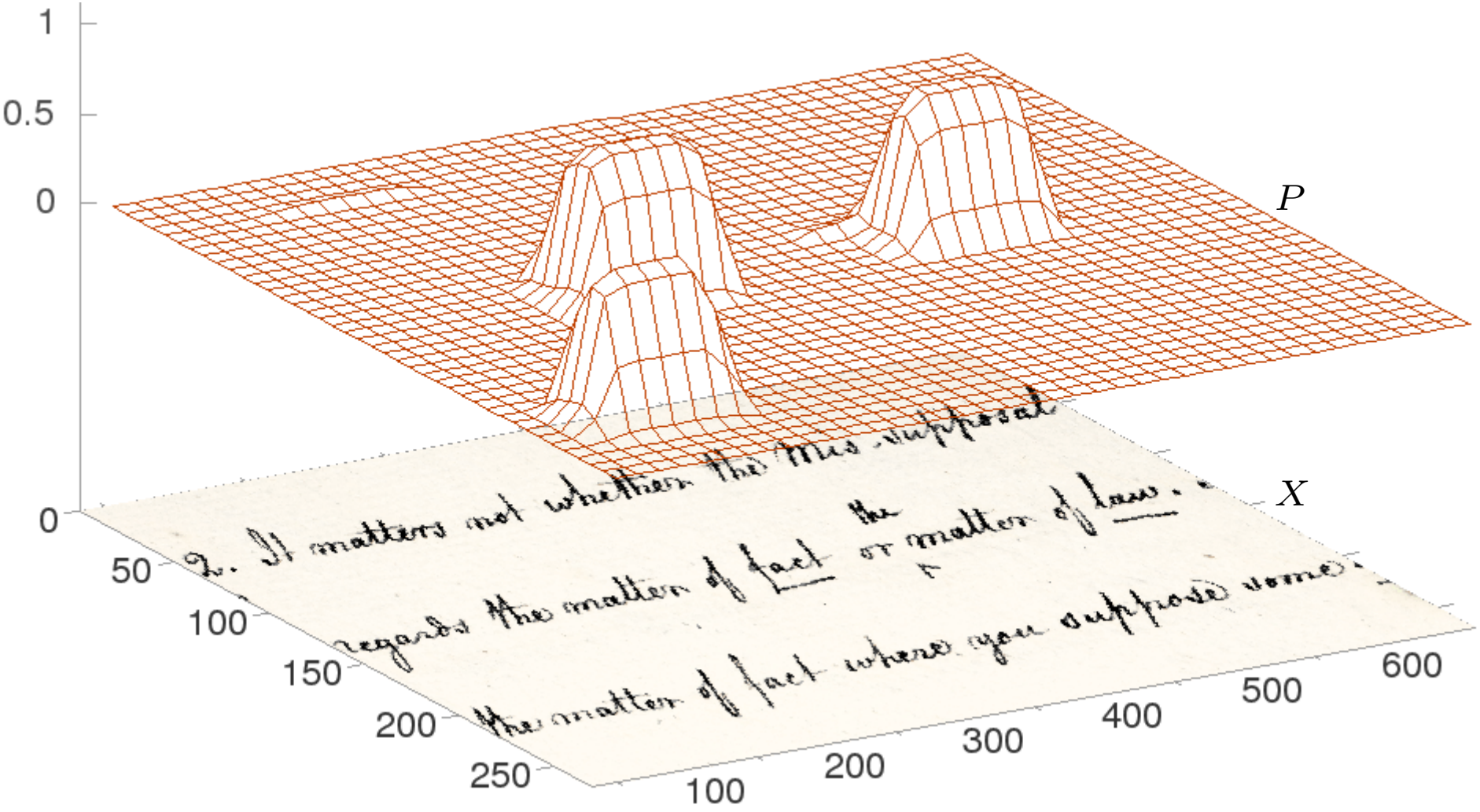
Textual access to Untranscribed Manuscripts

- Massive text image collections have been compiled by libraries and archives all over the world, but their textual content remains practically inaccessible
- If perfect or sufficiently accurate text image transcripts were available, image textual context could be straightforwardly indexed for plaintext textual access.
- But manual or even interactive-predictive, assisted transcription is entierely prohibitive to deal with massive image collections
- And fully automatic transcription results lack the level of accuracy needed for useful text indexing and search purposes

Good news: *indexing and textual search* can be directly carried out on *untrascribed images*, as we will see now.

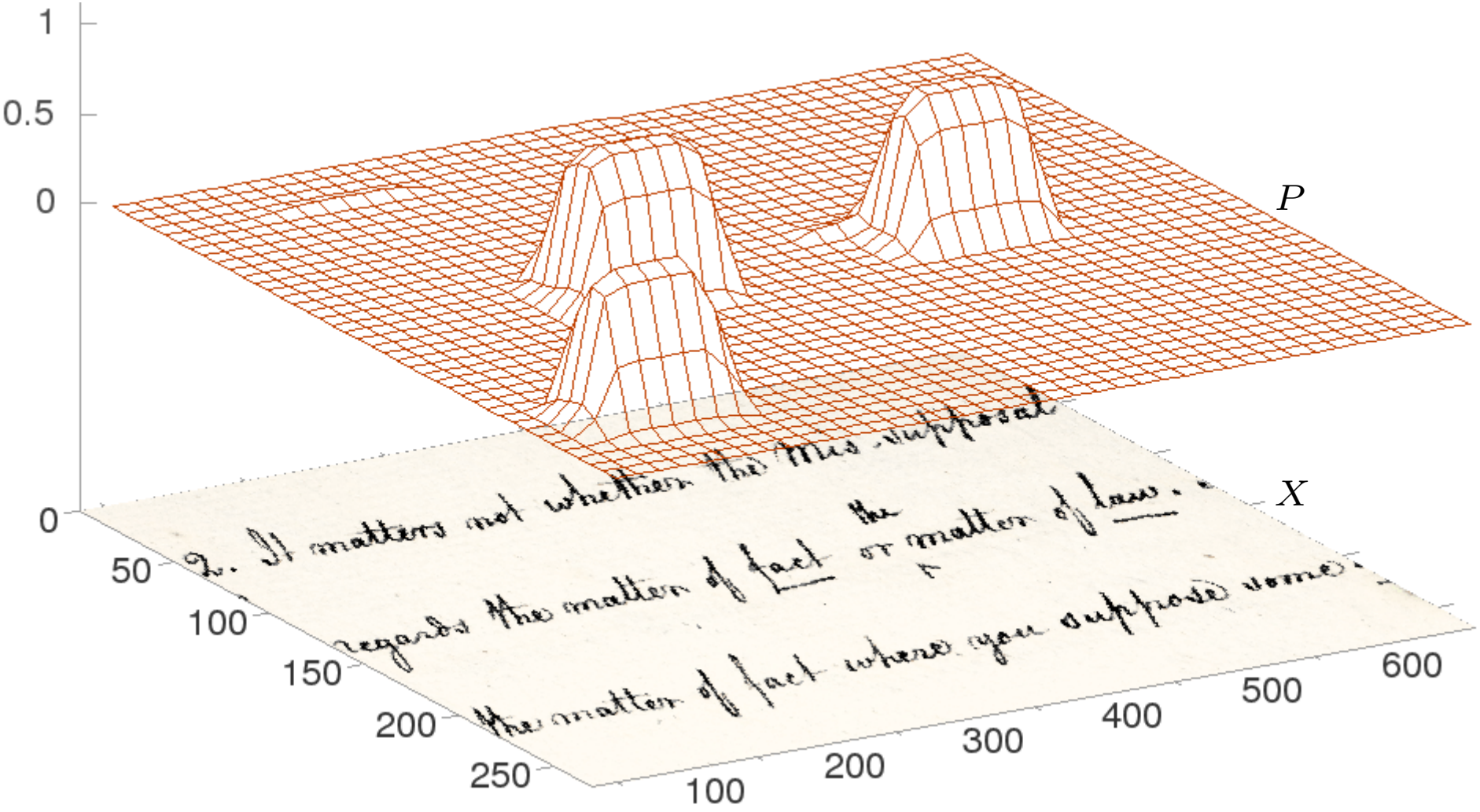
Index

Handwritten Text Indexing and Search: Pixel-level Posteriorgram



Pixel-level posterior probabilities (P) for a text image X and word $v =$"**matter**".

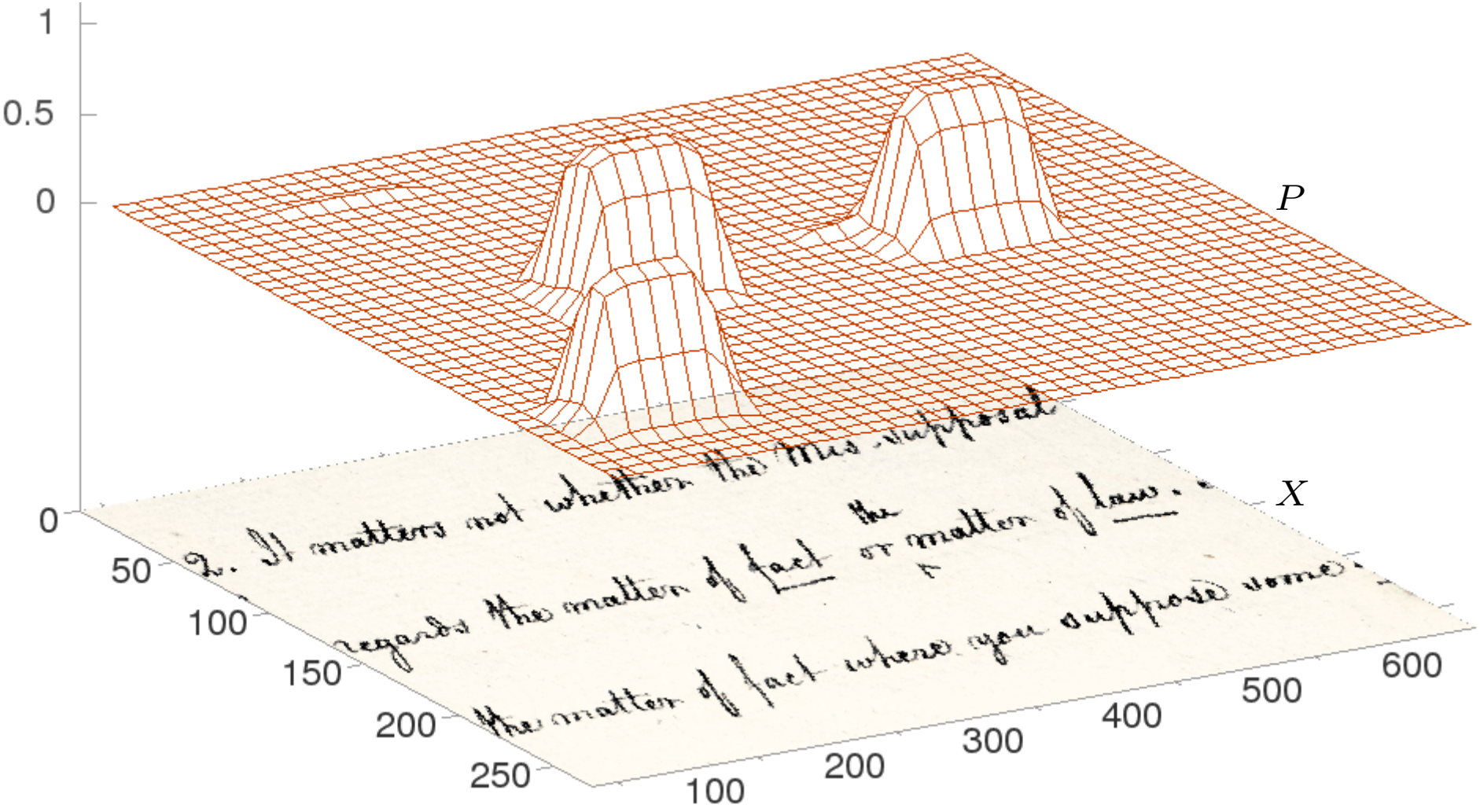
Handwritten Text Indexing and Search: Pixel-level Posteriorgram



Pixel-level posterior probabilities (P) for a text image X and word $v =$"**matter**".

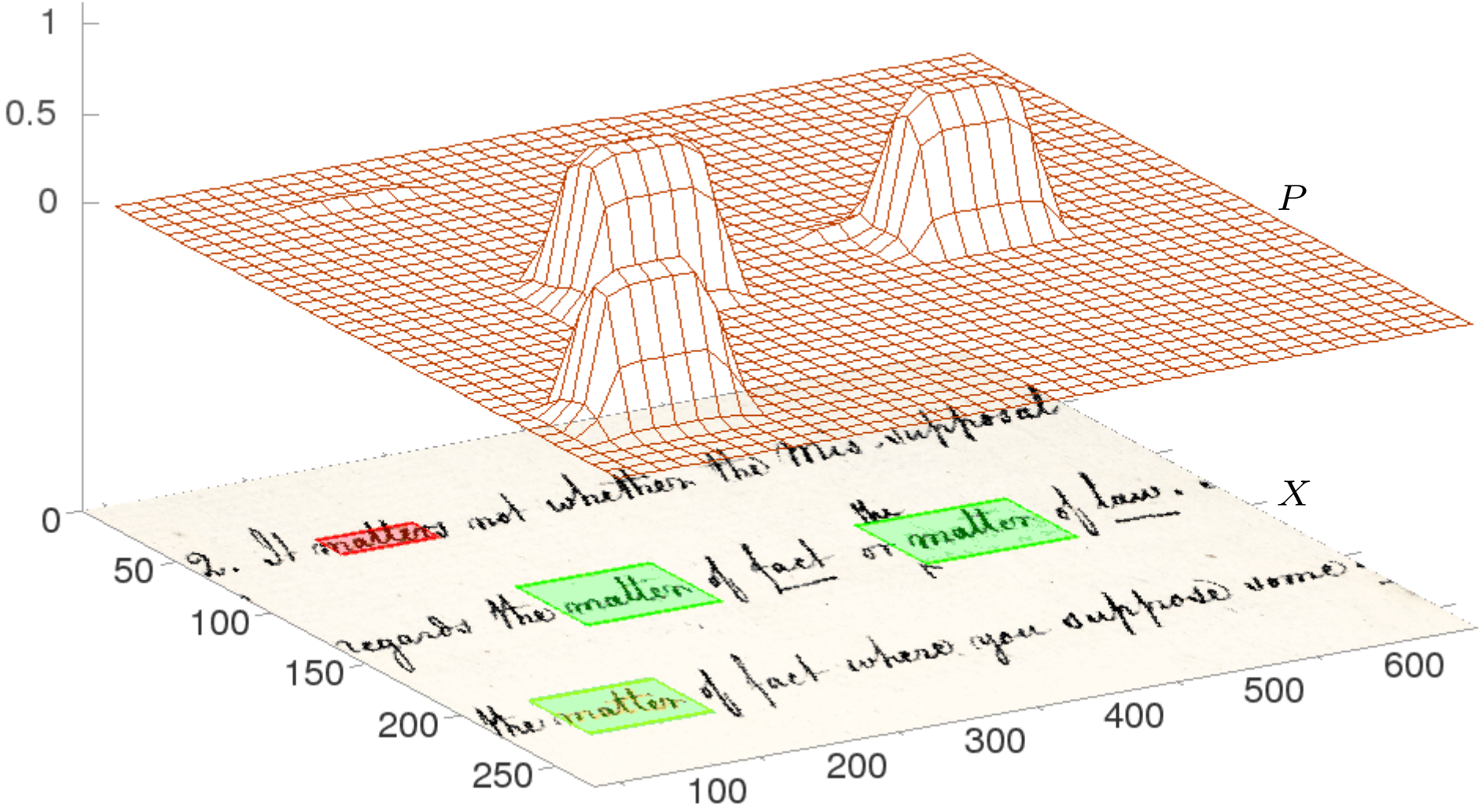
To compute P an accurate, contextual (n-gram based) *word classifier* can be used. In this example, this helped to achieve very low posteriors in a region of X around $(i = 100, j = 60)$, where a very similar (but *different*) word, "**matters**", is written.

Handwritten Text Indexing and Search: Pixel-level Posteriorgram



Directly computing and using a full pixel-level posteriorgram would entail a formidable computational load and would require prohibitive amounts of indexing storage.

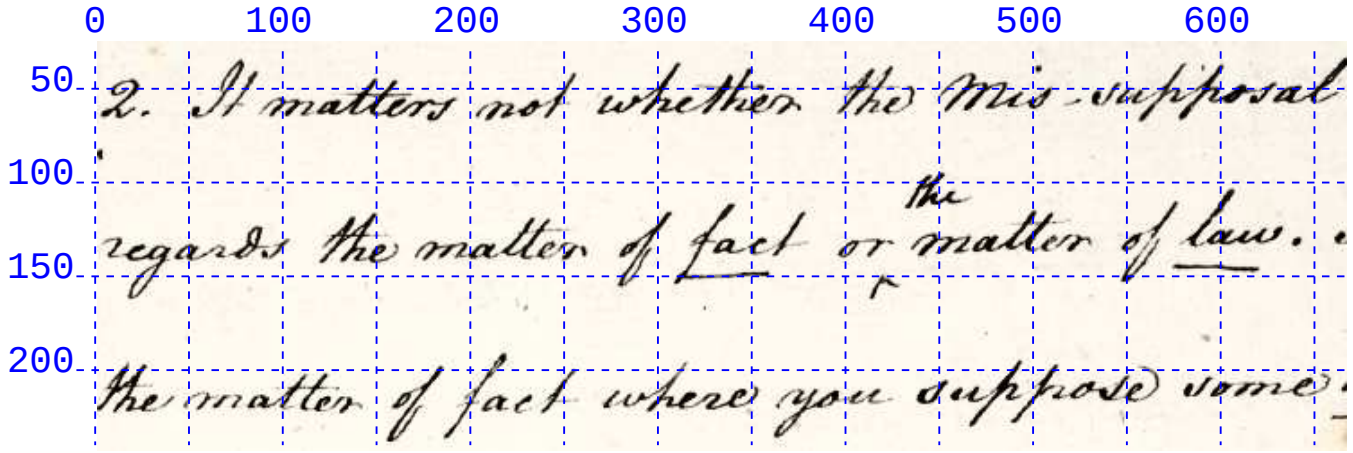
Probabilistic Word Indexing from the Posteriorgram



Directly computing and using a full pixel-level posteriorgram would entail a formidable computational load and would require prohibitive amounts of indexing storage.

But, for each word, image region *relevance probabilities* and *locations* are easily derived from the Posteriorgram – and used to probabilistically index the word in an efficient way.

Probabilistic Index: Example



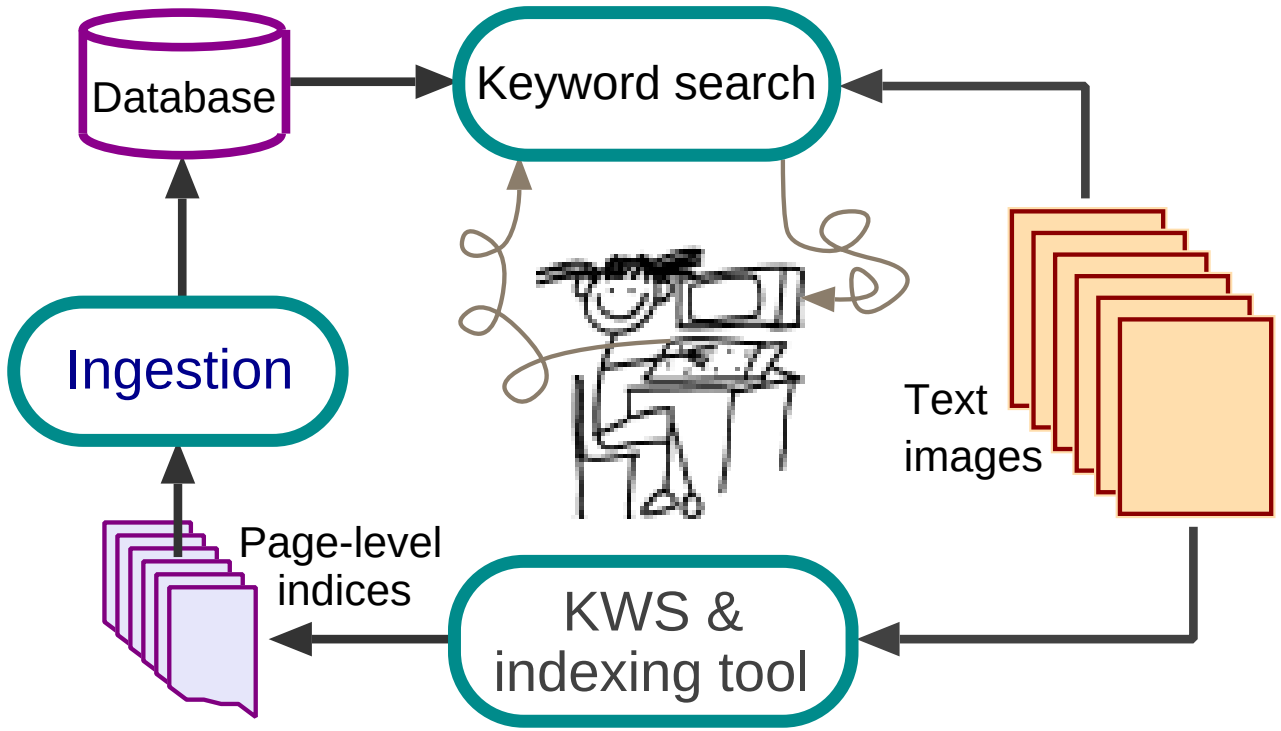

```
# pageID="Bentham-071-021-002-part"
#    keyword confid    bounding box
#
           2  0.929    1  36  20 31
          21  0.064    1  36  24 31
          IT  0.982   33  36  27 31
          IF  0.012   33  36  26 31
     MATTERS  0.989   77  36  99 31
      MATTER  0.011   77  36  93 31
         NOT  0.999  216  36   7 31
     WHETHER  1.000  256  36  99 31
         THE  0.997  389  36  33 31
MIS-SUPPOSAL  1.000  455  36 193 31

         THE  0.927  430  88  30 31
          HE  0.056  434  88  25 31
         ...  ...      ...     ...
```

```
REGARDS  0.857    5 115  84 31
REWARDS  0.138    5 115  90 31
    THE  0.993  110 115  43 31
 MATTER  0.998  160 115  93 31
     OF  0.996  271 115  23 31
   FACT  0.999  306 115  49 31
     OR  0.973  377 115  37 31
     ON  0.021  377 115  42 31
 MATTER  0.990  425 116 100 31
     OF  0.995  542 115  25 31
     BY  0.407  575 115  30 31
    ANY  0.175  575 115  55 31
    ...  ...      ...     ...
    LAW  0.032  575 115  36 31
    LAY  0.031  575 115  55 31
    ...  ...      ...     ...
    PAY  0.012  575 115  59 31
```

```
     THE  0.990    1 198  28 31
  MATTER  0.934   61 198  64 31
      OF  0.988  141 198  28 31
    FAST  0.367  182 198  62 31
     FAR  0.186  182 198  36 31
     ...  ...      ...     ...
    FACT  0.017  182 198  46 31
      AS  0.142  200 198  29 31
     HAS  0.022  200 198  29 31
   WHERE  0.992  255 198  90 31
     YOU  0.761  365 198  45 31
    YOUR  0.030  365 198  47 31
    GOES  0.064  372 198  45 31
 SUPPOSE  0.975  429 198 120 31
SUPPOSED  0.024  429 198 125 31
    SOME  0.834  570 198  78 31
  SOONER  0.016  576 198  83 31
     ONE  0.109  580 198  65 31
      ME  0.022  620 198  22 31
```

Index

Probabilistic Text Image Indexing and Search: System Diagram



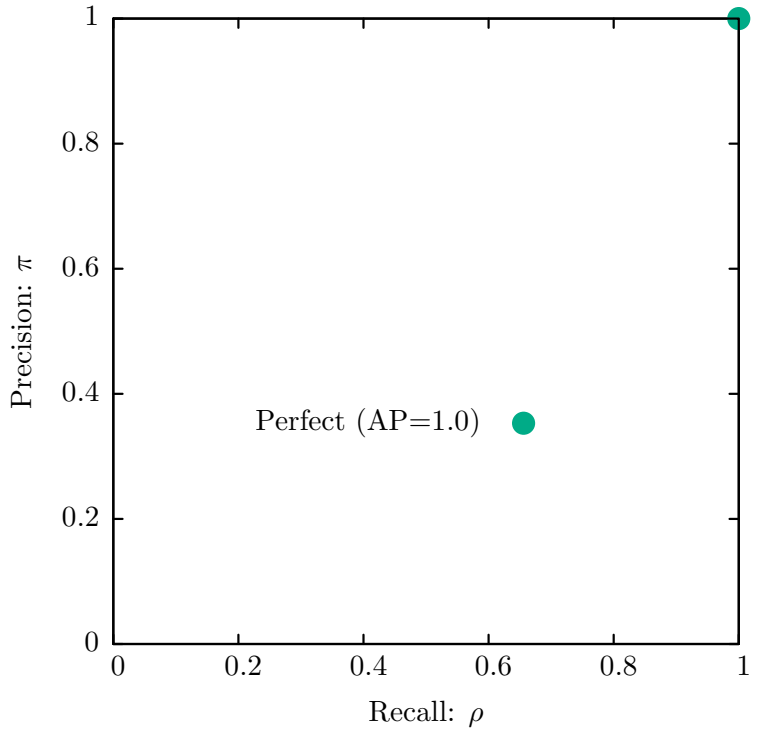
- *"KWS & indexing tool"*: Off-line pre-computation of probabilistic indices
- *"Ingestion"*: Off-line creation of the actual database. Typically a simple and computationally cheap process
- *"Keyword search"*: On-line user query analisys, find the requested information and present the retrived images. Short response times needed.

Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.

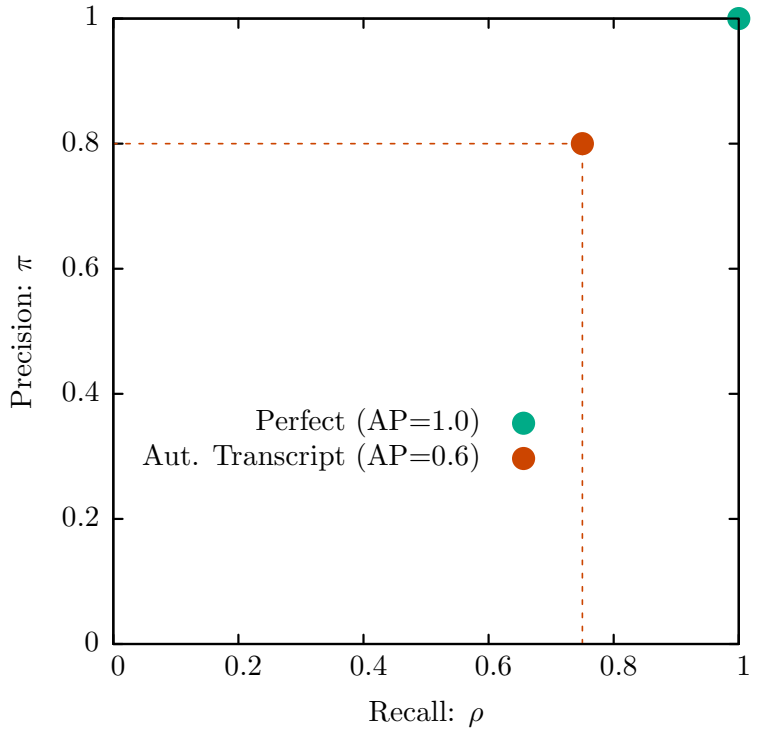


Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.



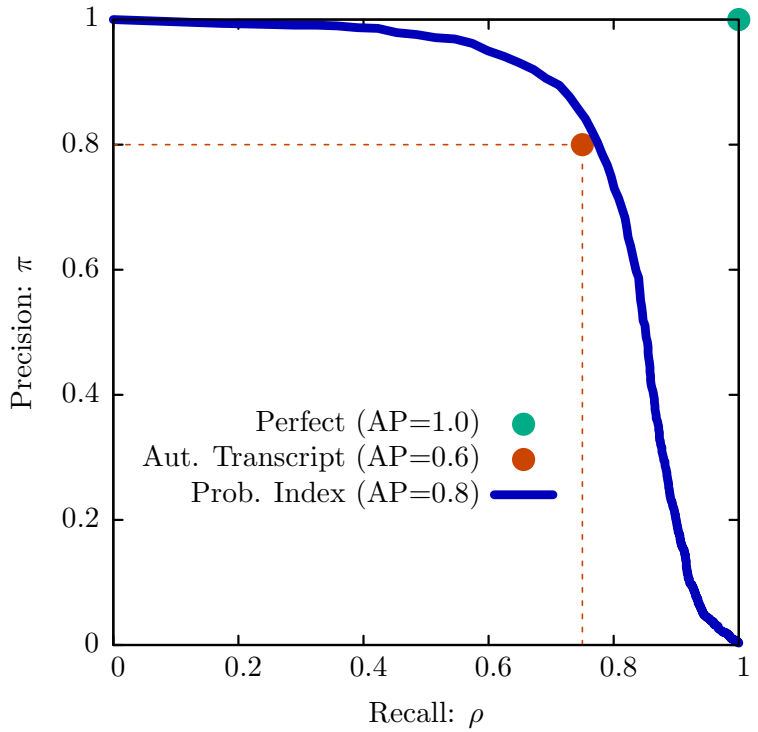
If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (pehaps something like $\rho = 0.75, \pi = 0.8$, with Averge Precision AP=0.6).

Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.



If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (pehaps something like $\rho = 0.75, \pi = 0.8$, with Averge Precision AP=0.6).

In contrast, probabilistic indexing allows for arbitrary precision-recall tradeoffs by setting a threshold on the system confidence (relevance probability)

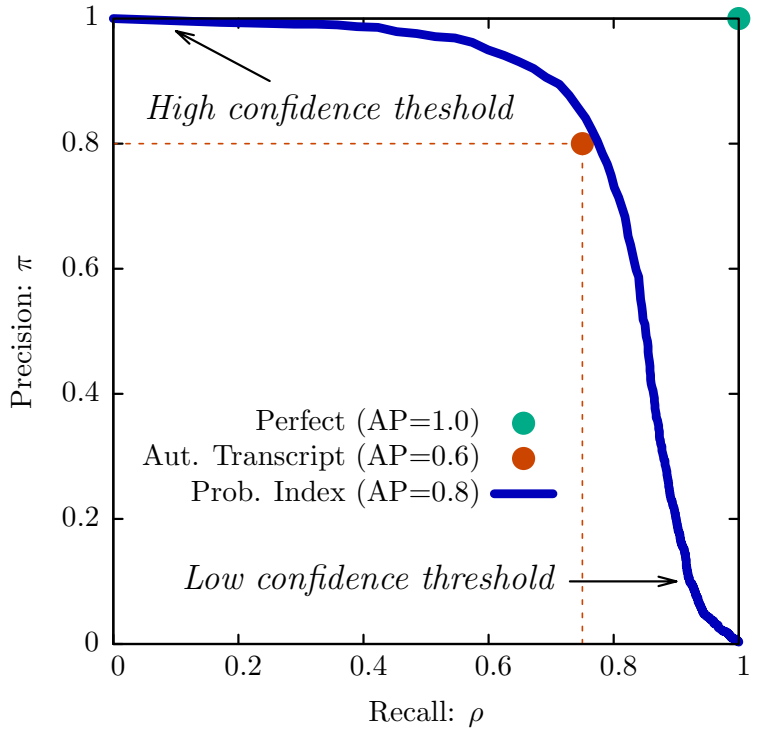
This flexible *"precision-recall tradeoff model"* obviously allows for better search and retrieval performance than naive plaintext searching on automatic noisy transcripts.

Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.



If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (pehaps something like $\rho = 0.75, \pi = 0.8$, with Averge Precision AP=0.6).

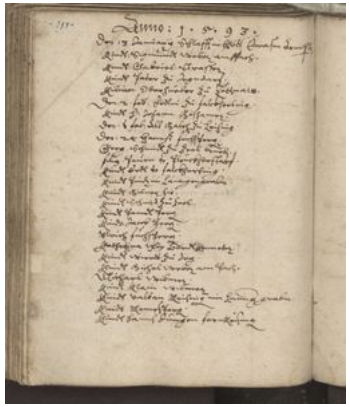
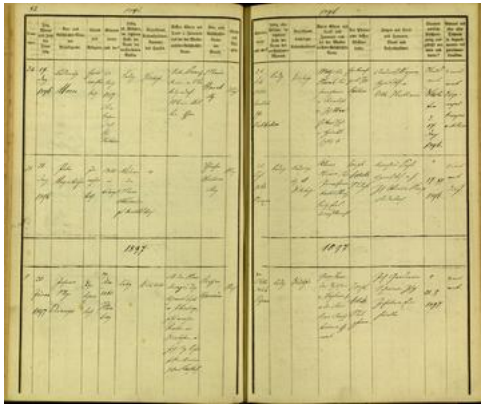
In contrast, probabilistic indexing allows for arbitrary precision-recall tradeoffs by setting a threshold on the system confidence (relevance probability)

This flexible *"precision-recall tradeoff model"* obviously allows for better search and retrieval performance than naive plaintext searching on automatic noisy transcripts.

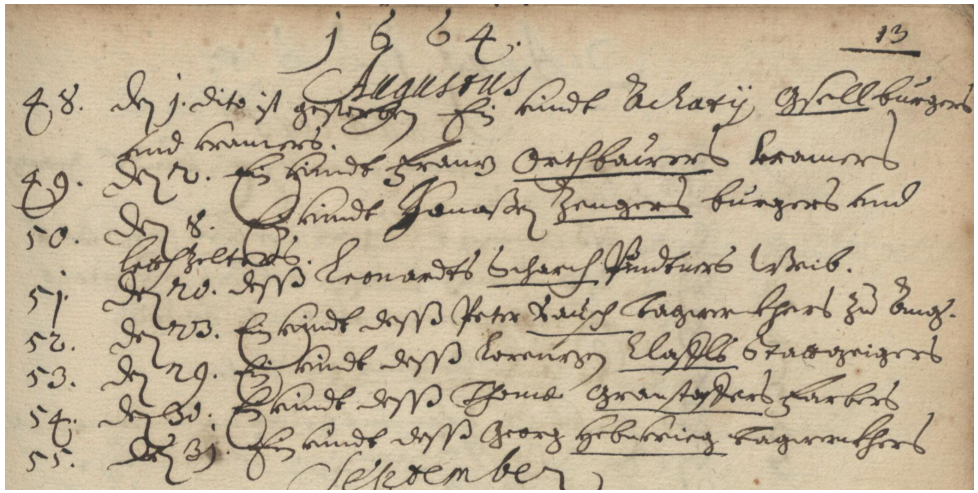
Index

"PASSAU" Dataset

XVI-XVIII century collection of historical records. $26,000$ images, written in German.

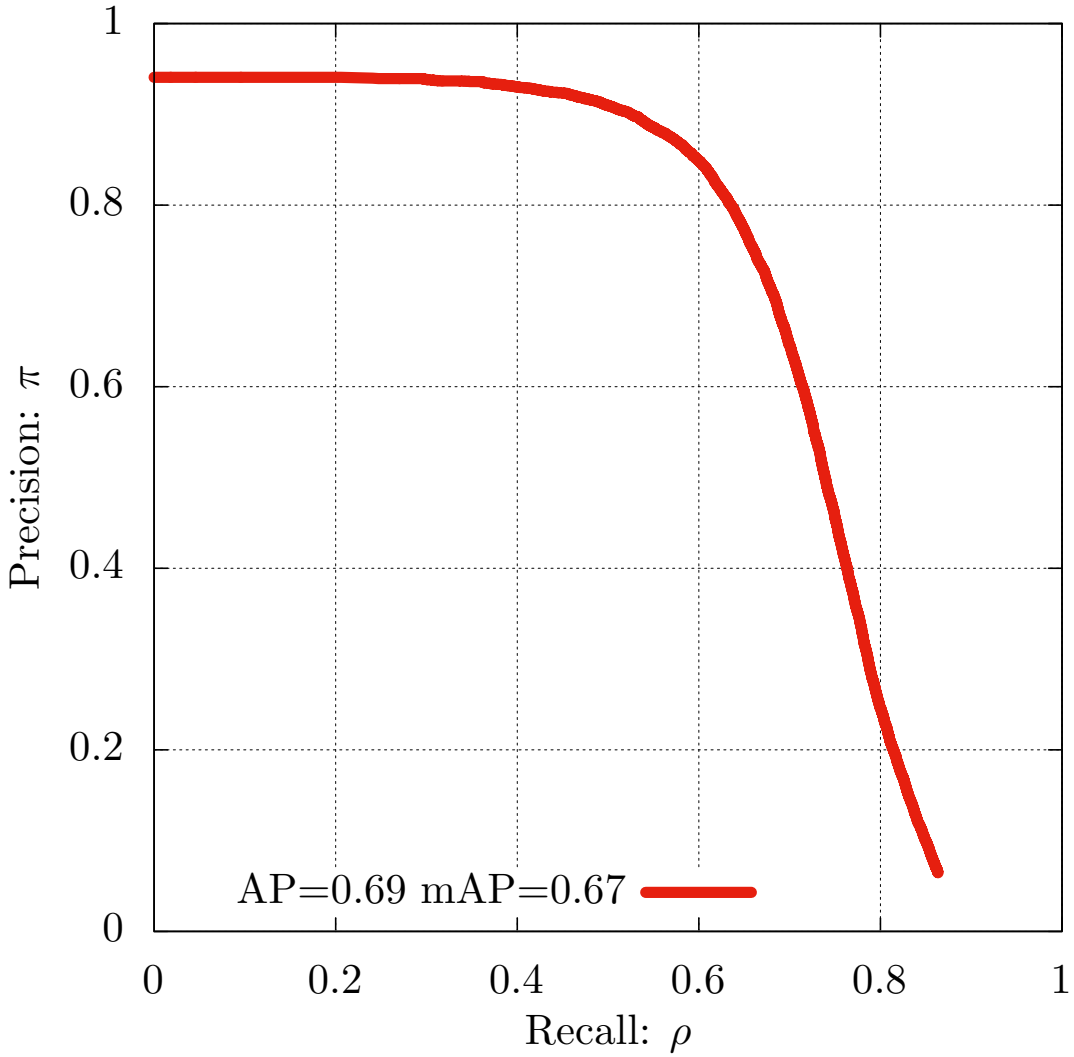
Preliminary experiments on 200 and 91 (train and test) pre-selected pages.



Number of:	Total
Pages	291
Lines	46 069
Running words	86 678
Lexicon size	23 024
Character set size	221

Remark: Text line detetction was fully supervised.

Passau: Evaluation Results



- Case & diacritics folded queries: $5\,836$.
- *Average & mean Average Precision* (AP and mAP).
- CLs obtained with character 6-gram LM and Viterbi decoding with beam width 15.

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Margareta || Margareth || Margaretha || Margaretham || Margaritha

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Margareta || Margareth || Margaretha || Margaretham || Margaritha

Joseph && *Maria*

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Margareta || Margareth || Margaretha || Margaretham || Margaritha

Joseph && *Maria*

Passau && Anna

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Margareta || Margareth || Margaretha || Margaretham || Margaritha

Joseph && *Maria*

Passau && Anna

(*Johann* || *Anna*) *1798*

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*

Margareta || Margareth || Margaretha || Margaretham || Margaritha

Joseph && *Maria*

Passau && Anna

(*Johann* || *Anna*) *1798*

(*Johann* || *Anna*) – *1798*

Passau: Indexing and Search Live Demonstration

PRHLT PASSAU Search Interface
http://transcriptorium.eu/demots/kws-Passau/

A small sample of query possibilities:

Apr || *April* || *Aprili* || *Aprilis*
Margareta || Margareth || Margaretha || Margaretham || Margaritha

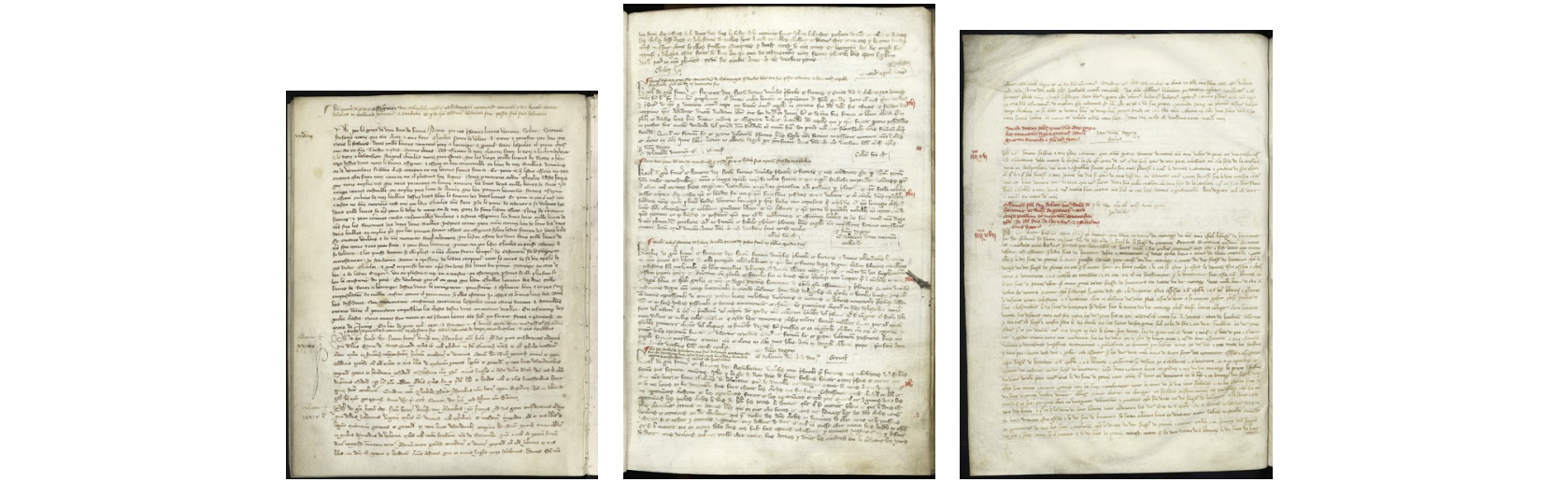
Joseph && *Maria*
Passau && Anna

(*Johann* || *Anna*) *1798*
(*Johann* || *Anna*) – *1798*
[*Anna Maria*]

Index

Chancery Probabilistic Index: Large Scale Indexing Example

XIV-XV century medieval registers produced by the French royal chancery. More than $70\,000$ document images written in Latin and French.



Philippes, etc. Savoir faisons à touz presenz et avenir que, comme ou traitié du mariage de nostre amé et feal le seigneur de Harecourt et de Ysabeaul, sa fame a present, file de nostre amé et feaul le seigneur de Partenay, certainnes convenances eussaint esté traite--es et acordées entre les dites parties par leur consoil de leurs amis, et sus ycelles convenances eust esté et fust entre eus aucun descort, c'est assavoir que li diz sires de Harecourt dysoit et maintenoit que deux miles livres à tournois de rente annuelle et perpetu--elle que li diz sires de Partenay li avoit promises et données pour cause dou dit mariage, à avoir du dit seigneur de Harecourt après le deceps du dit seigneur de Partenay, ou cas que il auroit hoir ou hoirs mâles, et le cas se estoit jà offert, li devoient estre assises à Chas-

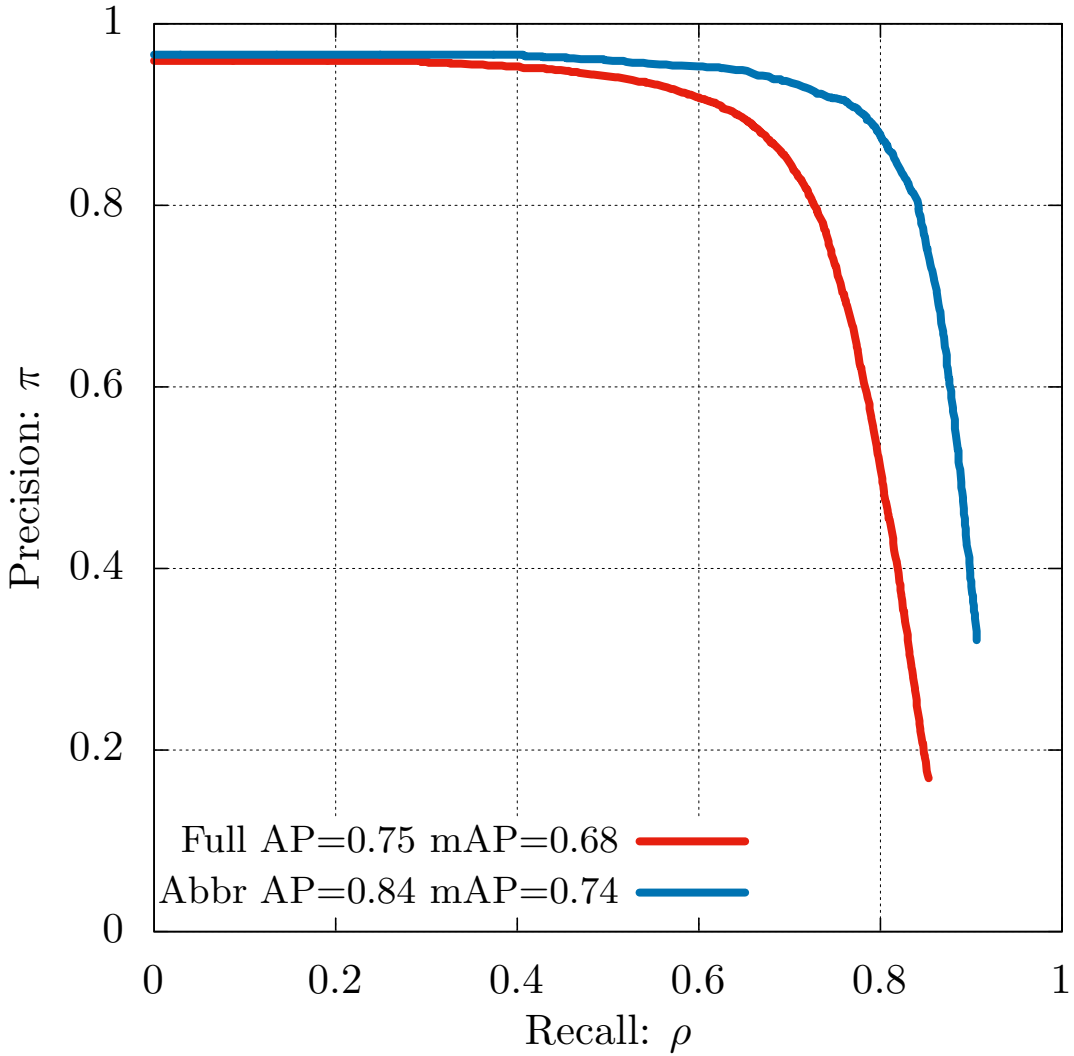
Chancery Probabilistic Index (as of Sep-2017)

Number of volumes	**167**
Number of pages	**67 413**
Number of automatically extracted lines	**3 035 550**
Estimated number of running words	**40 000 000**
Number of spots	266 301 333
Number of terms (different pseudo-words)	**28 211 224**
Spots per estimated running word (aprox.)	6.7
Storage used for images (Gb)	225
Storage used for the full index (Gb)	14
Memory used by the KWS search engine (Gb)	74
Typical query response time (seconds)	**0.1**

Recently, the last batch of about 15 000 Chancery images has been including, resulting in 199 volumes and about 83 000 page images indexed

Remark: Text line detection and extraction was fully automatic.

Chancery Probabilistic Index Evaluation: Laboratory Results



- Case & diacritics folded queries: $6\,506$ and 244 for the full and abbreviated words queries respectively.
- *Average & mean Average Precision* (AP and mAP).
- CLs obtained with character 5-gram LM and Viterbi decoding with beam width 5.

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

A small sample of query possibilities:

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

A small sample of query possibilities:

Eva

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

A small sample of query possibilities:

Eva

Alexandro

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

A small sample of query possibilities:

Eva

Alexandro

republica

res && publica

[res publica]

Chancery: Indexing and Search Live Demonstration

PRHLT HIMANIS Search Interface
http://prhlt-kws.prhlt.upv.es/himanis

A small sample of query possibilities:

Eva

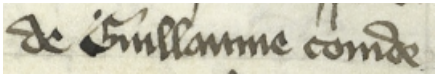
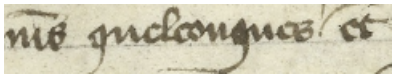
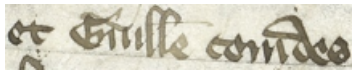
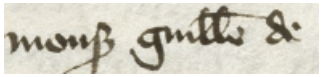
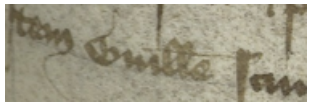
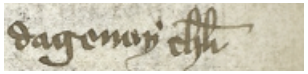
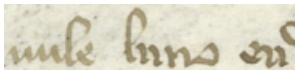
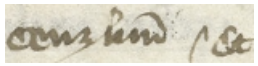
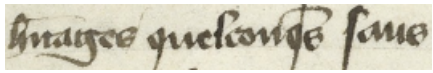
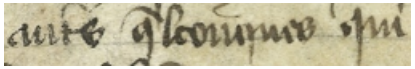
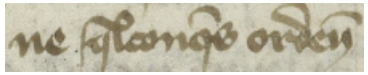
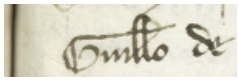
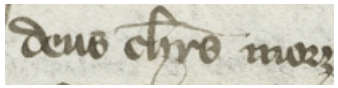
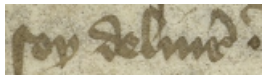
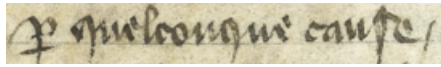
Alexandro

republica

res && publica

[res publica]

Spotting Abbreviated Words: Examples

Keyword	**Guillaume**	**chevalier**	**livres**	**quelconques**
Full form				
Abbreviated	  	  	  	  
False Positives				
Avg. Precision (AP)	0.79	0.89	0.79	0.91

Examples of modernized (expanded) keyword queries and corresponding spotting results. Selected examples of correct spotted images, both in full form and abbreviated, and examples of false positives, are shown for each query.

Index

Conclusions

- A probabilisitic framework has been introduced for indexing and searching large collections of untranscribed handwritten documents
- Empirical results in two historic collections exhibiting differnet challenges and levels of complexity assess the potential of this framework
- Two demonstrators have been implemented and made publicly available through the Internet for first-hand experience in real use
- Abreviations and other difficulties entailed by historical manuscripts are overcome: abbreviated words can get queried using just the expanded, "modernized" word forms, with excellent retrieval performance

Thanks for your attention !