Machine Learning Technologies for Handwritten Text Image Processing. Part I: Recognition and Indexing

Enrique Vidal & the PRHLT HTR team
evidal@prhlt.upv.es

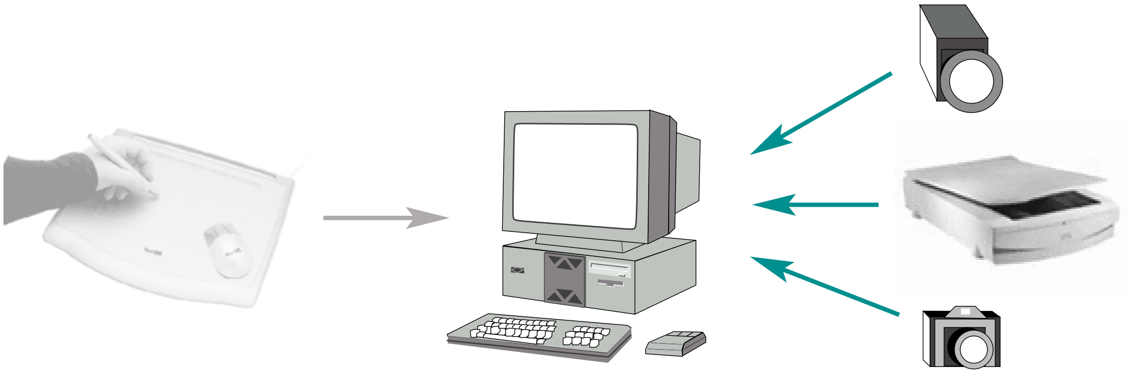
Pattern Recognition and Human Language Technology Reseach Center

Universitat Politècnica de València
Spain

June 2016
E.Vidal et al. – PRHLT/UPV

Index

Handwritten Text Recognition (HTR)



ON–LINE

Point sequence representation

(digital pen, tablet, etc.)

OFF–LINE

Bitmap (image) representation

(camera, scanner, video, etc.)

HTR and Historical Manuscripts

- Some decades ago, off-line HTR was thought to quickly become a research topic of little practical interest, since the use of text writtten on paper would soon became obsolete

  However . . .

- There are massive historical handwritten text collections stored in hundreds of kilometers of shelfs in archives and libraries

- According to some speculations, the amount of existing *handwritten* text is (much) larger than the total amount of *(original)* machine printed text, including digitally born documents!

- Important (textual) information is hidden behind digital images and these historical documents and thereby remain practically inaccessible

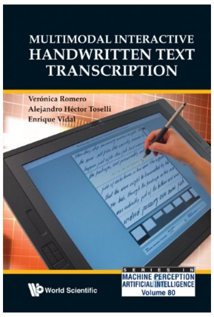
HTR may help alleviating this situation

Resources: Some Interesting Web Sites

- http://read.transkribus.eu
  The HTR READ Horizon 2000 European project
- http://transkribus.eu
  TRANSKRIBUS is a general purpose, collaborative document management and transcription tool, including automatic and assisted HTR, initially developped in TRANSCRIPTORIUM
- http://transcriptorium.eu
  The HTR TRANSCRIPTORIUM 7fp European project
- http://htk.eng.cam.ac.uk
  HTK is a time honored, well known toolkit for the deveopment of HTR systems based on N-gram language models and hidden Markov optical character models
- http://kaldi.sourceforge.net/about.html
  The (more modern, but also well known) KALDI speech recognition (and HTR) toolkit
- http://www.speech.sri.com/projects/srilm
  The SRI Language Modeling Toolkit (SRILM).

Resources: Main References Used in this Lecture

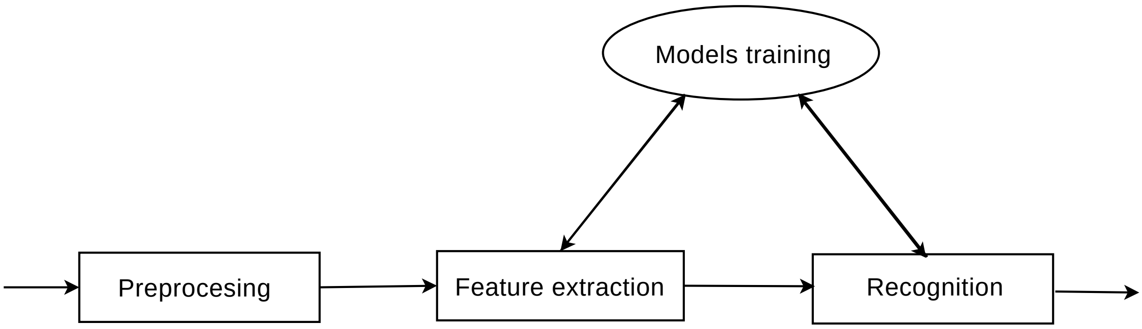
Two recent books on Interactive Pattern Recognition, Handwritten Text Recognition (HTR) and Interactive HTR:



- A.H.Toselli, E.Vidal, F.Casacuberta: *"Multimodal Interactive Pattern Recognition and Applications"*. Springer Verlag, 2011.
- V.Romero, A.H.Toselli and E.Vidal: *"Multimodal Interactive Handwritten Text Transcription"*, World Scientific, 2012.

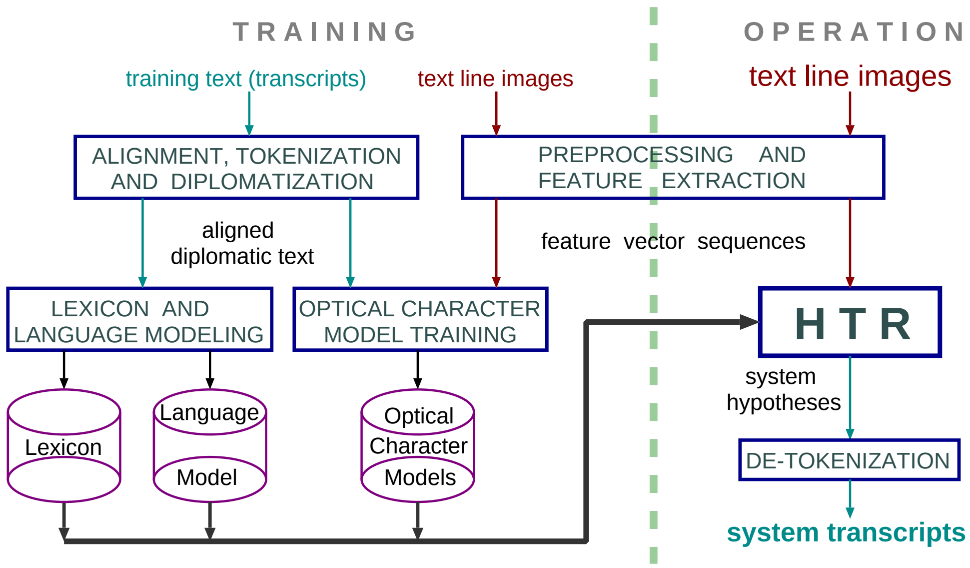
See more specific bibliography at the end of this lecture ▷42

HTR: Classical Pattern Recognition Architecture



- *Preprocessing*: noise removal, line detection and geometric normalizations
- *Feature Extraction*: sequences of attribute vectors representing local shape
- *Modeling*: optical (hidden Markov) models + (N-gram) language model
- *Recognition*: Viterbi search

Holistic, Segmentation-Free HTR System Overview



Preprocessing and Feature Extraction for Off-Line HTR

- **Page or text-block preprocessing**: *background removal*, *noise reduction*, *skew correction* and *text line detection*.
- **Line preprocessing**: *Slope/slant corrections* and (non-linear) *size normalization*.
- **Feature Extraction**: Process line-shaped images through a sliding window to obtain a *sequence of feature vectors*. Many approaches proposed; some examples:
  - Grey-level and its Gradient [Toselli et al.]
  - Grey level and local morphology heuristic features [Bunke et al.]
  - Moment-based normalization + PCA of column greylevels [Ney et al.]

Statistical Framework for HTR

Handwritten Text Recognition: Given a stream of feature vectors representing a text (line) image, x, and a set of models, $\mathcal{M}$ (optical character models, lexicon and language model), obtain a most probable transcript of x; i.e., a sequence of words $\widehat{w}$:

$$\widehat{w} \;=\; \arg\max_{w} P_{\mathcal{M}}(w \mid x)$$

Using the Bayes rule (and dropping $\mathcal{M}$ to simplify notation):

$$\widehat{w} \;=\; \arg\max_{w} \frac{P(w,x)}{P(x)} \;=\; \arg\max_{w} P(w)P(x \mid w)$$

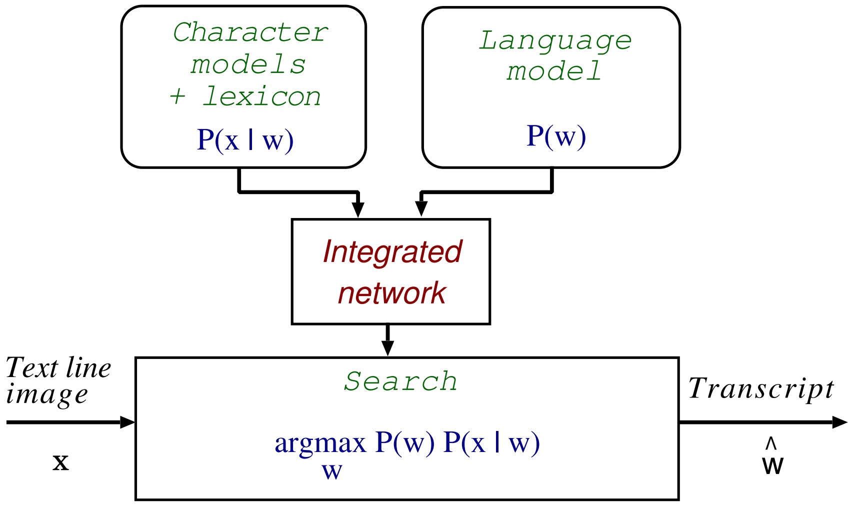
Popular models:

- $P(w)$: *N-Gram Language Model*
- $P(x \mid w)$: *Optical character HMMs* [recently also (recurrent) NNs]

Balancing models impact in practice: *Grammar Scale Factor*

$$\widehat{w} \;=\; \arg\max_{w} P(x \mid w)^{(1-\gamma)} \cdot P(w)^{\gamma} \;\equiv\; \arg\max_{w} P(x \mid w) \cdot P(w)^{\alpha}$$

HTR Integrated Architecture



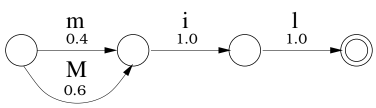
Search engine:

THE VITERBI ALGORITHM (+ beam search + ...)

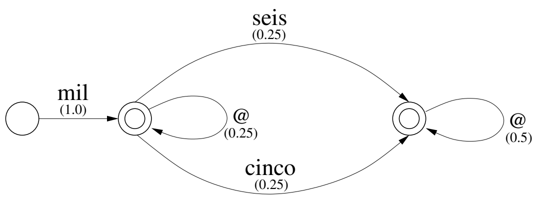
(also called “token passing” algorithm)

Lexicon, Language Model and HMM Integration (illustration)

word model

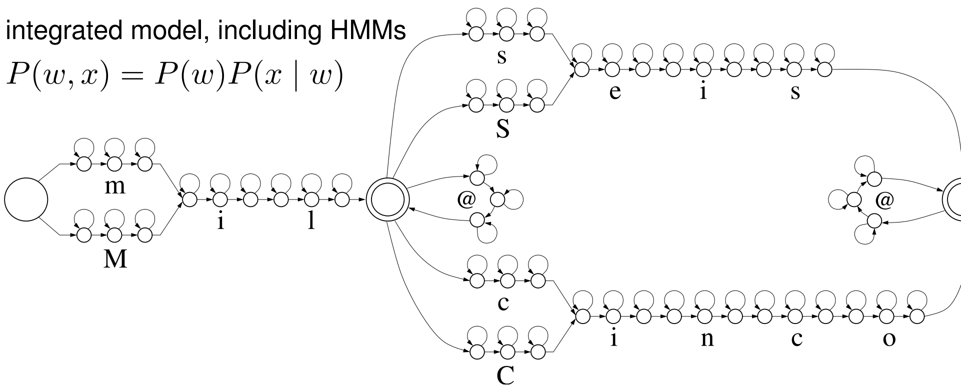


language model $P(w)$



integrated model, including HMMs

$P(w, x) = P(w)P(x \mid w)$



Training HTR Models

- *Language Model*, $P(w)$: N-gram training from the *tokenized* transcripts of the text images (and possibly other relevant "external" texts)

- *Lexicon*: set of words in the tokenized training text (possibly extended with relevant "external" vocabularies), spelled in terms of characters, including one (or more) white-space "character(s)"

- *Optical character HMMs*: *"embedded Baum-Welch training"* from pairs of text line images and their corresponding transcripts. No segmentation of the training images into words or characters is needed

HTR Experiments with Real Historical Manuscript Collections

- PLANTAS: XVII century botanical specimen manuscript collection of seven volumes written by a single hand in Old Spanish – kindly provided by the BNE

- ESPOSALLES: XVII century Marriage License records written by several hands in old Catalan and other languages

- HATTEM: XV century Medieval Manuscript composed of 573 sheets written by a single hand in Dutch

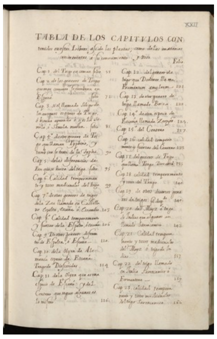
- REICHSGERICHT: XVIII century court decisions from the High Court of Germany, written by several hands in German

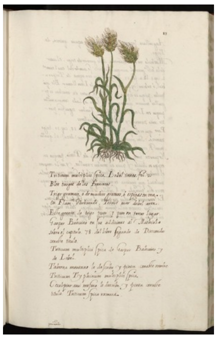
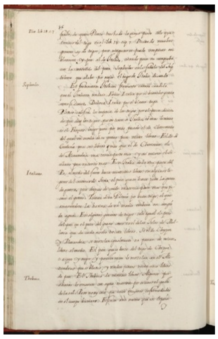
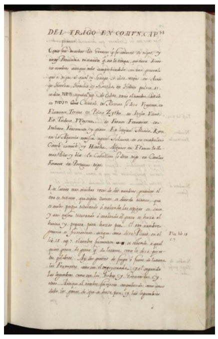
- BENTHAM: XVIII/XIX centuries collection of over $4,000$ volumes of drafts and notes, written by several hands in English

- AUSTEN: XVIII century Juvenilia manuscripts by Jane Austen (single hand in English) – kindly provided by the BL

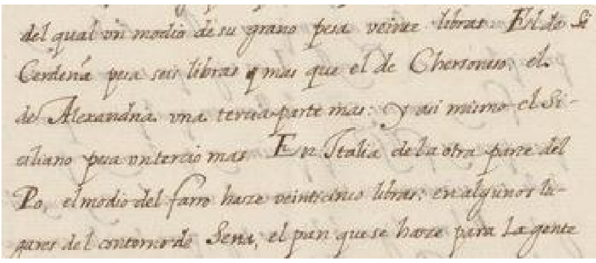
“PLANTAS” Dataset

XVII century Botanical Specimen Manuscript Collection of seven volumes written by a single writer in Old Spanish



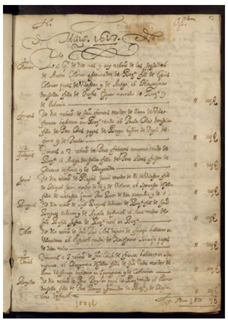
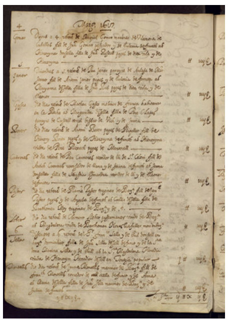
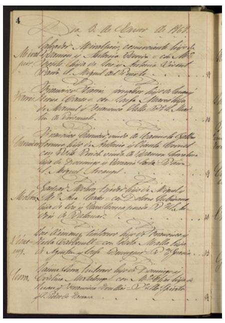
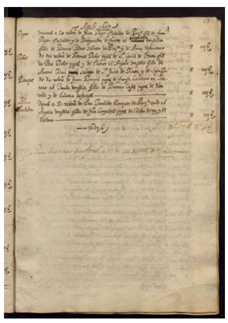
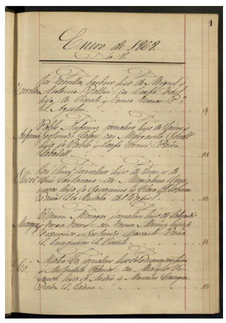
Experiments on the first volume



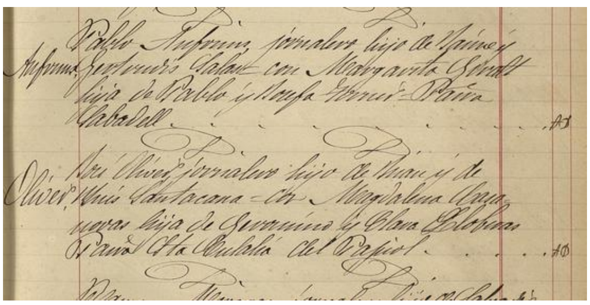
Number of:	Total
Pages	871
Lines	19 544
Running words	197 617
Lexicon size	21 148
Character set size	77

“ESPOSALLES” Dataset

XVII century Marriage License collection of several volumes

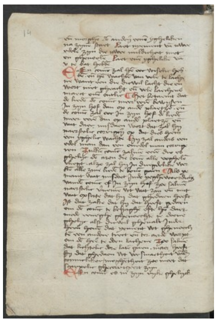
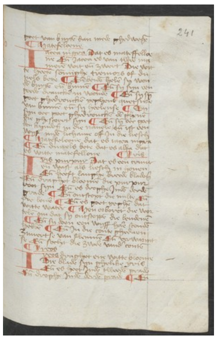
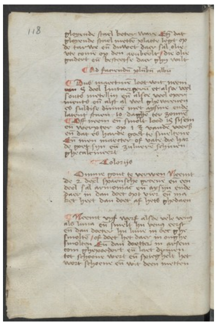
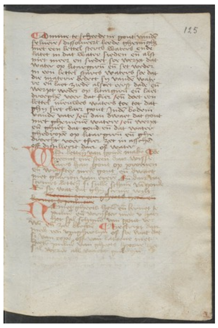
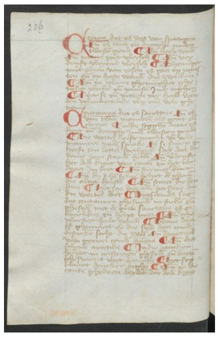
Experiments on the volume 69 written by a single writer in old Catalan



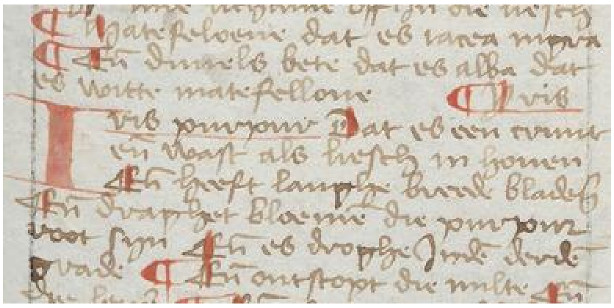
Number of:	Total
Pages	173
Lines	5 447
Running words	60 777
Lexicon size	3 465
Running characters	328 229
Character set size	85

“HATTEM” Dataset

XV century Manuscript composed of 573 sheets written by a single writer in Dutch

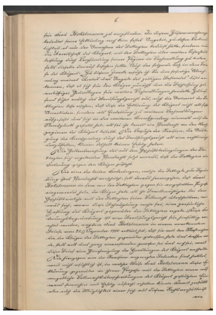
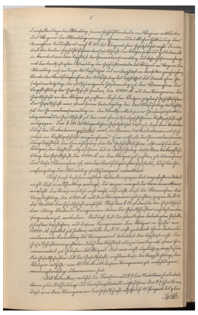
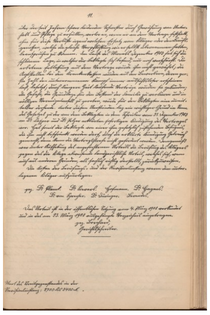
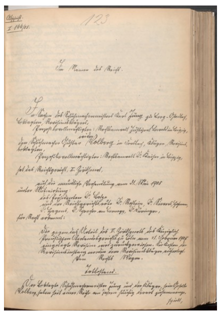
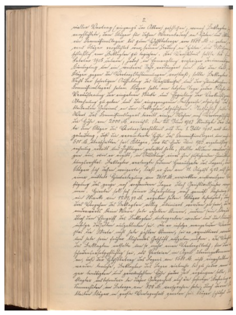
Experiments on 40 randomly selected pages

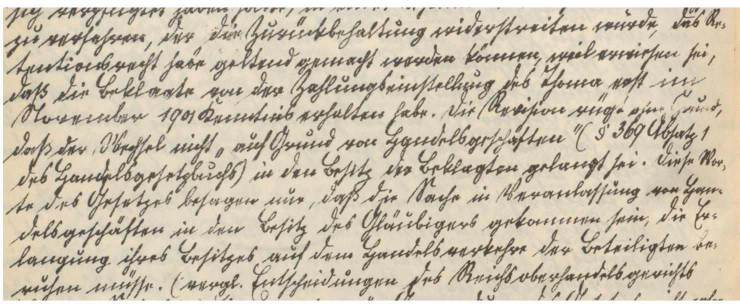


Number of:	Total
Pages	40
Lines	1 552
Running words	10 330
Lexicon size	2 602
Running characters	42 712
Character set size	60

“REICHSGERICHT” Dataset

XVIII century court decisions from the High Court of Germany, written by several hands

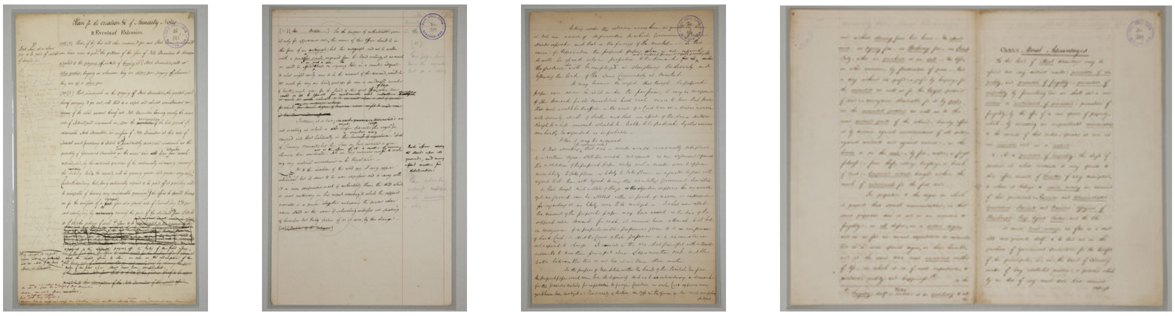








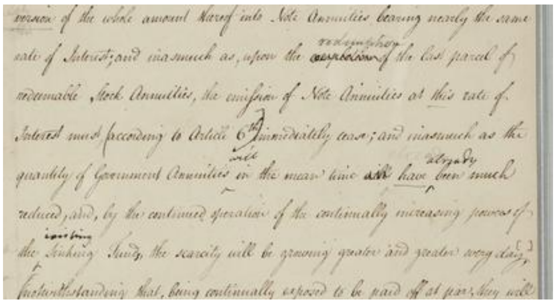
Number of:	Total
Pages	114
Lines	4 106
Running words	31 545
Dataset lexicon	8 108
Running characters	239 762
Character set size	92

"BENTHAM" Dataset

XVIII century collection of over $4,000$ volumes of drafts and notes, written by several writers in English



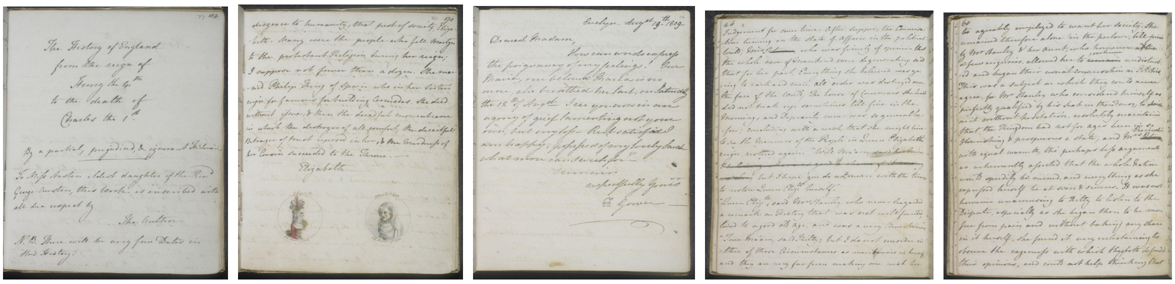
Experiments on a first batch of 433 pre-selected page images


version of the whole amount thereof into Note Annuities bearing nearly the same rate of Interest; and inasmuch as, upon the redemption of the last parcel of redeemable Stock Annuities, the emission of Note Annuities at this rate of Interest must according to Article 6 immediately cease; and inasmuch as the quantity of Government Annuities in the mean time will already have been much reduced, and, by the continued operation of the continually increasing powers of the Sinking Fund, the scarcity will be growing greater and greater every day; notwithstanding that, being continually exposed to be paid off at par, they will

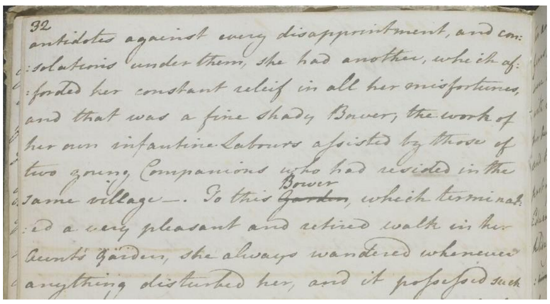
Number of:	Total
Pages	433
Lines	11 473
Running words	106 905
Lexicon size	9 717
Running characters	550 674
Character set size	86

"AUSTEN" Dataset

Jane Austen's *Juvenilia*: XVIII century single hand manuscript



Experiments on Volume The Third


32
antidotes against every disappointment, and consolations under them, she had another, which afforded her constant relief in all her misfortunes, and that was a fine shady Bower, the work of her own infantine Labours assisted by those of two young Companions who had resided in the same village. To this ~~Garden~~ Bower, which terminated in a very pleasant and retired walk in her Aunts Garden, she always wandered whenever anything disturbed her, and it possessed such

Number of:	Total
Pages	128
Lines	2 693
Running words	25 291
Dataset lexicon	3 567
Running characters	118 881
Character set size	81

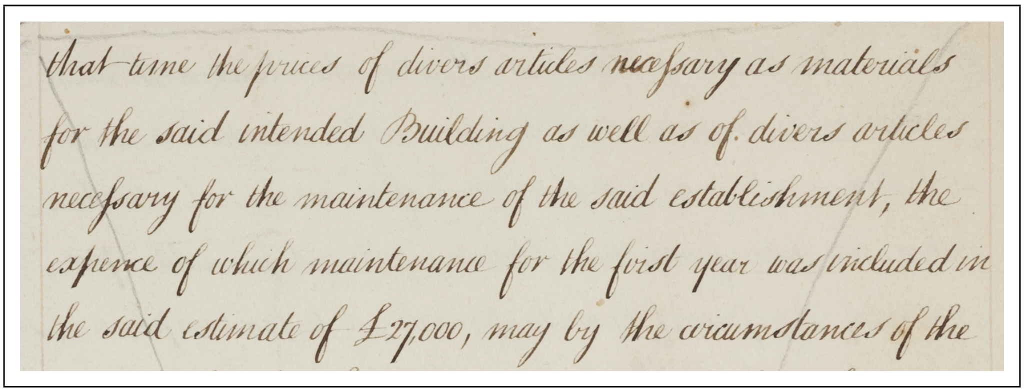
HTR Empirical Results (as of 2014)

- PLANTAS: Training: OMs with 224 pages, LM Lex. 21K words. Test: 647 pages.
  WER = 33.4 % CER = 16.0 % Running OOV rate: 12%
- ESPOSALLES: Cross validation on 173 pages, LM Lexicon ≈ 3.3K words.
  WER = 16.1% CER = 9.9% ROOV: 5%
- HATTEM: Cross-validation on 40 pages, LM Lexicon ≈ 2.5K words.
  WER = 33.8 % CER = 15.2% ROOV: 20%
- REICHSGERICHT: Training: OMs with 88 pgs, LM Lexicon 5K words. Test: 26 pgs.
  WER = 33.3 % CER = 12.9 % ROOV: 10%
- BENTHAM: Training: OMs with 400 pages, LM Lex. 10K words. Test: 33 pages.
  WER = 24.6 % CER = 13.8 % ROOV: 5.3%
- AUSTEN: Training: OMs with 50 pages, LM Lexicon 20K words. Test: 78 pages.
  WER = 35.3 % CER = 17.1 % ROOV: 3.6%
  - AUSTEN: *No training*; just using BENTHAM models WER = 45.0 % CER = 25.5 %
  - AUSTEN: *Training with both* AUSTEN *and* BENTHAM WER = 24.2 % CER = 11.7 %

LMs: *Open-vocabulary bi-grams*. WER/CER: percentage of mis-recognized words/characters

Is Current HTR Accuracy Useful? (example 1 from Bentham)

HTR models trained with 350 pages. Lexicon: 8 660 words. WER≈7%, CER≈1%

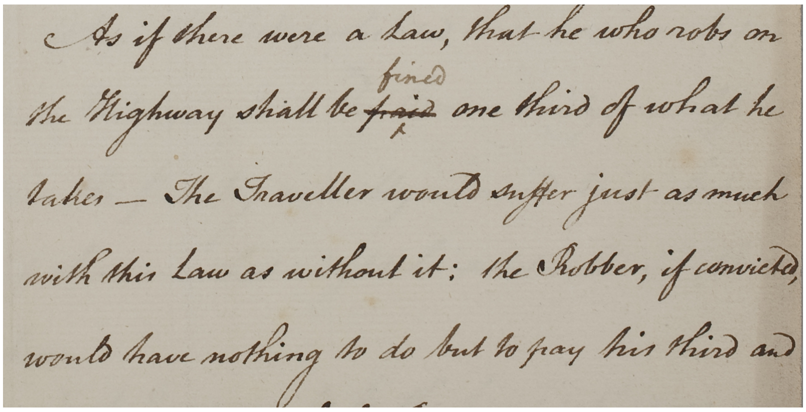

that time the prices of divers articles necessary as materials
for the said intended Building as well as of divers articles
necessary for the maintenance of the said establishment, the
expence of which maintenance for the first year was included in
the said estimate of £27,000, may by the circumstances of the

That time the prices of divers articles necessary as materials
for she said intended Building as well as of divers articles
necessary for the maintenance of the said establishment • the
expence Of which maintenance for the first year was included in
the said estimate of £ 27,000 • may by the circumstances of the

Tokens with two or less character wrong in blue/red; with more than two, all in red

Is Current HTR Accuracy Useful? (example 2 from Bentham)

HTR models trained with 350 pages. Lexicon: 8 660 words. WER≈24%, CER≈8%

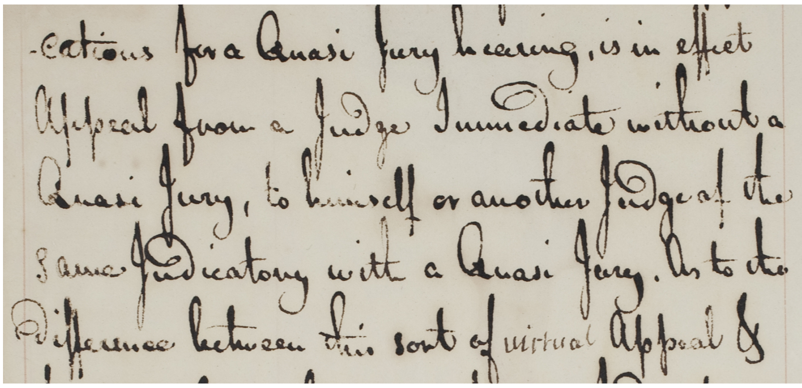

As if there were a Law, that he who robs on
the Highway shall be ~~paid~~ fined one third of what he
takes — The Traveller would suffer just as much
with this Law as without it: the Robber, if convicted,
would have nothing to do but to pay his third and

As if there were a law • that he who room or
the Highway shall be pass one third of what he
taken _ The Traveller would suffer just as much
wish this Law as without it • the Robber • of convicts• •
would have nothing to do but so pay his third and

Tokens with two or less character wrong in blue/red; with more than two, all in red

Is Current HTR Accuracy Useful? (example 3 from Bentham)

HTR models trained with 350 pages. Lexicon: 8 660 words. WER≈52%, CER≈30%


-cations for a Quasi Jury hearing, is in effect
Appeal from a Judge Immediate without a
Quasi Jury, to himself or another Judge of the
same Judicatory with a Quasi Jury. As to the
Difference between this sort of virtual Appeal &

-tations for a Clause Jury because • as in effect
Appeal from a Judge Su ideal without a
Cause Jury • to himself on another Towns Of •••
To see threatens with a Thing especially • •• to the
difference between This sort of with a Appeal •

Tokens with two or less character wrong in blue/red; with more than two, all in red

Is Current HTR Accuracy Useful?

- Accuracy of fully automatic HTR could be enough for some (or many?) applications involving not too difficult documents
- Even if transcriptions are not perfect, they could be used to derive adequate *metadata* that would roughly describe document contents
- Very accurate *word spotting* can be easily implemented using similar segmentation–free, N-gram/HMM technology as in HTR.

However. . .

⋆ Current automatic HTR accuracy is *not enough for high quality* transcription of most (historical) handwritten text images of interest

– *Human post-editing can be very expensive* and hardly acceptable by profesional transcribers (paleographers)

+ *Computer Assisted, Interactive-Predictive processing* offers promise for *significant improvements in practical performance and user acceptance*.

Computer–Assisted Transalation of Text Images (CATTI): example

	x							
STEP-1	$\hat{s} \equiv \hat{w}$	antiguas	cuidadelas	que	en	el Castillo	sus	llamadas
	p'	antigu						
	κ	os						
	p	antiguos						
STEP-2	$\hat{s}$		ciudadanos	que	en	el Castillo	sus	llamadas
	p'	antiguos	ciudadanos	que	en			
	κ					Castilla		
	p	antiguos	ciudadanos	que	en	Castilla		
FINAL	$\hat{s}$						se	llamaban
	p'	antiguos	ciudadanos	que	en	Castilla	se	llamaban
	κ							
	$p \equiv T$	antiguos	ciudadanos	que	en	Castilla	se	llamaban

Post–editing Word Error Rate WER: 6/7 (86%)
CATTI Word Stroke Ratio (WSR): 2/7 (29%), assuming whole-word corrections
Estimated Effort Reduction (EFR): $1 - 29/86$ (66%).

Statistical Framework for CATTI

CATTI is an instance of Interactive Pattern Recognition (IPR):

- the *input*, x, is a feature vector stream representing a line image,
- the *human feedback* is a *transcription prefix*, here called p,
- a system hypothesis is a suitable continuation of p, here called *transcription sufix*, s.

$$\hat{s} = \arg\max_s P(s \mid x, p) = \arg\max_s P(x \mid p, s) \cdot P(s \mid p)$$

Modeling:

- $P(x \mid p, s)$: *optical HMMs*
- $P(s \mid p)$: *prefix-conditioned N-Gram Language Model*

Search:

- Direct, by repeated Viterbi decoding
  $\Rightarrow$ Accurate, but prohibitively slow
- Using *Word-Graphs obtained from one Viterbi decoding*
  $\Rightarrow$ Fast, at the expense of some accuracy loss

CATTI Empirical Results (as of 2014)

- HATTEM: Cross-validation on 40 pages
  WER = 33.8 % CER = 15.2 % WSR = 26.8 % EFR: 20.7 %
- REICHSGERICHT: Training: OMs with 88 pgs, LM Lexicon 5K words. Test: 26 pgs
  WER = 33.3 % CER = 12.9 % WSR = 25.1 % EFR: 24.6 %
- BENTHAM: Training: OMs with 400 pages, LM Lex. 10K words. Test: 33 pages.
  WER = 24.6 % CER = 13.8 % WSR = 17.2 % EFR: 28.0 %
- AUSTEN: Training: OMs with 50 pages, LM Lexicon 20K words. Test: 78 pages
  WER = 35.4 % CER = 17.1 % WSR = 22.0 % EFR: 37.7 %

WER/CER: percentage of mis-recognized words/characters.
WSR: Percentage of word-level corrections to achieve ground truth transcripts.
EFR: *"Estimated Effort Reduction"*.
Experiments with *open-vocabulary* lexica and bi-gram LMs.

- Estimated effort is reduced by 70–80% (100-WSR) *wrt pure manual* transcription
- In contrast with *post-edditing*, CATTI is much more user friendly, since it allows the user to be always in command of the transcription process
- CATTI significantly reduces the estimated effort *wrt* post–editing (approx 20-40% EFR)

HTR and CATTI Demonstrations

- It is just a *“demo”* ! not intended for real operation (other systems do that)
- Everithing is *real*. No tricks to make demo look better than real
- Web client-server architecture:
  Web browser front-end, CATTI back-end server
- Off-line CATTI decoder based on wordgraphs
- Several tasks of increassing complexity

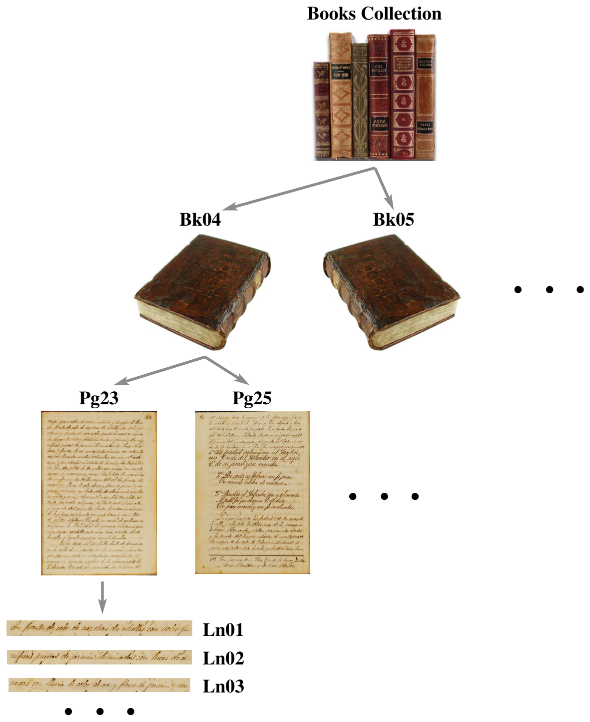
Keyword Indexing and Search in Untranscribed Text Images

- Many massive handwritten text document collections are available in archives and libraries, but their textual contents remain practically inaccessible, “buried” behind thousands of terabytes of high-resolution images
- If perfect or sufficiently accurate text image transcripts were available, image textual content could be strightforwardly indexed for plaintext textual access
- But fully automatic transcription results lack the level of accuracy needed for reliable text indexing and search purposes
- And manual or even computer-assited transcription is entierely prohibitive to deal with massive image collections

Good news: indexing and search can be directly implemented on the images themselves, *without explicitly resorting to image transcripts*.

Indexing and Search: A Hierarchical Model

- Indexing large document collections call for a *hierarchical organization* of indices
- The lowest herarchy level should consist of sufficiently small and practically meaningful *image regions*, such as *lines*
- The *precission-recall trade-off search model* requires *word confidence measures* to be properly defined at each level of the hierarchy
- Confidence measures must be *properly normalized* and *homogeneous* across hierarchy levels
- A *statistical KWS framework* is introduced to support the computation of the required confidence measures



Text Image KWS Statistical Framework: 2-D Posteriorgram

Main concept: *Posterior word probability at pixel level, or "2-D Posteriorgram"*:

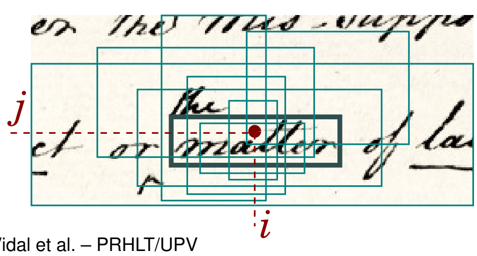
$$P(v \mid X, i, j), \quad 1 \leq i \leq I, \ 1 \leq j \leq J, \quad v \in V$$

where X is a $I \times J$ sized *text image*, V is a *vocabulary* and (i, j) a *pixel* of X.

$P(v \mid X, i, j)$ denotes the probability that a word v is written in a subimage of X which includes the pixel (i, j). It can be directly computed by *margizalization*:

$$P(v \mid X, i, j) = \sum_B P(v, B \mid X, i, j) \approx \frac{1}{K(i,j)} \sum_{B \in \mathcal{B}(i,j)} P(v \mid X, B)$$

where $\mathcal{B}(i,j)$ is the set of all the $K(i,j)$ reasonably shaped and sized (and assumedly equiprobable) boxes or subimages of X which include the pixel (i, j).



A few possible boxes $B \in \mathcal{B}(i,j)$. For $v =$ "**matter**", the thick-line box will provide the highest value of $P(v \mid X, B)$, while most of the other boxes will contribute only (very) low values to the sum.

What is exactly $P(v \mid X, B)$?

Computing the 2-D Posteriorgram by word classification

The 2-D Posteriorigram:

$$P(v \mid X, i, j) \;\approx\; \frac{1}{K(i,j)} \sum_{B \in \mathcal{B}(i,j)} P(v \mid X, B)$$

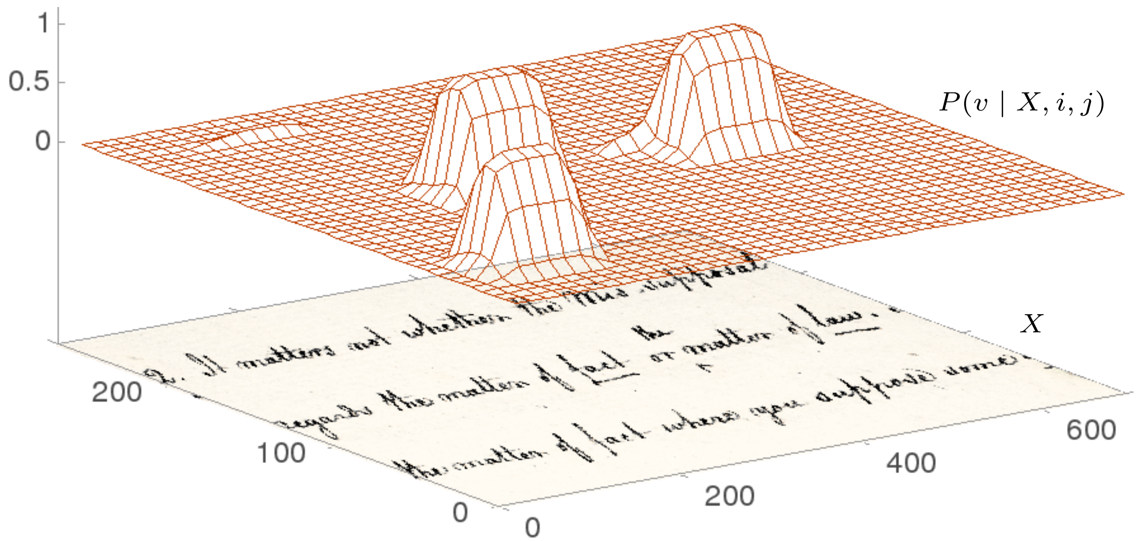
$P(v \mid X, B)$ is the posterior probability (implicitly or explicitly) used by any *isolated word image classifier*; i.e, any system capable of solving the following classification problem for a *presegmented* word subimage of X bounded by B:

$$\hat{v} \;=\; \arg\max_{v \in V} \; P(v \mid X, B)$$

Clearly, the better the classifier the better the estimated posterirorgram.

Notice: *Directly obtaining a full 2-D posteriorgram in this way entails a formidable amount of computation, but $P(v \mid X, i, j)$ can be very efficient computed by clever combinations of subsampling of (i, j) and choices of $\mathcal{B}(i, j)$* [see later].

Pixel-level Posteriorgram (illustration)



2-D Posteriorgram, $P(v \mid X, i, j)$, for a text image X and word $v =$"**matter**".

An accurate, contextual (n-gram based) *word classifier* was used to compute $P(v \mid X, B) \; \forall B \in \mathcal{B}(i, j)$. This resulted in very low posteriors in a region of X around $(i = 100, j = 200)$, where a very similar word, "**matters**", is written.

Image Region KWS

- Posteriorgrams can be directly used for KWS: Given a threshold $\tau \in [0, 1]$, a word $v \in V$ is spotted in all image positions where $P(v \mid X, i, j) > \tau$. Varying τ, adequate *precision–recall* tradeoffs can be achieved

- But, for indexing purposes, we need the probability that a word v is written within a pre-specified image region, such as a page, a column, or a line

  A popular (but wrong!) idea: For a text image region X, use the word posterior probability $P(v \mid X)$

  This is *ill-defined*, because $\sum_{v \in V} P(v \mid X) = 1$

  . . . but, for each of the (many) different words v actually written in X, we ideally want $P(v \mid X)$ to be close to 1 : the sum should ideally be $>>> 1$!

What is an adequate posterior probability for image region KWS ?

Image Region KWS: Proper Probabilistic Formulation

Let X be a given *image region* and $R \in \{\text{yes}, \text{not}\}$ a *binary* random variable.

We define the "R–posterior", $P(R \mid X, v)$, which denotes the probability that X is *relevant* for v; i.e., v is written somewhere in X.

It is computed as [this is the short history – see formal details here ▷34]:

$$P(R \mid X, v) \;\approx\; \max_{i,j} P(v \mid X, i, j)$$

. . . a formal result which is also intuitively meaningful (see page ▷33)!

Now:

$$\sum_{v \in V} P(R \mid X, v) = m$$

where m (generally much greater than 1) is the expected number of words from V written in the image region X.

Choosing Adequate Minimal Image Regions: Line-level KWS

Lines are useful image regions for indexing and search in practice; and they allow for *efficient computation* by *clever vertical subsampling* and *choosing* $\mathcal{B}(i,j)$:

- *Vertical subsampling*: In general, it amounts to just guessing a proper line height and then runing a vertical-sliding window of this height with some overlap
- Choosing $\mathcal{B}(i,j)$: For a line-level region, blocks needed to compute the posteriorfram by marginalization can be just defined by *horizontal segmentation*

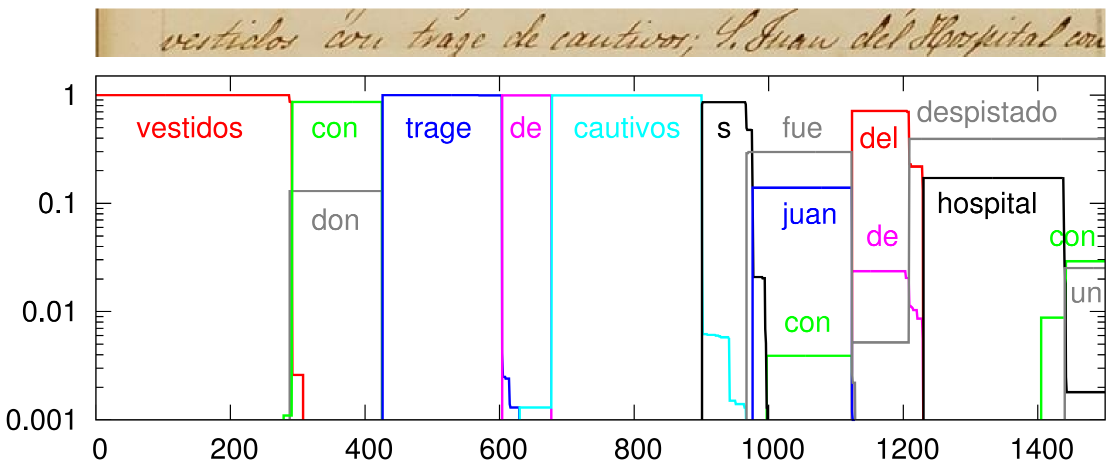
Line-level posteriorgrams are very efficiently computed using *Word Graphs*, obtained as a byproduct of Viterbi or "token-passing" decoding of line images.

This has two important benefits in order to compute posteriorgrams by marginalization:

- *Optical* (HMM) *Character Models* and (N-gram) *Language Models* are used to provide very accurate, contextual word classification probabilities, $P(v \mid X, B)$
- WGs provide lots of alternative horizontal word-level segmentations, which directly define $\mathcal{B}(i,j)$

Line region R–posteriors are directly computed from the corresponding posteriorgrams. They can in turn be easy and consistently combined to obtain *page-level* R–posteriors (...and so on for *chapters*, *books*, etc., for *hierarchical indexing*)

Line-level Posteriorgram (illustration)

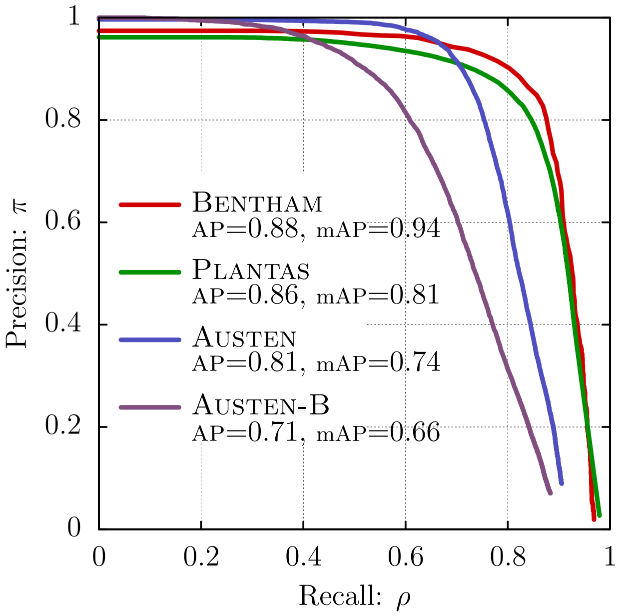


For a given line-level image region, x (on top), the posterior probabilities $P(v \mid x, i)$ of a few words (v) are shown as a function of the horizontal image position (i).

These posteriors are computed by marginalization over large amounts of horizontal word segmentation hypotheses provided by a Word Graph obtained from x.

HTIS Laboratory Results on Several Collections

– *Recall-Precision* curves
– *Average Precision* (AP)
– *Mean Average Precision* (mAP)



Datasets training and test details

- **BENTHAM:** *Multi-hand*. *Training*: 400 pages from Bentham, 87 char. HMMs, 2-gram LM trained on Bentham texts; Lexicon 9 341 tokens. *Test*: 33 pages; query set: 6 962 keywords
- **PLANTAS (VOL-I):** *Single hand*. *Training*: 224 pages from *Plantas*, 77 char. HMMs, 2-gram LM trained with the training set + book glossary transcripts. Lexicon 11 561 tokens. *Test*: 647 pages; query set: 9 945 keywords
- **AUSTEN:** *Single hand*. *Training*: 50 Austen pages, 81 char. HMMs, 2-gram LM trained on Austen texts; Lexicon 20K tokens. *Test*: 78 pages; query set: 2 281 keywords
- **AUSTEN-B:** *Single hand*. ***No training***; using Bentham character HMMs, lexicon and LM. *Test* & query set: Same as for **AUSTEN**

Handwritten Text Images Indexing and Search: Demonstration

- It is just a *"demo"*! not (yet) intended for real operation. But everything is *real* – no tricks to make demo look better than real
- Line-level indexing according to the *precision-recall trade-off model*:

  Rather than exact searching, search is carried out with a *confidence threshold*, specified by the user as part of the query in order to meet the required *precision-recall trade-off*
- Word confidence scores are based on pixel-level probabilities and computed for *line-shaped regions*. Spotted word positions are marked only approximately
- Several collections: AUSTEN, PLANTAS, WIENSANKTULRICH, . . . etc.

Conclusions

HTR Holistic optical and language modeling statistical technology:

- Accuracy of fully automated HTR can be enough for some applications and, in general, as a tool for building *metadata* for rough contents description

Interactive-Predictive HTR technology:

- Current fully automatic HTR accuracy is not enough for high quality transcription of most handwritten historical text images of interest
- Human post-editing can be very expensive and hardly acceptable by profesional transcribers (paleographers)
- *Computer Assisted, Interactive-Predictive* HTR offers promise for *significant improvements in practical performance and user acceptance*

Keyword Search technology:

- Accurate *keyword indexing and search* in untranscribed images based on confidence scores obtained using holistic HTR models and techniques

Bibliography

- F. Jelinek. "Statistical Methods for Speech Recognition". MIT Press, 1998.
- I. Bazzi, R. Schwartz, J. Makhoul. "An Omnifont Open-Vocabulary OCR System for English and Arabic". IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) Vol.21 pp.495-504, 1999.
- A. Vinciarelli, S. Bengio, H. Bunke. "Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models". IEEE Trans. on PAMI, Vol.26, pp.709-720, 2004.
- A. H. Toselli, A. Juan, D. Keysers, J. Gonzlez, I. Salvador, H. Ney, E. Vidal and F. Casacuberta. "Integrated Handwriting Recognition and Interpretation using Finite-State Models". Int. Journal of Pattern Recognition and Artificial Intelligence, 18(4):519-539, June 2004.
- E. Vidal, L. Rodriguez, F. Casacuberta and I. García-Varea: "Interactive Pattern Recognition". 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-07), Volume 4892 of LNCS, pp.60-71. Brno, Czech Republic, June 2007.
- A.H. Toselli, V. Romero, L. Rodríguez and E. Vidal. "Computer Assisted Transcription of Handwritten Text". 9th Int. Conference on Document Analysis and Recognition (ICDAR 2007), pp.944-948. IEEE Computer Society, Curitiba, Paraná (Brazil), September 2007.
- A.H. Toselli, V. Romero, M. Pastor and E. Vidal. "Multimodal interactive transcription of text images". Pattern Recognition, Vol.43, N.5, pp.1814–1825, April 2010.
- A.H.Toselli, E.Vidal, F.Casacuberta: "Multimodal Interactive Pattern Recognition and Applications". Springer Verlag, 2011.
- V.Romero, A.H.Toselli and Vidal: "Multimodal Interactive Handwritten Text Transcription", World Scientific, 2012.
- D.Martín-Albo, V.Romero, E.Vidal. "Escritoire: a Multi-Touch Desk with e-Pen Intput for Capture, Management and Multimodal Interactive Transcription of Handwritten Dcouments". 7th Iberian Conference on Pattern Recognition and Image Analysis, proc. pp.471-478. Springer, 2015