ICFHR–2010 Tutorial:

Multimodal Computer Assisted Transcription of Handwriting Images

III – Multimodality in Computer Assisted Transcription

Alejandro H. Toselli & Moisés Pastor & Verónica Romero

{ahector,moises,vromero}@iti.upv.es

Pattern Recognition and Human Language Technology Group

Instituto Tecnológico de Informática – Universidad Politécnica de Valencia

Spain

November 2010

A BLANK PAGE

Tutorial Contents and Schedule

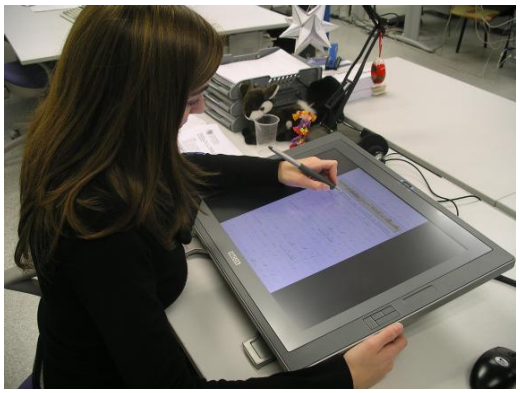
Index

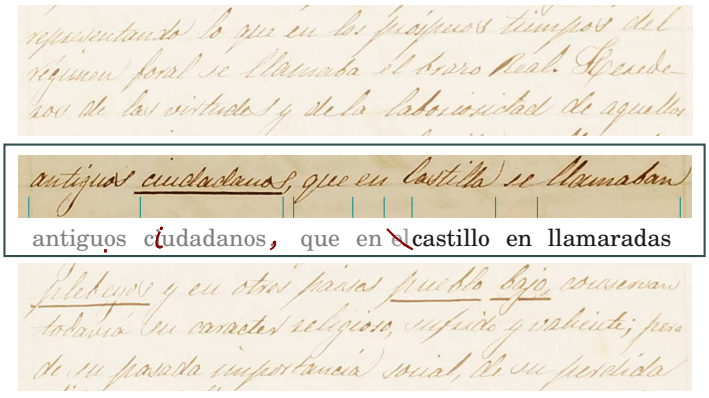
Multimodal Interaction for Text Transcription

- Human feedback for CATTI has been assumed to come in the form of keyboard and mouse actions
- At the expense of loosing the deterministic accuracy of this traditional input modality, more ergonomic multimodal interfaces are possible: e.g., voice, gaze tracking, etc.
- A very natural modality for Text Transcription is on-line HTR on a touchscreen: *Multimodal CATTI (MM-CATTI)*
- The increased ergonomy comes at the cost of new errors expected from the decoding of the feedback signals

Solving the *multimodal interaction problem* requires to achieve a *modality synergy* where both main and feedback data streams help each-other to optimize overall accuracy.

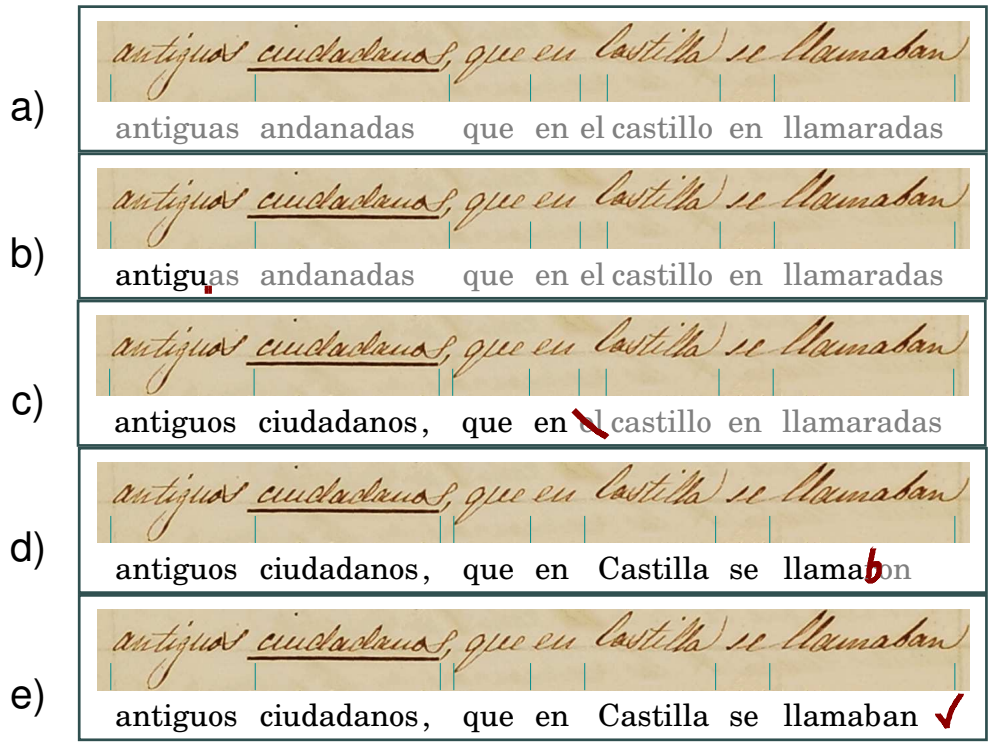
A multimodal desktop for interactive Text Transcription





On-line HTR on a touchscreen offers a very natural modality for Interactive Text Image Transcription: The original image, the on-going transcription and the user corrective feedback penstrokes are jointly displayed for the user comfort.

Multimodal Interaction for Text Transcription (illustration)



General statistical framework for MM-CATTI

Let:

- x be the input image
- p be a user-validated transcription prefix
- t be the user feedback (on-line *touchscreen pen strokes*)
- s' be the previous system-suggested suffix; The feedback t is aimed at accepting or amending parts of s' and/or at adding more text
- κ be some *keystrokes* typed by the user to correct (other) parts of s' and/or to add more text

Now the system has to suggest a new suffix, $\hat{s}$, as a continuation of the prefix p, conditioned by the on-line pen strokes t and the typed text κ.

The general problem is to find $\hat{s}$ given x, p and κ and considering all possible *decodings*, d, of t (i.e., letting d be a hidden variable):

$$\hat{s} = \underset{s}{argmax} \sum_{d} \Pr(s, d \mid x, p, s', t, \kappa)$$

A more realistic MM-CATTI framework

In the general formulation, the user can type with independence of the result of the on-line handwritten decoding process. This is not realistically useful in practice.

More realistic scenario: Wait for the decoding $\hat{d}$ of the touchscreen data t, prior to typing κ. These keystrokes will typically be used to fix possible on-line handwritten recognition errors in $\hat{d}$. Each interaction step can be formulated in three phases:

1. The system relies on the input image and the previous transcription prefix, in order to search for a good transcription suffix, as in CATTI:

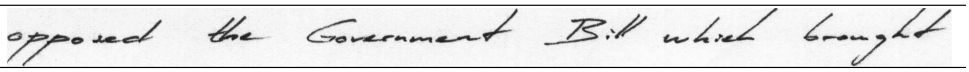
$$\hat{s} = \underset{s}{argmax}\, \Pr(s \mid x, p)$$

2. Given $\hat{s}$, the user produces (may be null) on-line pen strokes t and the system decodes t into an optimal character, word (or word sequence), $\hat{d}$:

$$\hat{d} = \underset{d}{argmax}\, \Pr(d \mid x, p, \hat{s}, t)$$

3. The user enters amendment keystrokes κ, if necessary, and consolidates a new prefix, p, based on the previous p, $\hat{d}$, κ and parts of $\hat{s}$.

MM-CATTI operation

	x						
STEP-0	p						
STEP-1	$\hat{s} \equiv \hat{w}$	**opposite**	**this**	**Comment**	**Bill**	**in that**	**thought**
	p', t	~~opposed~~	this	Comment	Bill	in that	thought
	$\hat{d}$	opposed					
	κ						
	p	**opposed**					
STEP-2	$\hat{s}(\equiv s')$		**the**	**Government**	**Bill**	**in that**	**thought**
	p', t	**opposed**	**the**	**Government**	**Bill**	which	thought
	$\hat{d}$					whack	
	κ					**ich**	
	p	**opposed**	**the**	**Government**	**Bill**	**which**	
FINAL	$\hat{s}(\equiv s')$						**brought**
	p', t	**opposed**	**the**	**Government**	**Bill**	**which**	**brought**
	κ						
	$p \equiv T$	opposed	**the**	**Government**	**Bill**	which	**brought**

POST-EDIT: 6 corrections
CATTI: 2 keyboard corrections
MM-CATTI: 3 corrections; 2 touch-screen (red) + 1 keyboard (underlined)

Decoding the feedback data for MM-CATTI

Assuming independence between t and $x, p, \hat{s}$ given d, the feedback penstroke decoding can be rewritten as:

$$\begin{aligned}\hat{d} &= \underset{d}{argmax}\, \Pr(d \mid x, p, \hat{s}, t) = \underset{d}{argmax}\, \Pr(d, x, p, \hat{s}, t) \\ &= \underset{d}{argmax}\, \Pr(x, p, \hat{s}) \cdot \Pr(d \mid x, p, \hat{s}) \cdot \Pr(t \mid d, x, p, \hat{s}) \\ &= \underset{d}{argmax}\, \Pr(d \mid x, p, \hat{s}) \cdot \Pr(t \mid d)\end{aligned}$$

- $\Pr(t \mid d)$ modelled by HMM models of the word(s) in d
- $\Pr(d \mid x, p, \hat{s})$ provided by a language model constrained by the input image x, by the previous prefix p and by the suffix $\hat{s}$

In practice, a *Grammar Scale Factor* is generally used:

$$\hat{d} = \underset{d}{argmax}\, \Pr(d \mid x, p, \hat{s})^{\alpha} \cdot \Pr(t \mid d)^{(1-\alpha)}$$

Several assumptions and constraints can be adopted for $\Pr(d \mid x, p, \hat{s})$

MM-CATTI feedback decoding LM constraints

Different level-constrained cases can be considered:

- Most complex case – use all the conditions in $\Pr(d \mid x, p, \hat{s})$, including the input image x (not considered for now).
- Simplest, *baseline* case – pure on-line HTR, ignoring all the available conditions: $\Pr(d \mid x, p, \hat{s}) \approx P(d)$ (standard N-gram).
- *Error-conditioned* case – Consider the wrong words that user tries to correct $\Pr(d \mid x, p, \hat{s}) \approx P(d \mid s_e)$.
- *Compromise*: Only consider the information provided by the prefix p and the off-line HTR prediction $\hat{s}$: $\Pr(d \mid x, p, \hat{s}) \approx P(d \mid p, \hat{s})$.

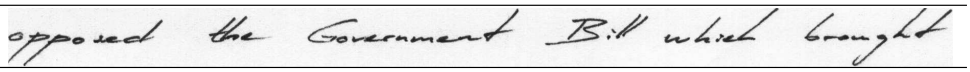
MM-CATTI feedback decoding LM constraints

Compromise: Only consider the information provided by the prefix p and the off-line HTR prediction $\hat{s}$: $\Pr(d \mid x, p, \hat{s}) \approx P(d \mid p, \hat{s})$.

This simplifies the feedback pen-strokes decoding, while still achieving an effective synergy between the main and the feedback modalities:

- the user accepts an initial correct part of $\hat{s}$, $\hat{s}_a$
- this validates a new overall correct prefix $p' = p\ \hat{s}_a$ and signals the first erroneous word(s) of $\hat{s}$, $\hat{s}_e$
- the user produces some (may be null) pen strokes t aimed to amend $\hat{s}_e$
- the multimodal (on-line HTR) decoder is constrained to find a transcription of t which is a suitable continuation of p' and depends on the wrong word(s) $\hat{s}_e$: $P(d \mid p, \hat{s}) \equiv P(d \mid p', \hat{s}_e)$

MM-CATTI operation

	x						
STEP-0	p						
STEP-1	$\hat{s} \equiv \hat{w}$	**opposite**	**this**	**Comment**	**Bill**	**in that**	**thought**
	p', t	opposed	this	Comment	Bill	in that	thought
	$\hat{d}$	opposed					
	κ						
	p	**opposed**					
STEP-2	$\hat{s}(\equiv s')$		**the**	**Government**	**Bill**	**in that**	**thought**
	p', t	**opposed**	**the**	**Government**	**Bill**	which	thought
	$\hat{d}$					whack	
	κ					ich	
	p	**opposed**	**the**	**Government**	**Bill**	**which**	
FINAL	$\hat{s}(\equiv s')$						**brought**
	p', t	**opposed**	**the**	**Government**	**Bill**	**which**	**brought**
	κ						
	$p \equiv T$	opposed	**the**	**Government**	**Bill**	which	**brought**

POST-EDIT: 6 corrections
CATTI: 2 keyboard corrections
MM-CATTI: 3 corrections; 2 touch-screen (red) + 1 keyboard (underlined)

MM-CATTI Language Modeling and Search

To simplify notation, p and e are used instead of p' and $\hat{s}_e$; where e can be a sequence of one or more words or characters (just 1 in practice).

Assuming e is a single word, the decoder should produce a *word hypothesis* $\hat{d}$ for the pen strokes t.

Language model constrains can be approached using n-grams, depending on each multimodal scenario considered.

- *Baseline* $P(d)$: does not take into account any interaction-derived information and since only whole-word touchscreen corretions are assumed, only uni-grams actually make sense.

- *Error-conditioned* model $P(d \mid e)$:

$$P(d \mid e) = \begin{cases} 0 & d = e \\ \dfrac{P(d)}{1 - P(e)} & d \neq e \end{cases}$$

MM-CATTI Language Modeling and Search

- Compromise: The prefix-and-error conditioned MM-CATTI language model can be approached as:

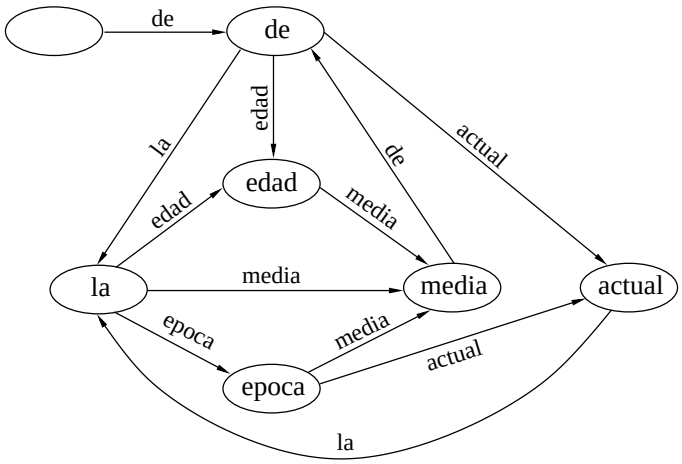
$$P(d \mid p, e) \approx \begin{cases} 0 & d = e \\ \dfrac{P(d \mid p^k_{k-n+2})}{1 - P(e \mid p^k_{k-n+2})} & d \neq e \end{cases}$$

  where k is the length of p.

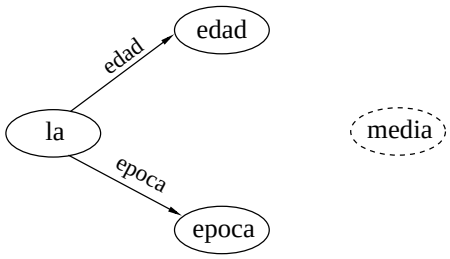
- The information provided by the interaction process should allow for feedback decoding improved accuracy.

Example of building a MM-CATTI dynamic 2-gram LM

Original Bi-gram model $P(w)$



Conditioned Bi-gram model $P(d \mid p, e)$



Prefix = "`la`"
Disabled word = "`media`"

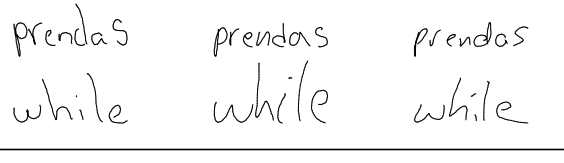
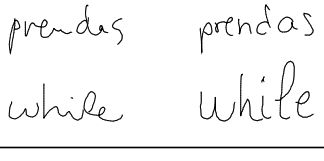
From the original 2-gram model (used by the off-line HTR system) a prefix-and-error conditioned 2-gram sub-model is derived, which takes as initial state that corresponding to the prefix "`la`". This simplified language model is used by the on-line HTR feddback sub-system to recognize the handwritten word "`edad`", intended to replace the wrong off-line misrecognized word "`media`", disabled by this model.

MM-CATTI experimental framework

- As in the case of CATTI, all the experiments assume interaction happens only at the word level; that is each interaction step involves the correction of a single, whole word from the system-predicted suffix
- Evaluating the MM-CATTI feedback subsystem requires on-line, touch-screen data. This data is the very essence of human interaction
- But, for experimental purposes, we can not afford the permanent availability of users to test the system
- The production of on-line feedback data can be properly simulated using a publicly available on-line handwritten corpus: UNIPEN
- Furthermore, this allows the experiments to be fully reproducible by strictly using public resources

Simulating touch-screen feedback with the UNIPEN corpus

- Every word that the user should write on the touchscreen needs to be obtained from UNIPEN data
- The touchscreen data for each needed word are generated by concatenating the data from random *character* samples from three UNIPEN categories: `1a` (digits), `1c` (lowercase letters) and `1d` (symbols)
- To increase realism, each needed word is produced only using characters belonging to the same writer, from three randomly chosen writers
- Examples of words generated using characters from three UNIPEN test writers, along with samples of the same words written by two real writers:

Words from concatenated UNIPEN chars	Real word writing
	

Touch-screen feedback and UNIPEN datasets

Training: 42 symbols and digits and 1 000 most frequent English and Spanish words generated from UNIPEN characters of 17 writters.

Test: For each off-line HTR task: number of *on-line* unique words and word instances needed as feedback to correct the off-line HTR word errors made by the plain off-line HTR system.

Task	Unique words	Word instances
ODEC-M3	378	753
IAMDB	510	755
CS	648	1 196
Total	1 536	2 704

Overall statistics of the UNIPEN training and test data used in the experiments

Number of different:	Train	Test	Lexicon
writers	17	3	-
digits (1a)	1 301	234	10
letters (1c)	12 298	2 771	26
symbols (1d)	3 578	3 317	32
total characters	17 177	6 322	68

MM-CATTI on-line HTR decoding accuracy

Writer average MM-CATTI feedback decoding error rates for the different corpora and three language models: plain unigram (U, *baseline*), error-conditioned unigram (U_e) and prefix-and-error conditioned bigram (B_e). The relative accuracy improvements for U_e and B_e with respect to U are shown in the last column.

Corpus	Lexicon	Feedback ER (%)			Relative Improv. (%)	
		U	U_e	B_e	U_e	B_e
ODEC-M3	2 790	5.1	5.0	3.1	2.0	39.2
IAMDB	8 017	4.6	4.3	3.5	6.5	23.9
CS-*page*	2 623	6.4	6.2	5.8	3.1	9.3

U, U_e and B_e correspond to $P(d)$, $\Pr(d|e)$ and $P(d|p,e)$, respectively.
The *baseline* is U, with no interaction-derived constraints.

CATTI / MM-CATTI empirical results

From left-to-right: post-editing corrections (WER), interactive corrections needed (WSR), contributions of both input modalities: on-line touch-screen (TS) and keyboard (KBD), and overall estimated effort reduction (EFR). All results are percentages

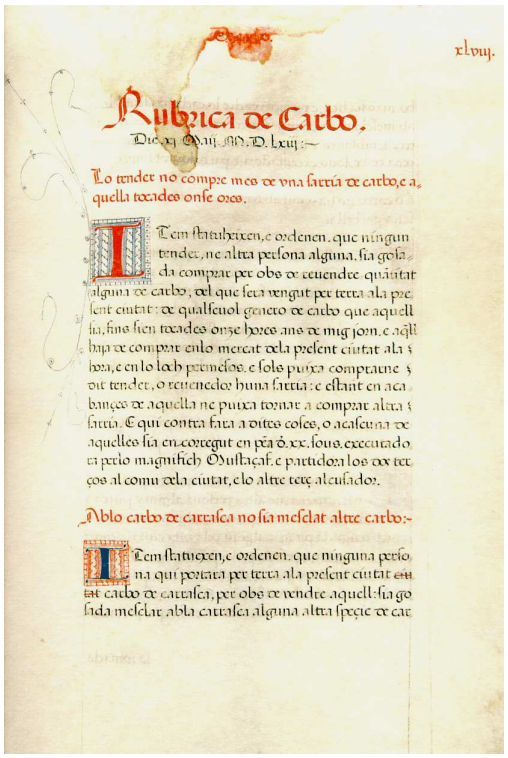
Corpus	Post-edit	CATTI	MM-CATTI		Overall EFR	
	WER	WSR	WSR_U	WSR_{B_e}	CATTI	MM-CATTI (U / B_e)
ODEC	22.9	18.9	19.7	19.5	17.5	14.0 / 14.8
IAMDB	25.3	21.1	22.1	21.8	16.6	12.6 / 13.8
CS-*page*	28.5	26.9	28.6	28.4	5.6	-0.04 / 0.40

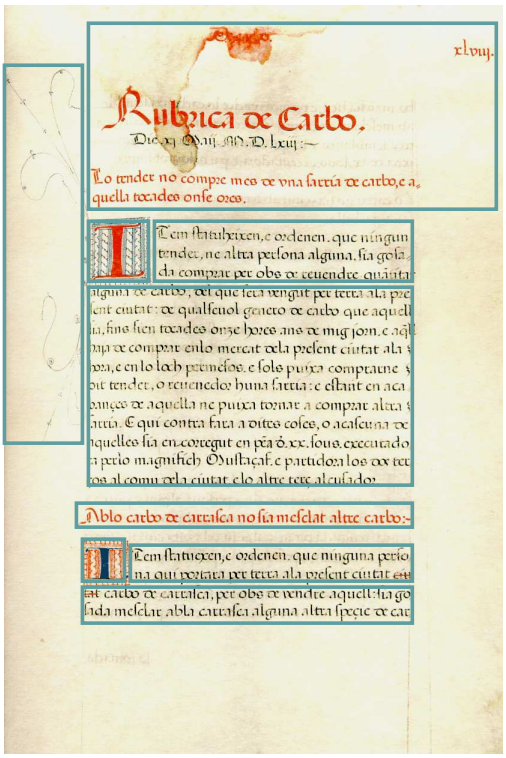
- WER: word error rate
- MM-CATTI values corresponded to $P(d)$ and $P(d \mid p,e)$ respectively
- WSR: word error rate and word stroke ratio. In the MM-CATTI case, these take into account the total amount of interactions (touchscreen + keystroke inputs)
- EFR: overall estimated effort reduction
  (assuming equal effort for TS and KBD interactions)

Multimodal Interaction for other Document Processing tasks?

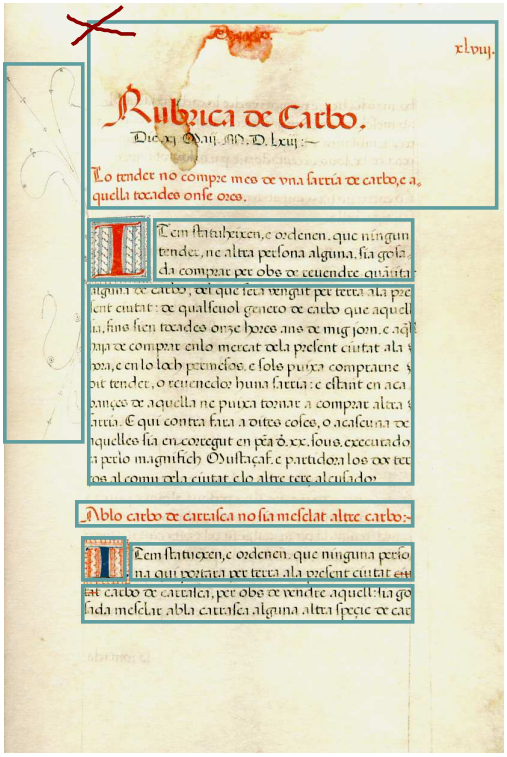
- Image Restoration and Enhancing
- ***Layout Analysis***
- Skew Correction
- Line Detection
- Slant Correction
- Size Normalization
- ... more?

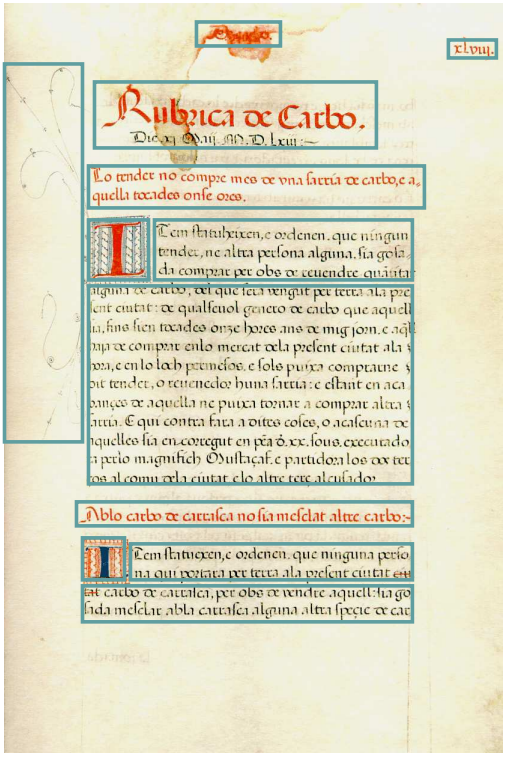
Multimodal Interaction for Document Layout Analysis?





Multimodal Interaction for Document Layout Analysis?





General Conclusions of the Tutorial

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

Bibliography

- E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera and C. Martínez. "Computer-assisted translation using speech recognition". IEEE Trans. on Audio, Speech and Language Processing, 14(3):941-951, 2006.
- 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 and E. Vidal. "Computer Assisted Transcription of Text Images and Multimodal Interaction", 5th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-08), Utrecht, The Netherlands, September 2008.

Index

MM-CATTI DEMO