Handwritten Text Recognition for Ancient Documents

A. Juan, V. Romero, J.A. Sánchez, N. Serrano, A.H.Toselli, E. Vidal
Institut Tecnològic d'Informàtica
Universitat Politècnica de València

Workshop on Applications of Pattern Analysis, September 2010

HTR for Ancient Documents

Problem: Off-line transcription of handwritten text images

Applications:

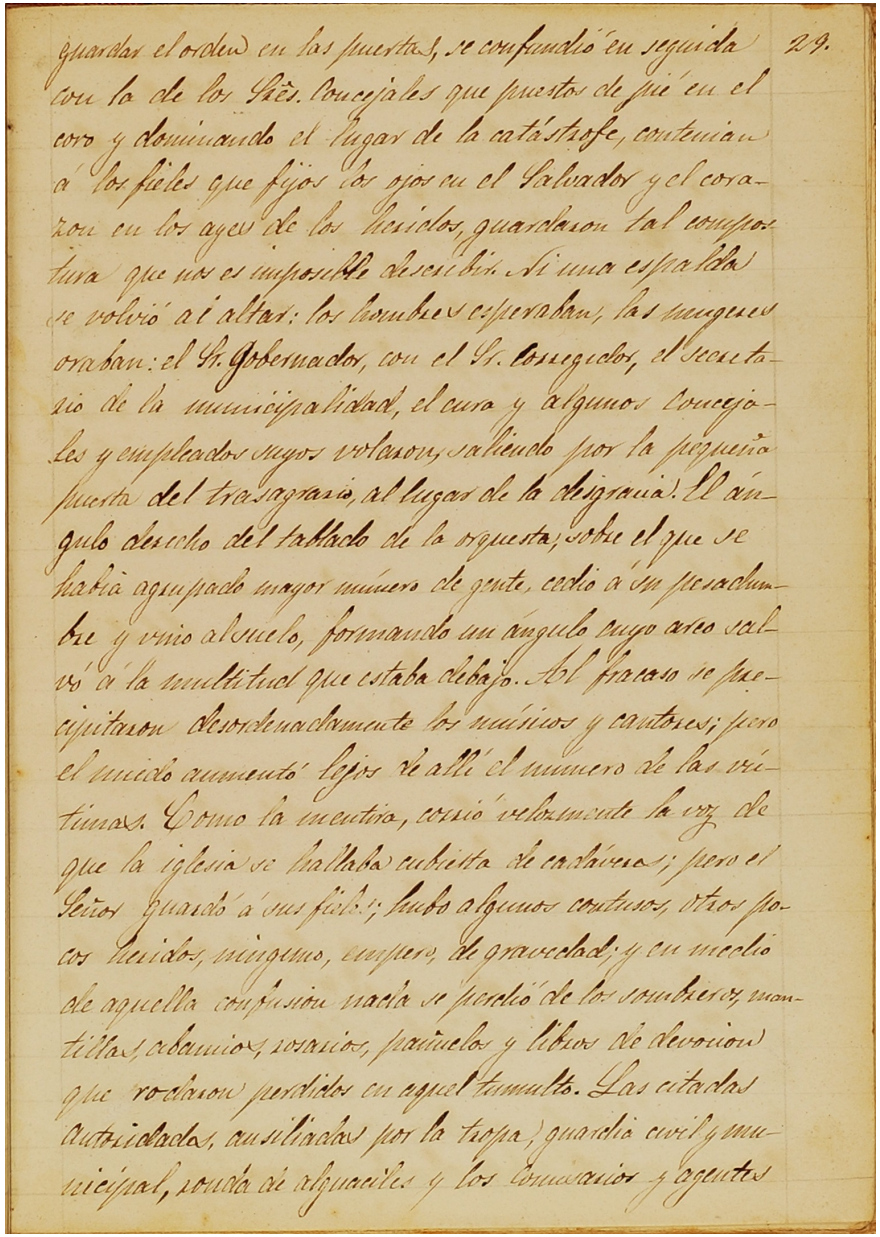
- Historical document collections residing in libraries, museums and archives
- Literary masterpieces

Approaches:

➢ OCR technologies are very far from offering useful solutions HTR

➢ Off-line HTR without any prior segmentation . . .

. . . but a Computer Assisted Interactive process



HTR Technology Overview

➢ **Document image preprocessing:**

- skew correction
- background removal
- noise reduction

➢ **Line image extraction:**

- slant correction
- non-linear size normalization

➢ **HMM training/decoding:**

- Lines represented as sequences of feature vectors
- Characters modeled by continuous density HMM
- Words modeled by stochastic finite-state automata
- Text lines modeled by using N-grams
- Baum-Welch algorithm for training
- Viterbi algorithm for decoding

MM-CATTI: MultiModal Interactive Transcription

Scenario: Totally correct transcriptions with minimal human effort

Approach: Multimodal Computer Assisted Transcription of Text Images

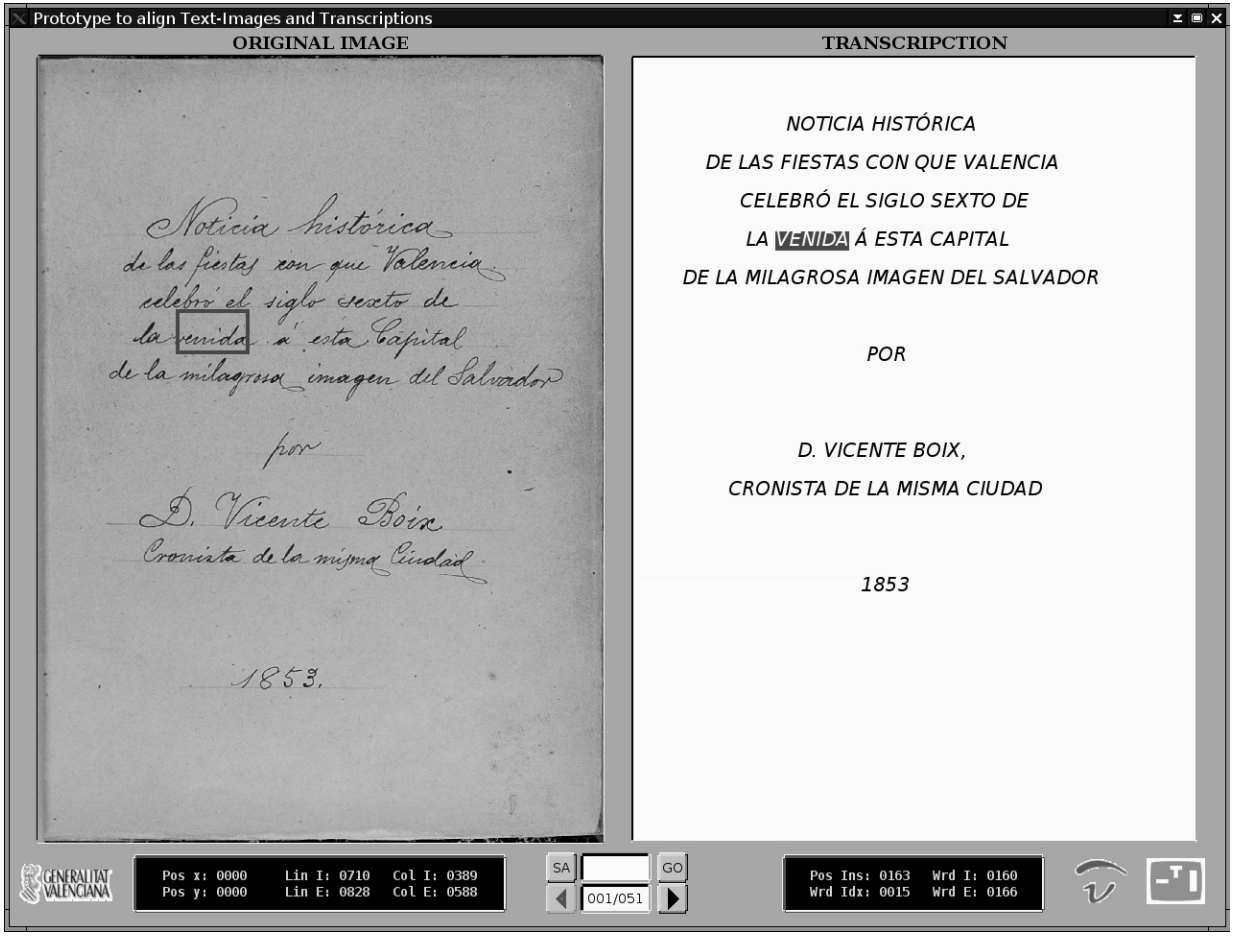
System:
- Interactive-predictive HTR
- Multimodal interfaces

	x								
INTER-0	p								
INTER-1	$\widehat{s} \equiv \widehat{w}$	**antiguos**	**cuidadores**	**que**	**en**	**el Castillo**	**sus**	**llamadas**	
	p'	**antiguos**							
	v		ciudadanos						
	p	**antiguos**	**ciudadanos**						
INTER-2	$\widehat{s}$			**que**	**en**	**el Castillo**	**sus**	**llamadas**	
	p'	**antiguos**	**ciudadanos**	**que**	**en**				
	v					Castilla			
	p	**antiguos**	**ciudadanos**	**que**	**en**	**Castilla**			
FINAL	$\widehat{s}$						**se**	**llamaban**	
	v								#
	$p \equiv t$	**antiguos**	***ciudadanos***	**que**	**en**	***Castilla***	**se**	**llamaban**	

Aligning Text-Images and Transcriptions

Scenario: Handwritten text images and their proper transcriptions are available

Approach: Viterbi alignment based on HMM



GIMP-based Interactive Transcription Tool

Scenario: HTR applications in which obtaining perfect transcriptions is not mandatory

Approach: Train from partially supervised transcriptions

System:

- Interactive-predictive HTR built on top of the GIMP
- Constantly updated models with previously annotated documents
- Active Learning techniques for sample selection
- Maximum error allowed is specified in advance