# Handwritten Text Recognition for Ancient Documents

**Alfons Juan** AJUAN@ITI.UPV.ES
**Verónica Romero** VROMERO@ITI.UPV.ES
**Joan Andreu Sánchez** JANDREU@ITI.UPV.ES
**Nicolás Serrano** NSERRANO@ITI.UPV.ES
**Alejandro H. Toselli** AHECTOR@ITI.UPV.ES
**Enrique Vidal** EVIDAL@ITI.UPV.ES
***Institut Tecnològic d'Informàtica***
***Universitat Politècnica de València***
***Camí de Vera s/n, 46022 València, Spain***

**Editors:** Tom Diethe, Nello Cristianini, John Shawe-Taylor

## Abstract

Huge amounts of legacy documents are being published by on-line digital libraries world wide. However, for these raw digital images to be really useful, they need to be transcribed into a textual electronic format that would allow unrestricted indexing, browsing and querying. In some cases, adequate transcriptions of the handwritten text images are already available. In this work three systems are presented to deal with this sort of documents. The first two address two different approaches for semi-automatic transcription of document images. The third system implements an alignment method to find mappings between word images of a handwritten document and their respective words in its given transcription.

## 1. Introduction

Huge historical document collections residing in libraries, museums and archives are currently being digitalized for preservation purposes and to make them available worldwide through large, on-line digital libraries. However, efforts should also focus on technologies aimed at reducing the human effort required for the annotation of the raw images with informative content. In the case of text images, which are among the most numerous and interesting, the most informative annotation level is their (paleographic) transcription into an adequate textual electronic format that would provide new ways of indexing, consulting and querying these documents.

Given the kind of (typically handwritten) text images involved in ancient documents, currently available OCR text recognition technologies are very far from offering useful solutions to the transcription problem, since, in the vast majority of the ancient text images of interest, characters can by no means be isolated automatically. The required technology should be able to recognize all text elements (sentences, words and characters) as a whole, without any prior segmentation of the image into these elements. This technology is generally refered to as "*off-line Handwritten Text Recognition*" (HTR) (Marti and Bunke, 2001).

For (high quality, modern) unrestricted text images, current HTR state-of-the-art (research) prototypes provide accuracy levels that range from 20 to 60% word error rate (Marti and Bunke, 2001; Toselli et al., 2004). Clearly, if high- or moderate-quality transcriptions are needed, the only possibility to use these systems is by acknowledging the need of a human-expert revision or "*post-editing*" of the system results. Given the high error rates involved, such a post-editing solution is quite inefficient and uncomfortable and is not generally accepted by expert transcribers. Therefore, only semi-automatic or *computer-assisted* solutions can be currently foreseen to cope with the text-image transcription bottleneck which hinders useful accessibility to old document data. In (von Ahn et al., 2008) collaborative human correction of an OCR system was succesfully applied to old text printed documents.

Depending on the specific requirements of the different document transcription tasks, two kind of produced transcription qualities are distinguished: transcriptions containing some controlled amount of errors and totally correct transcriptions. Transcriptions of the first type can be used as metadata for indexing, consulting and querying documents, while the second type corresponds to the conventional paleographic transcriptions of manuscripts. In this work we describe two HTR systems that incorporate human experts in the correction process.

The GIDOC system (Serrano et al., 2010), which is described in Section 3 allows us to reduce error location effort, by highlighting recognized words with low confidence. The user can supervise and correct these words, if needed. This partial supervision does help the system to adaptively train its statistical models for greater accuracy.

If totally correct transcriptions are needed, the *multimodal interactive-predictive approach* can be used (Toselli et al., 2009; Romero et al., 2009a) as an effective alternative to post-editing. In this approach the HTR system and the human transcriber tightly cooperate to generate the final transcription. The MM-CATTI system is presented in Section 4.

While the vast majority of legacy documents are currently only available in the form of digital images, for a significant amount of these documents (manually produced) transcriptions are already available. This fact has motivated the development of *automatic alignment* techniques which generate a mapping between each line and word on a document image and its respective line and word on its electronic transcript. Section 5 presents an image-transcription alignment system which carries out this task.

For each of the three previously systems, we identify a concrete real scenario in which the system could be useful, or we identify potential users who would be benefited of the technology that is introduced.

## 2. HTR Technology Overview

The implementation of the systems described in this work involves three common different parts: document image preprocessing, line image feature extraction and Hidden Markov Model (HMM) training/decoding (Bazzi et al., 1999).

Document image preprocessing encompasses the following steps: first, skew correction, background removal and noise reduction (Kavallieratou and Stamatatos, 2006) are performed on each document page image. Next, the page image is divided into separate line images profile (Marti and Bunke, 2001). Finally, slant correction and non-linear size normalization are applied to each extracted line image.

As our recognition and alignment systems are based on HMMs, each preprocessed line image is represented as a *sequence of feature vectors*. The feature extraction module applies a grid to divide line image into squared cells. In each cell, three features are calculated: normalized gray level and horizontal and vertical gray level derivatives (Toselli et al., 2004). Columns of cells or "frames" are processed from left to right and a feature vector is constructed for each frame by stacking the three features computed in its constituent cells.

Characters are modeled by continuous density left-to-right HMMs whose number of HMM states and Gaussian densities per state is determined by tuning empirically the system on several corpora. Once a HMM topology has been adopted, the model parameters can be easily trained from images of continuously handwritten text lines (without any kind of word or character segmentation) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called forward-backward or Baum-Welch re-estimation.

Lexical entries (words) are modelled by stochastic finite-state automata, representing all possible concatenations of individual characters to compose the word. The concatenation of words to form text lines or sentences is statistically modelled using *N-grams*.

All these character, word and text models are easily *integrated* into a single *global* model on which a search process is performed for decoding each input sequence of feature vectors into an output sequence of recognized words. This search is efficiently carried out by using the Viterbi algorithm, which provides detailed word alignment information as a byproduct.

## 3. GIDOC: Gimp-based Interactive Transcription of Old Text Documents

In some handwritten text recognition applications, obtaining perfect (error free) transcriptions is not mandatory. If few errors are allowed, the system role is to reduce the user effort beyond simply ignoring words, and to adapt the system correctly from partially correct transcriptions. This approach has been developed and tested in the GIDOC system (Serrano et al., 2009, 2010), where we showed that systems trained from partially supervised transcriptions achieve results similar to those of systems trained from fully supervised transcriptions. Moreover, the maximum error allowed in final transcriptions can be pre-specified, resulting in further user effort reduction.

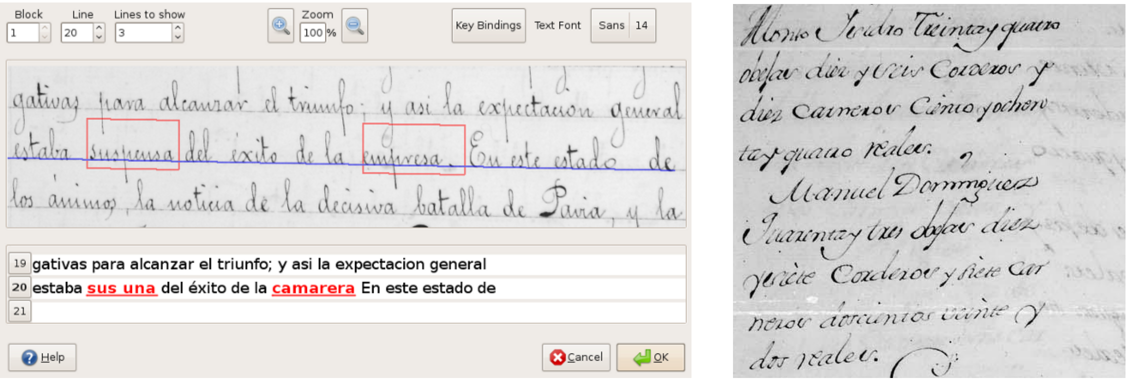
GIDOC[1] is a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. It is built on top of the well-known GNU Image Manipulation Program (GIMP). Homogeneous documents are annotated with GIDOC, by supervising hypotheses drawn from HTR models that are constantly updated with the increasing number of previously annotated documents.

In the proposed framework, the user specifies the maximum error allowed in the final (after human supervision) transcriptions. Next, each text line image is processed in turn, by first predicting its most likely transcription, and then automatically locating probable system errors. In order to locate these errors, GIDOC resorts to (word-level) confidence measures, which are calculated as posterior word probabilities estimated from word graphs.

[1] `http://prhlt.iti.es/gidoc.php`

Active Learning techniques (Settles, 2009) for sample selection, as well as semisupervised learning techniques for model adaptation are incorporated in GIDOC. Figure 1 in the left shows an example of GIDOC prototype.

As previously mentioned, GIDOC approach could be useful for accessing documents for which perfect transcription is not mandatory. For example, "Catastro de Ensenada" is a vast collection of handwritten documents related to property register of Spain from XVIII century that is accessible through a web portal at "Portal de Archivos Españoles"[2] (see Figure 1 right). Each old town has its own entry in this register and, sometimes it is consulted for checking properties, or for historical studies. It makes no sense to transcribe all this collection, but a useful alternative would be to have available a tool like GIDOC, that could help to check these documents.



Block 1 Line 20 Lines to show 3 Zoom 100 % Text Font Sans 14

19 gativas para alcanzar el triunfo; y asi la expectacion general
20 estaba sus una del éxito de la camarera En este estado de
21

Alonso Serrano Treinta y quatro
obejas diez y seis Corderos y
diez Carneros Cinco y ochen
ta y quatro reales.
Manuel Dominguez
Quarenta y tres obejas diez
y siete Corderos y siete Car
neros doscientos veinte y
dos reales.

Figure 1: Left: GIDOC interactive transcription dialog. Right: Partial example of a document of the "Catastro de Ensenada" collection.

## 4. MM-CATTI: MultiModal Interactive Transcription of Text Images

In contrast with the objectives of GIDOC, here the aim is to achieve *totally correct* transcriptions with minimal human effort. This is the approach using the "Computer Assisted Transcription of Text Images (CATTI)" technology (Toselli et al., 2009; Romero et al., 2009a). In the CATTI framework, the user is involved in the transcription process since she is responsible of validating and/or correcting prefixes of increasing size from the HTR output (Toselli et al., 2009). The protocol that rules this process can be formulated as an iterative process iterated until a correct transcription is obtained (see Figure 2).

In order to improve human transcriber productivity and to make the previous defined protocol friendlier for the user, "pure" *mouse action feedback* was studied in detail in (Romero et al., 2009b).

Furthering the goal of making the iteration process friendlier to the user, more ergonomic multimodal interfaces (MM-CATTI) such as touchscreen have been studied[3]. It is worth noting, however, that the use of this more ergonomic feedback modality comes at the cost of new, additional interaction steps needed to correct possible feedback decoding errors.

[2] http://pares.mcu.es

[3] htt://catti.iti.upv.es

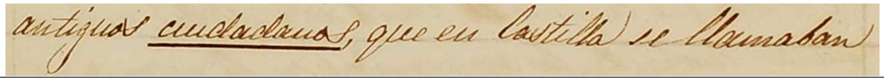
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $x$ |  | | | | | | | |
| INTER-0 | $p$ | | | | | | | | |
| INTER-1 | $\widehat{s} \equiv \widehat{w}$ | **antiguos** | **cuidadores** | **que** | **en** | **el Castillo** | **sus** | **llamadas** | |
| | $p'$ | **antiguos** | | | | | | | |
| | $v$ | | 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** | |

Figure 2: Example of CATTI interaction to transcribe an image of the Spanish sentence *"antiguos ciudadanos que en Castilla se llamaban"* (7 words). Initially the prefix $p$ is empty, and the system proposes a complete transcription $\widehat{w}$ of the input image $x$. This initial, pure automatic transcription has 5 words which need to be corrected. In each iteration step the user reads the current transcription and accepts a correct prefix $p'$. Then, he or she types in some word, $v$, to correct the erroneous text that follows the validated prefix, thereby generating a new prefix $p$ (the accepted one $p'$ plus the word $v$ added by the user). At this point, the system suggests a suitable continuation $\widehat{s}$ of this prefix $p$ and this process is repeated until a complete and correct transcription of the input image is reached. In the final transcription, $t$, the 2 words typed by the user are underlined. In this example the estimated post-editing effort would be 5/7 (71%), while the corresponding interactive estimate is 2/7 (29%). This results in an estimated user effort reduction of 59%.

To have available perfect transcription is fundamental when dealing with handwritten text that are important literary masterpieces or historical manuscripts (like correspondence between high level institutions). Transcribing this documents is usually carried out by particular paleographers, public institutions[4], or private companies[5], who would be greatly benefited by the CATTI technology.

## 5. Aligning Text-Images and Transcriptions

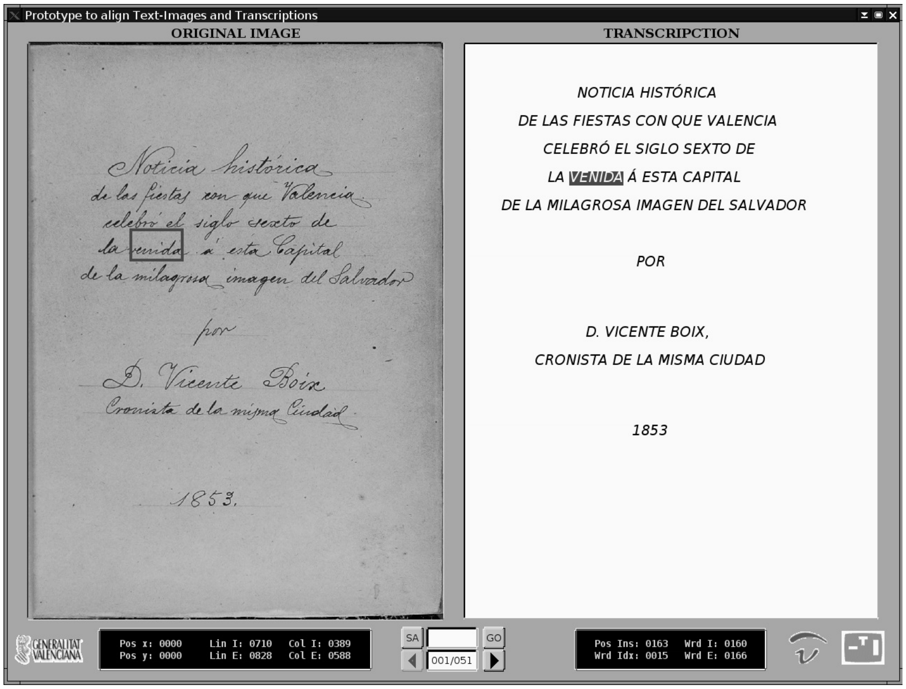
As mentioned in the introduction, in many cases both the handwritten text images and their proper (manually produced) transcriptions (in ASCII or PDF format) are available. Here we present a system that aligns these documents and their transcripts; i.e. it generates a mapping between each word image on a document page with its respective ASCII word on its transcript (Toselli et al., 2007).

Two different levels of alignment can be defined: line level and word level. Line alignments attempt to obtain beginning and end positions of lines in transcribed pages that do not have synchronized line breaks. This information allows users to easily visualize the page

[4] See "Centre for Manuscript and Print Studies" at `http://ies.sas.ac.uk/cmps/index.htm`

[5] See MICRONET at `http://admyte.com/contenido.htm`

image documents and their corresponding transcriptions. Moreover, using these alignments as segmentation ground truth, large amounts of training and test data for segmentation-free cursive handwriting recognition systems become available. On the other hand, word alignments allow users to easily find the place of a word in the manuscript image when reading the corresponding transcript (see Figure 3). The alignment method implemented here relies on the Viterbi decoding approach to HTR based on HMMs outlined in Section 2.



Noticia histórica
de las fiestas con que Valencia
celebró el siglo sexto de
la venida á esta Capital
de la milagrosa imagen del Salvador
por
D. Vicente Boix
Cronista de la misma Ciudad
1853.

NOTICIA HISTÓRICA
DE LAS FIESTAS CON QUE VALENCIA
CELEBRÓ EL SIGLO SEXTO DE
LA VENIDA Á ESTA CAPITAL
DE LA MILAGROSA IMAGEN DEL SALVADOR

POR

D. VICENTE BOIX,
CRONISTA DE LA MISMA CIUDAD

1853

```
Pos x: 0000    Lin I: 0710   Col I: 0389
Pos y: 0000    Lin E: 0828   Col E: 0588
```

```
Pos Ins: 0163   Wrd I: 0160
Wrd Idx: 0015   Wrd E: 0166
```

Figure 3: Screen-shot of the alignment prototype interface displaying an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word in the transcript (right).

Currently, it is more and more usual to have available digitalized documents of historical masterpieces belonging to relevant writers or researchers [6] that can be checked by the general public. The technology described in this section can used to make easier the reading of this sort of documents.

## 6. Concluding remarks

In this paper, three systems have been described: two of them aiming at assisting in the transcription of handwriting old documents, while the third focuses on the alignment between handwritten text images and their corresponding transcriptions.

One of the assisted transcription systems (GIDOC) employs a simple yet effective method to find an optimal balance between recognition error and supervision effort in handwrit-

[6] For example, some William Shakespeare's manuscripts digitalized documents are available at `http://www.amazon.com`, and some C. Darwin's manuscripts togheter with their transcription can be seen at `http://darwin-online.org.uk/manuscripts.html`.

ing text recognition. The other assisted transcription system (MM-CATTI) is based on an interactive-predictive framework which combines the efficiency of automatic HTR systems with the accuracy of the experts in the transcription of ancient documents. Finally, for handwritten manuscripts whose transcriptions are already available, the presented alignment system maps every line and word image on the manuscript with its respective line and word on the electronic (ASCII or PDF) transcript.

The interactive-predictive transcription systems that have been introduced will allow us to explore interesting learning scenarios. The interactive framework is an appropriate scenario for online learning by taking immediate profit of the feedback provided by the user. In addition , this framework will allow us to explore active learning techniques by choosing actively the most informative text image to be annotated. In addition, multimodal input can also be explored as a new form of user interaction.

## Acknowledgments

Work supported by the EC (FEDER, FSE) and the Spanish Government (MICINN, ”Plan E”, under grants MIPRCV ”Consolider Ingenio 2010”, MITTRAL TIN2009-14633-C03-01, iTrans2 TIN2009-14511 and FPU AP2007-02867), and Generalitat Valenciana (Prometeo/2009/014 and ACOMP/2010/051).

## References

I. Bazzi, R. Schwartz, and J. Makhoul. An Omnifont Open-Vocabulary OCR System for English and Arabic. *IEEE Trans. on PAMI*, 21(6):495–504, 1999.

E. Kavallieratou and E. Stamatatos. Improving the quality of degraded document images. In *Proc. of DIAL*, pages 340–349, Washington, USA, 2006.

U.-V. Marti and H. Bunke. Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. *IJPRAI*, 15(1):65–90, 2001.

V. Romero et al. Interactive multimodal transcription of text images using a web-based demo system. In *Proc. of IUI*, pages 477–478. Florida, 2009a.

V. Romero et al. Using mouse feedback in computer assisted transcritpion of handwritten text images. In *Proc. of ICDAR*, Barcelona (Spain), 2009b.

N. Serrano, D. Pérez, A. Sanchis, and A. Juan. Adaptation from partially supervised handwritten text transcriptions. In *Proc. of ICMI-MLMI*, Cambridge (USA), 2009.

N. Serrano, A. Sanchis, and A. Juan. Balancing error and supervision effort in interactive-predictive handwritten text recognition. In *Proc. of IUI*, Hong Kong (China), June 2010.

B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

A. H. Toselli, A. Juan, and E. Vidal. Spontaneous Handwriting Recognition and Classification. In *Proc. of ICPR*, pages 433–436, Cambridge (UK), 2004.

A. H. Toselli, V. Romero, and E. Vidal. Viterbi Based alignment between Text Images and their Transcripts. In *Proc. of LaTeCH*, pages 9–16. Prague (Czech Republic), 2007.

A. H. Toselli, V. Romero, M. Pastor, and E. Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825, 2009.

Luis von Ahn et al. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468, 2008.