State, an Assisted Document Transcription System*

David Llorens
dllorens@lsi.uji.es

Andrés Marzal
amarzal@lsi.uji.es

Federico Prat
fprat@lsi.uji.es

Juan Miguel Vilar
jvilar@lsi.uji.es

Departamento de Lenguajes y Sistemas Informáticos
Universitat Jaume I
12071 Castelló (Spain)

ABSTRACT

State is an interactive system for ancient and handwritten document transcription with several input modalities for entering and correcting text. It has a flexible architecture that allows easy connection to different OCR systems.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Input devices and strategies; I.7.5 [**Document Capture**]: Optical character recognition (OCR); J.5 [**Arts and Humanities**]: Literature—*Ancient document transcription*

General Terms

Documentation, Human Factors

Keywords

Ancient documents, handwriting, text transcription.

1. OVERVIEW OF THE SYSTEM

Ancient and handwritten document transcription are important research areas for OCR systems [1, 2, 4]. Transcribing ancient documents presents important difficulties due to the state of the physical support of the texts, which usually has suffered from the passage of time and inadequate storage conditions; and the texts themselves have unusual fonts containing symbols not used anymore. Handwritten texts are also notoriously difficult due to the high variability in writing styles. This makes it necessary to shift the emphasis from automatic transcription into assisted transcription.

We have built State [3], a multimodal platform to assist the user over the whole process: from the conditioning of the scanned images of the document to the correction of the transcription provided by an OCR system. The user benefits from different input and interaction methods: keyboard, mouse, and stylus. Experiments have shown reductions in the time needed for transcribing a page of up to 50% [5].

*Work partially supported by the Spanish *Ministerio de Ciencia e Innovación* (TIN2006-12767 and *Consolider Ingenio 2010* CSD2007-00018), and *Fundació Caixa Castelló-Bancaixa* (P1·1B2006-31).

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
ACM 978-1-60558-772-1/09/11.

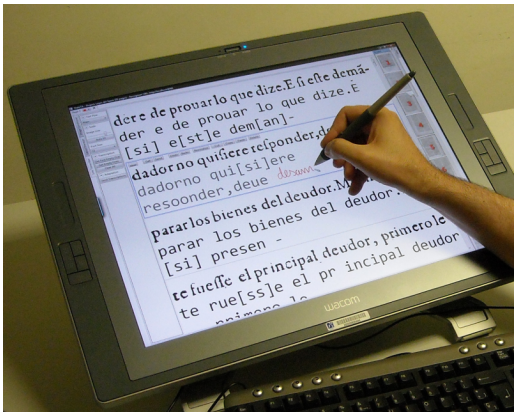

Figure 1: State on a pen-sensitive display.

State is written in C# on .NET 3.5. It is currently in use in the *Arxiu Jaume I*, a repository of ancient law documents, and in the *Biblioteca Virtual Miguel de Cervantes*, the main virtual library for Spanish cultural legacy.

State comprises a graphical front-end that can be connected to different text transcription back-ends. The front-end and the back-end are separate processes, possibly in different machines, that communicate via HTTP and XML.

2. THE FRONT-END

The front-end of State can be operated on a pen-sensitive display (Fig. 1). It works with projects, which are sets of images of the pages to be transcribed to which information is added, including the image processing operations applied, their layout, and their transcription.

The user can open several pages simultaneously. For each one, a pipeline of three stages is built: Image Conditioner, Layout Manager, and Assisted Line Transcriber.

The Image Conditioner (Fig. 2) offers commands to remove noise and to enhance the text in the image. This is used to correct possible scan defects or poor condition of documents due to moisture, torn pages, fading ink, staining, etc.

The Layout Manager (Fig. 3) offers commands to detect and edit the page layout, which is a hierarchical structure containing text flows composed of blocks and lines.

Finally, the Assisted Line Transcriber (Fig. 4) allows the user to obtain transcriptions of the lines using a transcrip-

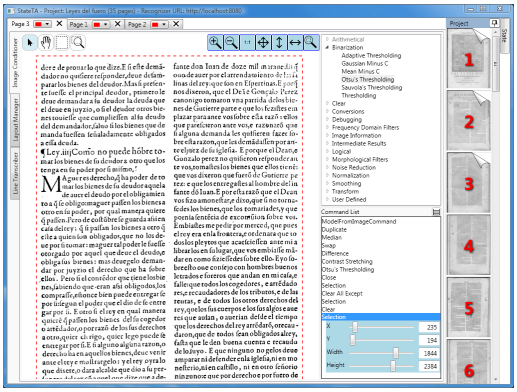

Figure 2: The Image Conditioner.

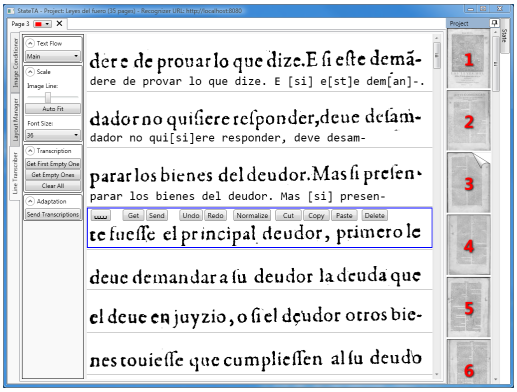

Figure 4: The Assisted Line Transcriber.

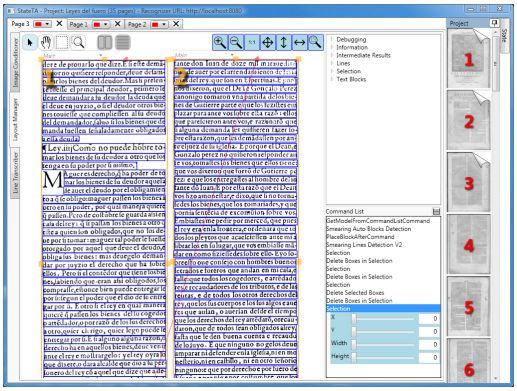

Figure 3: The Layout Manager.

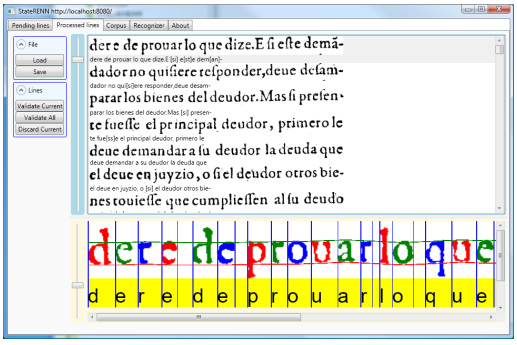

Figure 5: The Interactive Line Editor.

tion back-end and to manually correct them. The user can also send back any line and its corrected transcription to inform the back-end that they can be used for improving the decoder. The Assisted Line Transcriber handles pen, keyboard, and mouse input. It can be used with a Tablet PC, digitizing tablet, or pen-sensitive screen. The text transcription is shown under the image of the corresponding lines, so validation is quick and easy. The user can use both the keyboard and the stylus to edit the transcriber output.

3. THE BACK-ENDS

The front-end communicates with the back-end in order to transcribe lines or to provide it with text samples in order to improve its performance. This communication is performed using the XML format over the HTTP protocol.

Figure 5 shows the Interactive Line Editor that is used to extract character instances to train the transcription back-end for ancient documents. A pen-based segmentation and transcription editor assists the user in validating input data.

This back-end also contains an Interactive Corpus Editor that manages character instances. Its interface lets the user to delete or relabel character samples when appropriate.

The transcription back-end for ancient documents uses a dynamic programming Two-Level like segmentation and decoding algorithm that classifies segments using the Nearest-Neighbor rule on PCA coded character images. We have also experimented with a transcription back-end for handwritten documents which is based on HMMs and Neural Networks.

4. CONCLUSIONS

We have presented a system for assisting in the transcription of documents. It greatly improves productivity and it is specially useful for documents presenting difficulties like old typographies or handwritten texts.

5. REFERENCES

[1] F. Le Bourgeois and H. Emptoz. DEBORA: Digital AccEss to BOoks of the RenAissance. *International Journal on Document Analysis and Recognition*, 9(2-4):193–221, 2007.

[2] M. Droettboom, R. Ferguson, and I. Fujinaga. Gamera. `http://ldp.library.jhu.edu/projects/gamera`, 2007.

[3] Albert Gordo, David Llorens, Andrés Marzal, Federico Prat, and Juan Miguel Vilar. STATE: A multimodal assisted text-transcription system for ancient documents. In *The Eigth IAPR Workshop on Document Analysis Systems*, Nara (Japan), September 2008.

[4] METAe. The Metadata Engine Project. `http://meta-e.aib.uni-linz.ac.at`, 2003.

[5] Juan Miguel Vilar, María José Castro-Bleda, Francisco Zamora-Martínez, Salvador España-Boquera, Albert Gordo, David Llorens, Andrés Marzal, Federico Prat, and Jorge Gorbe. A flexible system for document processing and text transcription. To appear in Proceedings of CAEPIA'09.