The GIDOC Prototype

N. Serrano, L. Tarazón, D. Pérez, O. Ramos Terrades and A. Juan

DSIC/ITI, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain

Abstract. Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. In this paper, an efficient interactive-predictive transcription prototype called GIDOC (Gimp-based Interactive transcription of old text DOCuments) is presented. GIDOC is a first attempt to provide integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. It is based on GIMP and uses advanced techniques and tools for language and handwritten text modelling. Results are given on a real transcription task on a 764-page Spanish manuscript from 1891.

1 Introduction

Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. It might be carried out by first processing all document images off-line, and then manually supervising system transcriptions to edit incorrect parts. However, current techniques for automatic page layout analysis, text line detection and handwriting recognition are still far from perfect [10, 2, 1], and thus post-editing system output is not clearly better than simply ignoring it.

A more effective approach to transcribe old text documents is to follow an interactive-predictive paradigm in which both, the system is guided by the user, and the user is assisted by the system to complete the transcription task as efficiently as possible. Following this approach, a system prototype called GIDOC (Gimp-based Interactive transcription of old text DOCuments) has been developed to provide user-friendly, integrated support for interactive-predictive layout analysis, line detection and handwriting transcription [4, 7].

GIDOC is designed to work with (large) collections of homogeneous documents, that is, of similar structure and writing styles. They are annotated sequentially, by (partially) supervising hypotheses drawn from statistical models that are constantly updated with an increasing number of available annotated documents. And this is done at different annotation levels. For instance, at the level of page layout analysis, GIDOC uses a novel text block detection method in which conventional, memoryless techniques are improved with a "history" model of text block positions [4]. Similarly, at the level of text line image transcription, GIDOC includes a handwriting recognizer which is steadily improved with a growing number of partially supervised transcriptions [7]. Also at this level, the user is allowed to decide on a maximum tolerance threshold for the recognition error (in non-supervised parts), and the system adjusts the required supervision effort on the basis of an estimate for this error [8].

Serrano N., TarazÃşn L., PÃľrez D., Ramos Terrades O. and Juan A.
The GIDOC Prototype.
DOI: 10.5220/0003028300820089
In *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems (ICEIS 2010)*, page
ISBN: 978-989-8425-14-0

This paper presents a comprehensive description of the GIDOC prototype, with special emphasis on parts not previously described [4, 7, 8]. After an overview of GIDOC in Section 2, its main functions are described in Sections 3 (block and line detection), 4 (HTK training) and 5 (transcription). Experiments are reported in Section 6, and conclusions are discussed in Section 7.

2 System Overview

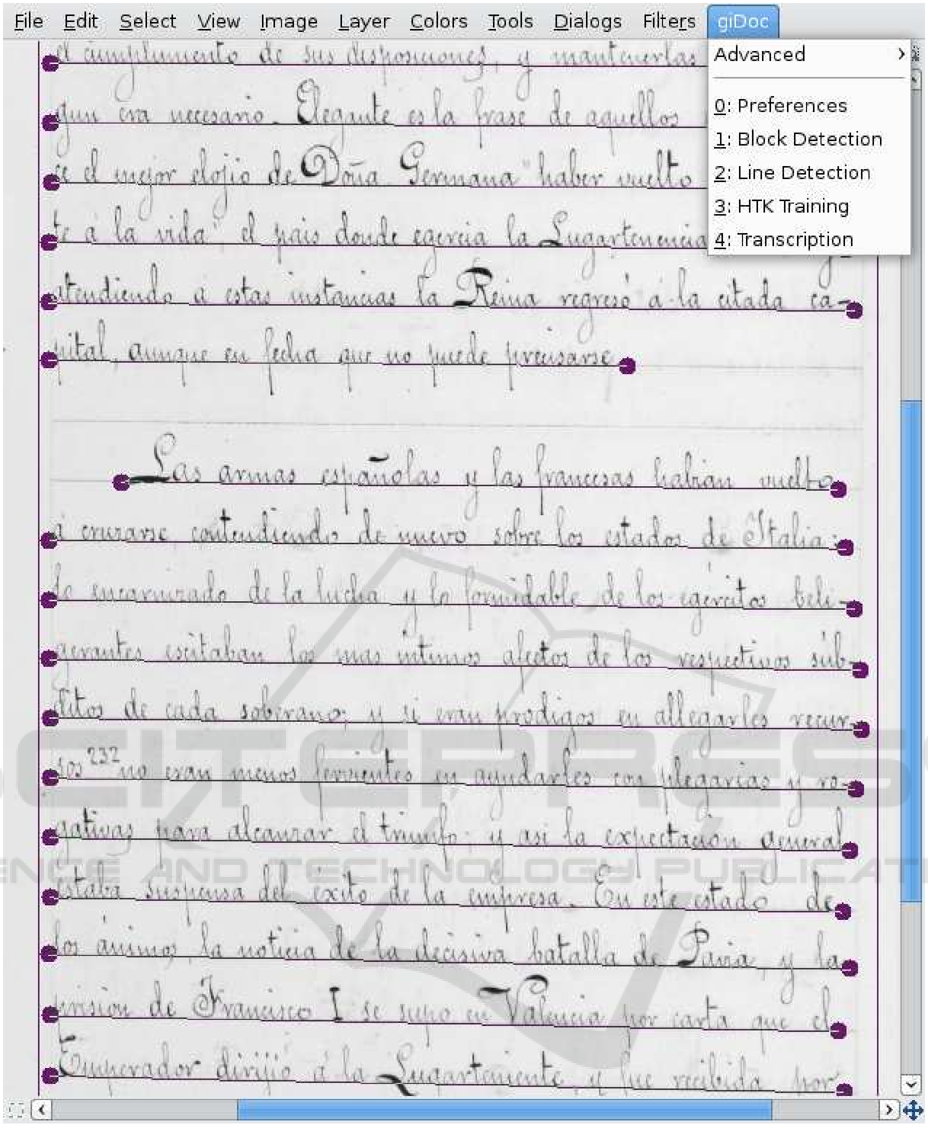
As indicated by its name, GIDOC has been implemented on top of the well-known GNU Image Manipulation Program (GIMP). As GIMP, GIDOC is licensed under the GNU General Public License, and it can be downloaded from [6]. To run GIDOC, we must first run GIMP and open a document image. GIMP will come up with its high-end user interface, which is often configured to only show the main toolbox (with docked dialogs) and an image window. GIDOC can be accessed from the menubar of the image window (see Fig. 1).

As shown in Fig. 1, the GIDOC includes six entries: *Advanced options, 0: Preferences, 1: Block Detection, 2: Line Detection, 3: HTK Training,* and *4: Transcription. Advanced options* is a second-level menu where experimental features are grouped. *Preferences* opens a dialog to configure global options, as well as more specific options for preprocessing, training and recognition. Some of them are discussed below together with menu entries after *Preferences.*

3 Block and Line Detection

During its development, GIDOC has been mainly tested on a old book in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines, as in the example shown in Fig. 1. As said in the introduction, GIDOC is designed to work with such homogeneous documents and, indeed, it takes advantage of their homogeneity. In particular, the *Block Detection* entry in the GIDOC menu uses a novel text block detection method in which conventional, memoryless techniques are improved with a “history” model of text block positions. Please see [4] for more information.

Given a textual block, the *Line Detection* entry in the GIDOC menu detects all its text baselines, which are marked as straight paths. The result can be clearly observed in the example of Fig. 1. Although each baseline has handlers to graphically correct its position, it is worth noting that the baseline detection method implemented works quite well, at least in pages like that of the example. It is a rather standard projection-based method [2]. First, horizontally-averaged pixel values or black/white transitions are projected vertically. Then, the resulting vertical histogram is smoothed and analyzed so as to locate baselines accurately. Two preprocessing options are included in *Preferences*, first, to decide on the histogram type (pixel values or black/white transitions), and second, to define the maximum number of baselines to be found. Concretely, this number is used to help the projection-based method in locating (nearly) blank lines.



el cumplimiento de sus disposiciones, y mantenerlas
quien era necesario. Elegante es la frase de aquellos
se el mejor elogio de Doña Germana haber vuelto
te a la vida el país donde ejercia la Lugartenencia
atendiendo a estas instancias la Reina regresó a la ciudad ca-
pital, aunque su fecha que no puede precisarse

Las armas españolas y las francesas habian vuelto
á cruzarse contendiendo de nuevo sobre los estados de Italia;
lo encarnizado de la lucha y lo formidable de los ejércitos beli-
gerantes excitaban los mas íntimos afectos de los respectivos súb-
ditos de cada soberano; y si eran pródigos en allegarles recur-
sos 232 no eran menos fervientes en ayudarles con plegarias y ro-
gativas para alcanzar el triunfo; y asi la expectacion general
estaba suspensa del éxito de la empresa. En este estado de
los ánimos, la noticia de la decisiva batalla de Pavia y la
prision de Francisco I se supo en Valencia por carta que el
Emperador dirijió á la Lugartenienta, y fue recibida por

Fig. 1. Image window showing GIDOC menu.

4 HTK Training

GIDOC is based on standard techniques and tools for handwritten text preprocessing and feature extraction, HMM-based image modeling, and language modeling [10]. Handwritten text preprocessing applies image denoising, deslanting and vertical size normalization to a given text (line) image. An illustrative example is given in Fig. 2. It can be configured through preprocessing options in *Preferences*. There is an option to

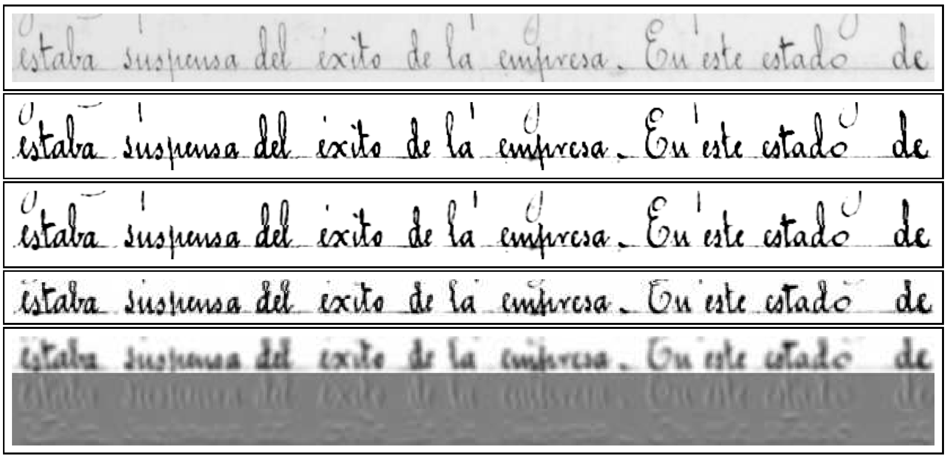


Fig. 2. Preprocessing and feature extraction of a text line image. From top to bottom: original image, denoising, deslanting, vertical size normalisation and feature extraction.

use instead a customized procedure, and two options to define (bounds for) the locations of the upper and lower lines, with respect to the baseline.

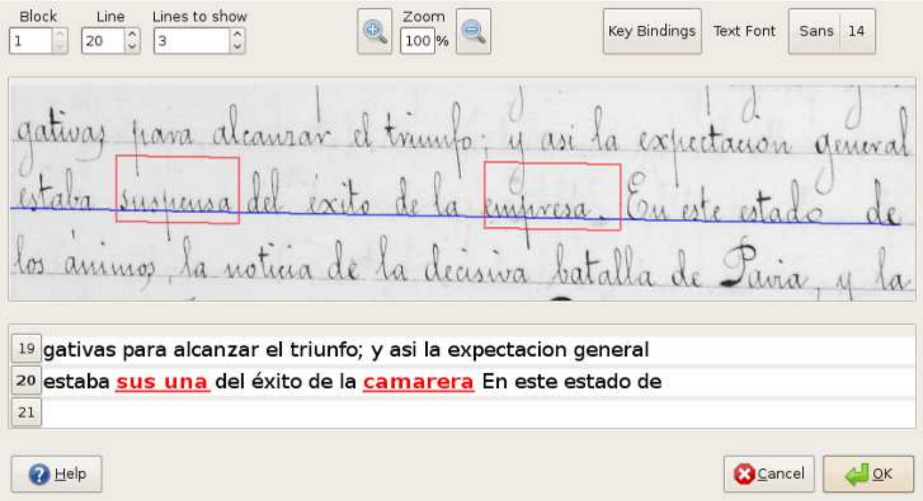
Feature extraction for HMM modeling consists in transforming the preprocessed image into a *sequence of (fixed-dimension) feature vectors*. There are two, well-known feature extraction methods available in GIDOC. The default method first divides the preprocessed image into a grid of square cells whose size is a small fraction of the image height (e.g. $1/20$). Then, each cell its characterized by its normalized gray level and, optionally, by its vertical and horizontal gray-level derivatives. See Fig. 2 for an example and [10] for more details. The alternative method moves a single-column window left-to-right over the image, and extracts 9 geometrical features at each position [1].

HMM image modeling is carried out with the well-known and freely available *Hidden Markov Model Toolkit (HTK).* [11]. Similarly, language modeling is implemented through the open source *SRI Language Modeling Toolkit (SRILM)* [9].

HTK Training reads the directory of task document images and, for each image, it extracts all its transcribed text lines, if any, together with their corresponding line images. Transcriptions are first preprocessed to isolate special characters (mainly punctuation signs) and expand abbreviations (e.g. *S.M.* is expanded to *Su Magestad*). Then, an n-gram language model is built from preprocessed transcriptions using a SRILM command which, by default, generates a bigram language model with Knesser-Ney discounting. On the other hand, extracted line images are preprocessed and transformed into sequences of feature vectors so as to train, using their corresponding transcriptions and HTK, continuous density (Gaussian) left-to-right HMMs at character level.

5 Transcription

The *Transcription* entry in the GIDOC menu opens the GIDOC interactive transcription dialog (see Fig. 3). It consists of two main sections: the image section, in the upper



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

Fig. 3. Interactive transcription dialog.

part, and the transcription section, in the bottom part. A number of text line images are displayed in the image section together with their transcriptions, if available, in separate editable text boxes within the transcription section. The *current* line to be transcribed or simply supervised is selected by placing the edit cursor in the appropriate editable box. Its corresponding baseline is emphasized (in blue color) and, whenever possible, GIDOC shifts line images and their transcriptions so as to display the current line in the central part of both the image and transcription sections. It is assumed that the user transcribes or supervises text lines, from top to bottom (or in any order desired), by entering text and moving the edit cursor with the arrow keys or the mouse. However, it is possible for the user to choose any order desired.

As can be seen in Fig. 3, each editable text box in the transcription section, has a button attached to its left. This button is labeled with the corresponding line number. By clicking on it, its associated line image is extracted, preprocessed, transformed into a sequence of feature vectors, and Viterbi-decoded using HTK and the models trained with *HTK training*. In this way, it is not needed to enter the complete transcription of the current line, but hopefully only minor corrections to the decoded output. Clearly, this is only possible if, first, text lines are correctly detected and, second, the HMM and language models are adequately trained, from a sufficiently large amount of training data. Therefore, it is assumed that transcription is carried out manually in early stages of a transcription task, and then is assisted as described here.

6 Experiments

During its development, GIDOC has been used by a paleography expert to annotate blocks, text lines and transcriptions on a new dataset called GERMANA [3]. GERMANA is the result of digitizing and annotating a 764-page Spanish manuscript from

1891, in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. The example shown in Fig. 1 corresponds to the page 144. GERMANA is solely written in Spanish up to page 180; then, the manuscript includes many parts that are written in languages different from Spanish, namely Catalan, French and Latin.

Due to its sequential book structure, the very basic task on GERMANA is to transcribe it from the beginning to the end, though here we only consider its transcription up to page 180. Starting from page 3, we divided GERMANA into 9 consecutive blocks of 20 pages each (18 in block 9) and, on average, 417 lines and 4687 running words. Then, from block 2 (pages 23–42) to block 9 (pages 163–180), each block was automatically transcribed by GIDOC trained with all preceding blocks. The results are plotted in Fig. 4, in terms of transcription Word Error Rate (WER). To avoid fluctuations due to varying test set complexity, the WER was also computed for a fixed block (block 9) after each GIDOC re-training, and the resulting WER curve has been added to Fig. 4. Also shown is the part of the WER due to the occurrence of out-of-vocabulary (OOV) words.

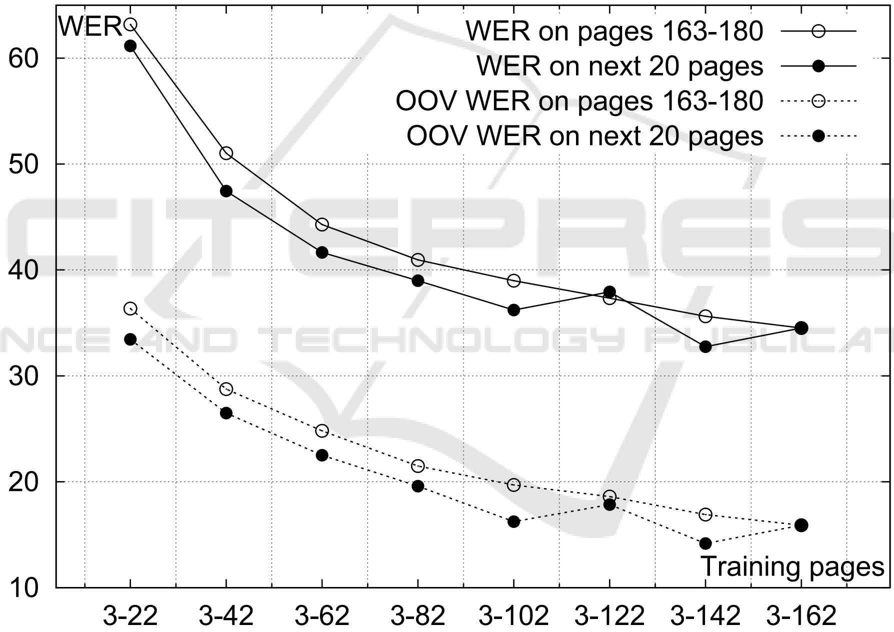


Fig. 4. Transcription Word Error Rate (WER) on GERMANA as a function of the pages already supervised and thus available for training (training pages). The WER is computed for both, the next 20 pages to supervise (solid line with black circles), and a fixed set comprising pages 163-180 (solid line with white circles). Also shown is the part of the WER due to the occurrence of out-of-vocabulary (OOV) words (dashed lines).

As expected, the WER decreases as the amount of training data increases. In particular, GIDOC achieves around 34% of WER for the last two blocks, which can be successfully used in computer-assisted transcription. The WER curve for block 9 does not differ significantly from that for the next block, though it appears that block 9 is a bit more complicated that all but one (block 7) of its preceding blocks. Regarding the OOV curves, it becomes clear that a considerable fraction of transcription errors is due to the occurrence of unseen words. More precisely, unseen words account for approximately 50% of transcription errors.

It is worth noting that preliminary WER results (only for the next block) have been already reported in [3] to accompany GERMANA description. In contrast to them, the WER and OOV curves reported here are slightly better on average (5.4% and 6.4%, respectively). This is mainly due to better modeling of word abbreviations and punctuation signs. Also, we have used an updated version of GERMANA baselines which are more accurately adjusted.

As with GERMANA, GIDOC has been used to annotate blocks, text lines and transcriptions on a more recent dataset called RODRIGO [5]. Although comparable in size to GERMANA, RODRIGO comes from a much older manuscript, from 1545, where the typical difficult characteristics of historical documents are more evident. In [5], experiments and results similar to those discussed here are reported.

7 Conclusions

A computer-assisted transcription prototype called GIDOC has been presented for handwritten text in old documents. GIDOC is a first attempt to provide integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. It is build on top of GIMP, and uses standard techniques and tools for handwritten text preprocessing and feature extraction, HMM-based image modeling, and language modeling. As GIMP, GIDOC is licensed under the GNU General Public License, and it can be freely downloaded from Internet. The effectiveness of GIDOC has been empirically demonstrated on the GERMANA database, which is also publicly available on Internet.

Acknowledgements

Work supported by the EC (FEDER, FSE), the Spanish Government (MEC, MICINN, ”Plan E”, under grants MIPRCV ”Consolider Ingenio 2010” CSD2007-00018, iTransDoc TIN2006-15694-CO2-01, MITTRAL TIN2009-14633-C03-01 and FPU AP2007-02867) and the Generalitat Valenciana (grant Prometeo/2009/014).

References

1. R. Bertolami and H. Bunke. Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41:3452–3460, 2008.
2. L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *International Journal on Document Analysis and Recognition*, 9, 2007.
3. Daniel Pérez, Lionel Tarazón, Nicolas Serrano, Francisco-Manuel Castro, Oriol Ramos-Terrades, and Alfons Juan. The GERMANA database. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 301–305, Barcelona (Spain), July 2009.

4. O. Ramos-Terrades, N. Serrano, A. Gordó, E. Valveny, and A. Juan. Interactive-predictive detection of handwritten text blocks. In *Document Recognition and Retrieval XVII (Proc. of SPIE-IS&T Electronic Imaging)*, pages 75340Q–(1–10), San Jose, CA (USA), January 2010.
5. N. Serrano, F. Castro, and A. Juan. The RODRIGO database. In *Proc. of the 8th Language Resources and Evaluation Conf. (LREC 2010)*, Valleta (Malta), May 2010.
6. Nicolás Serrano, Alfons Juan, et al. The GIDOC prototype. http://prhlt.iti.es/gidoc.php., 2009.
7. Nicolás Serrano, Daniel Pérez, Albert Sanchis, and Alfons Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. In *Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, pages 289–292, Cambridge, MA (USA), November 2009.
8. Nicolás Serrano, Albert Sanchis, and Alfons Juan. Balancing error and supervision effort in interactive-predictive handwriting recognition. In *Proc. of the 14th Int. Conf. on Intelligent User Interfaces (IUI 2010)*, Hong Kong (China), February 2010.
9. A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO (USA), 2002.
10. 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. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(4):519–539, 2004.
11. S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995.