Adaptation from Partially Supervised Handwritten Text Transcriptions*

Nicolás Serrano, Daniel Pérez, Albert Sanchis, and Alfons Juan
DSIC/ITI, Universitat Politècnica de València
Camí de Vera, s/n, 46022 València, Spain
{nserrano, dperez, asanchis, ajuan}@iti.upv.es

ABSTRACT

An effective approach to transcribe handwritten 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. This approach has been recently implemented in a system prototype called GIDOC, in which standard speech technology is adapted to handwritten text (line) images: HMM-based text image modelling, n-gram language modelling, and also confidence measures on recognized words. Confidence measures are used to assist the user in locating possible transcription errors, and thus validate system output after only supervising those (few) words for which the system is not highly confident. Here, we study the effect of using these partially supervised transcriptions on the adaptation of image and language models to the task.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*OCR; Document analysis*; I.5.5 [**Pattern Recognition**]: Implementation—*Interactive systems*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Computer-assisted Text Transcription, Confidence Measures, Document Analysis, Handwriting Recognition

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, state-of-the-art technologies for automatic page layout analysis, text line detection and handwritten text recognition are still far from perfect [7, 4, 1], and thus post-editing automatically generated 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 human supervisor, and the supervisor is assisted by the system to complete the transcription task as efficiently as possible. This computer-assisted transcription (CAT) approach has been successfully followed in the DEBORA [2] and iDoc [3] research projects, for old-style printed and handwritten text, respectively. In the case of iDoc, a CAT system prototype called GIDOC (Gimp-based Interactive transcription of old text DOCuments) has been developed to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription [3].

Here we will focus on the handwriting recognition part of GIDOC. As in the most advanced handwriting recognizers today, it is based on standard speech technology adapted to handwritten text images; that is, HMM-based text image modelling and n-gram language modelling. HMMs and the language model are trained from manually transcribed text lines during early stages of the transcription task. Then, each new text line image is processed in turn, by first predicting its most likely transcription, and then locating and editing system errors. In order to reduce the effort in locating these errors, GIDOC again resorts to standard speech technology and, in particular, to confidence measures (at word level), which are calculated as posterior word probabilities estimated from word graphs [6]. Recognized words below a given confidence threshold are marked as possible errors and the decision on how to proceed is left to the user. For instance, if a small number of transcription errors can be tolerated for the sake of efficiency, then the user might validate the system output after only supervising (a few) marked words. On the contrary, if at a first glance no significant portion of the text line seems to be correctly recognized, then the user might ignore system output and transcribe the whole text line manually.

Successively produced transcriptions can be used to better adapt image and language models to the task by, for instance, re-training them from the previous and newly acquired transcribed data. However, if transcriptions are only partially supervised, then (hopefully minor) recognition er-

*Work supported by the EC (FEDER/FSE) and the Spanish MCE/MICINN under the MIPRCV "Consolider Ingenio 2010" programme (CSD2007-00018), the iTransDoc project (TIN2006-15694-CO2-01) and the FPU scholarship AP2007-02867. Also supported by the UPV grant 20080033.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11 ...$10.00.

rors may go unnoticed to the user and have a negative effect on model adaptation. In this paper, we study this effect as a function of the degree of supervision, on two real handwriting transcription tasks of considerable complexity. We also consider three adaptation (re-training) strategies: from all data, only from high-confidence parts, and only from supervised parts. Re-training from high-confidence parts is inspired in the work of Wessel and Ney [8], in which confidence measures were successfully used to restrict unsupervised learning of acoustic models for large vocabulary continuous speech recognition. In this work, however, high-confidence parts include both, unsupervised words above certain confidence threshold, and supervised words. Also, they are used to re-train both, HMMs and the n-gram language model. On the other hand, in order to simulate user actions at different degrees of supervision, we propose a simple yet realistic user interaction model.

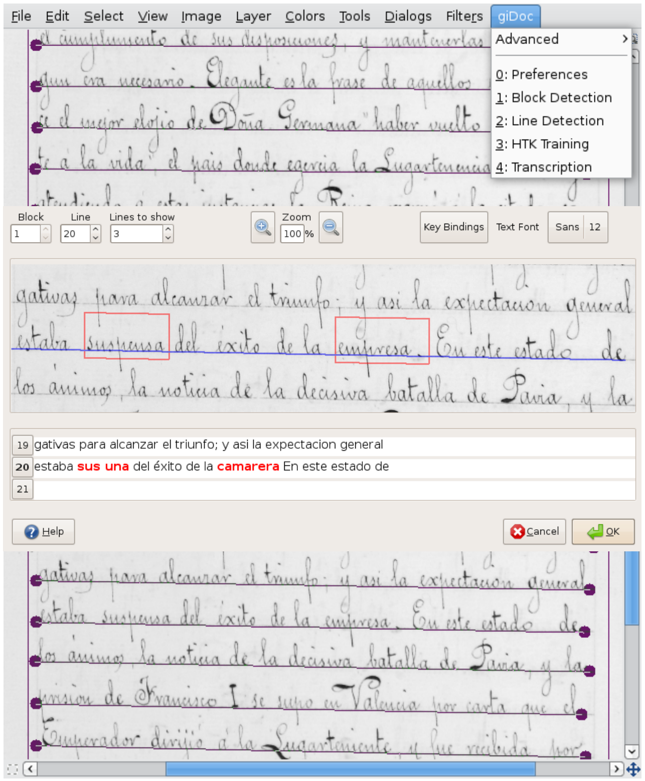
The paper is organized as follows. After a brief overview of GIDOC in Section 2, calculation of confidence measures is exemplified in Section 3, and the proposed user interaction model is described in Section 4. Experiments are reported in Section 5, while concluding remarks are drawn in Section 6.

2. GIDOC OVERVIEW

GIDOC is a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription [3]. It is built as a set of plug-ins for the well-known GNU Image Manipulation Program (GIMP), which has many image processing features already incorporated and, what is more important, a high-end user interface for image manipulation. 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 Figure 1).

As shown in Figure 1, the GIDOC menu includes six entries, though here only the last one, *Transcription,* is briefly described (see [3] for a more detailed description of GIDOC). The *Transcription* entry opens an interactive transcription dialog (also shown in Figure 1), which consists of two main sections: the image section, in the middle 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 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, by entering text and moving the edit cursor with the arrow keys or the mouse. However, the user may choose any order desired.

Note that each editable text box has a button attached to its left, which is labelled with its 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 HMMs and a language model previous trained. As shown in Figure 1, words in the current line for which the system is not highly confident are emphasized (in red) in both the image and transcription sections. It is then up to the user to supervise system output completely, or simply those words emphasized in red. He/she may accept, edit or discard the current system line transcription.



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

Figure 1: Interactive transcription dialog over an image window showing GIDOC menu.

3. CONFIDENCE MEASURES

As indicated in the introduction, confidence measures on recognized words are calculated as posterior word probabilities estimated from word graphs. Generally speaking, word graphs are used to represent, in a compact form, large sets of transcription hypothesis with relatively high probability of being correct. Consider the example in Figure 2, where a small (pruned) word graph is shown aligned with its corresponding text line image and its recognized and true transcriptions.

Each word graph node is aligned with a discrete point in space, and each edge is labelled with a word (above) and its associated posterior probability (below). For instance, in Fig. 2, the word “sus” has a posterior probability of 0.69 to occur between “estaba” and “un”, and 0.03 to occur between “estaba” and “con”. Note that all word posteriors sum to one at each point in space. Therefore, the posterior probability for a word w to occur at a specific point p is given by the sum of all edges labelled with w that are found at p; e.g. “sus” has a posterior probability of 0.72 at any point in which the two edges labelled with “sus” are simultaneously found. The confidence measure of a recognized word is calculated from these point-dependent posteriors, by simply maximizing over all points where it is most likely to occur

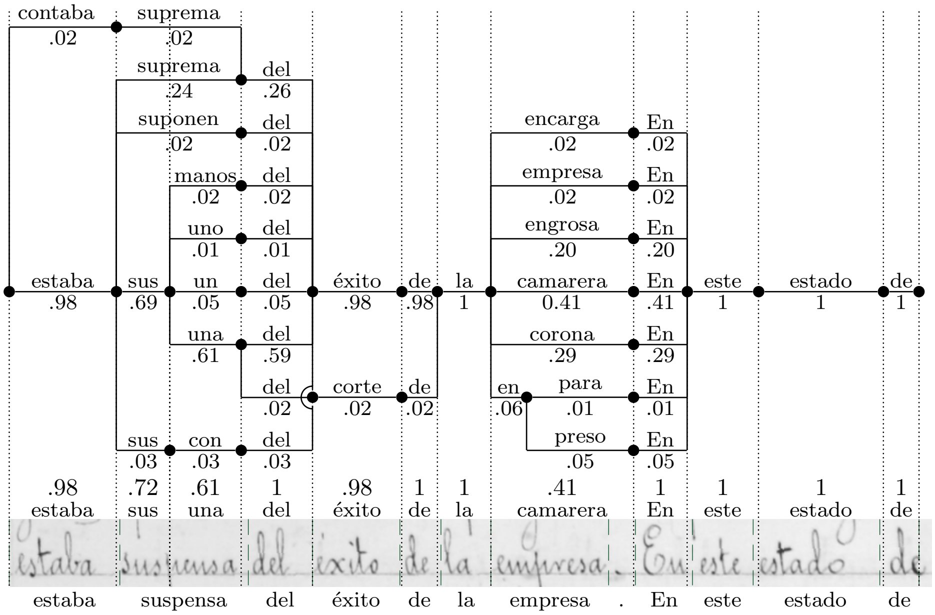


Figure 2: Word graph example aligned with its corresponding text line image and its recognized and true transcriptions. Each recognized word is labelled (above) with its associated confidence measure.

(Viterbi-aligned). As an example, each recognized word in Fig. 2 is labelled (above) with its associated confidence measure. Please see [6] for more details.

4. USER INTERACTION MODEL

As said in the introduction, in this paper we propose a simple yet realistic user interaction model to simulate user actions at different degrees of supervision. The degree of supervision is modelled as the (maximum) number of recognized words (per line) that are supervised: 0 (unsupervised), 1, . . . , ∞ (fully supervised). It is assumed that recognized words are supervised in non-decreasing order of confidence.

In order to predict the user actions associated with each word supervision, we first compute a minimum edit (Levenshtein) distance path between the recognized and true transcriptions of a given text line. For instance, the example text line image in Fig. 2 is also used in Fig. 3 to show an example of minimum edit distance path between its recognized and true transcriptions. As usual, three elementary editing operations are considered: substitution (of a recognized word by a different word), deletion (of a recognized word) and insertion (of a missing word in the recognized transcription). Substitutions and deletions are directly assigned to their corresponding recognized words. In Fig. 3, for instance, there is a substitution assigned to "sus", a deletion assigned to "una", and a second substitution that corresponds to "camarera". Insertions, however, have not direct assignments to recognized words and, hence, it is not straightforward to predict when they are carried out by the user. To this end, we first compute the Viterbi segmentations of the text line image from the true and recognized transcriptions. Given a word to be inserted, it is assigned to the recognized word whose Viterbi segment covers most part of its true Viterbi segment. For instance, in Fig. 3, the period is completely covered by "camarera", and thus its insertion is assumed to be done when "camarera" is supervised.

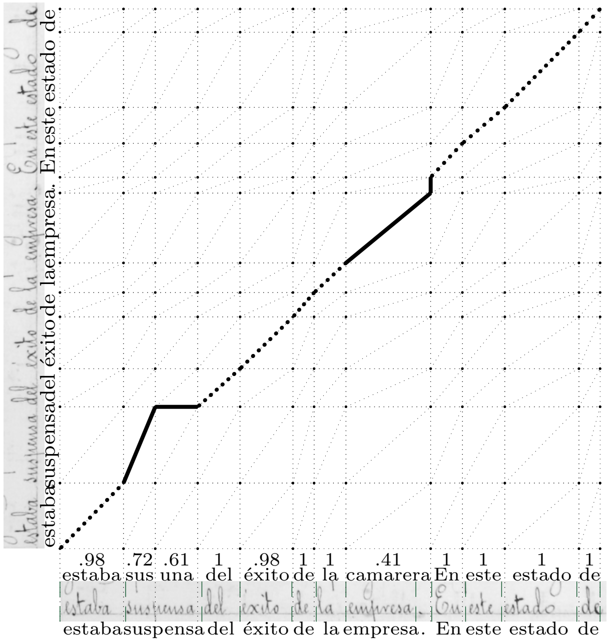

estaba suspensa del éxito de la empresa . En este estado de
.98 .72 .61 1 .98 1 1 .41 1 1 1 1
estaba sus una del éxito de la camarera En este estado de
estaba suspensa del éxito de la empresa . En este estado de

Figure 3: Example of minimum edit distance path between the recognized and true transcriptions of a text line image.

5. 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 [5]. 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). The first two blocks (pp. 3-42) were used to train initial image and language models from fully supervised transcriptions. Then, from block 3 to 8, each new block was recognized, partially supervised and added to the training set built from its preceding blocks. We considered three degrees of supervision: 0 (unsupervised), 1 and 3 supervised words per line. Also, as indicated in the introduction, we considered three adaptation (re-training) strategies: from all data, only from high-confidence parts, and only from supervised parts. The results are shown in Fig. 4 in terms of Word Error Rate (WER) on block 9 (pp. 163-180).

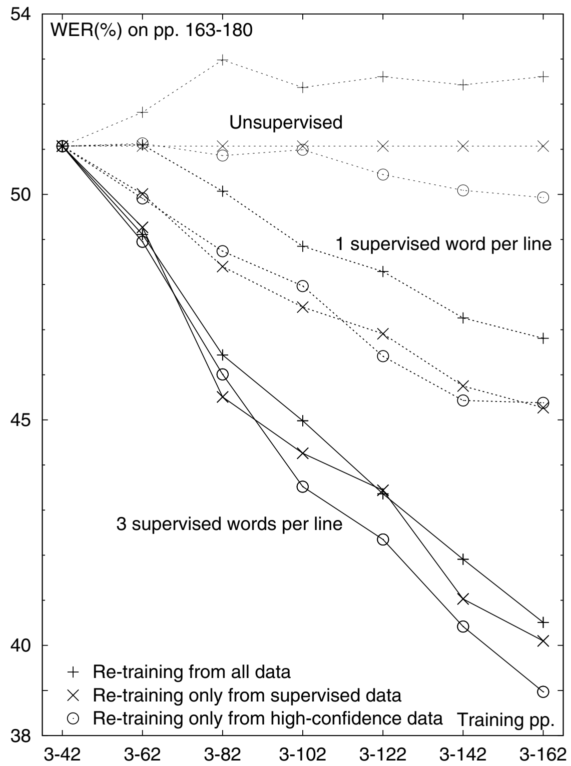


Figure 4: Test-set Word Error Rate (WER) on GERMANA as a function of the training set size (in pages), for varying degrees of supervision (supervised words per line).

From the results in Fig. 4, it becomes clear that baseline models can be improved by adaptation from partially supervised transcriptions, though a certain degree of supervision is required to obtain significant improvements. In particular, supervision of 3 words per line leads to a reduction of more than a 10% of WER with respect to unsupervised learning (baseline models), though there is still room for improvement since full supervision achieves a further reduction of 5% (34%). The adaptation strategy, on the other hand, has a relatively minor effect on the results. Nevertheless, it seems better not to re-train from all data, but only from high-confidence parts, or just simply from supervised parts.

Apart from the above experiment on GERMANA, we did a similar experiment on the well-known IAM dataset, using a standard partition into a training, validation and test sets [1]. The training set was further divided into three subsets; the first one was used to train initial models, while the other two were recognized, partially supervised (4 words per line) and added to the training set. The results obtained in terms of test-set WER are: 42.6%, using only the first subset; 42.8%, after adding the second subset; and 42.0%, using also the third subset. In contrast to GERMANA, there is no significant reduction in terms of WER after adding partially supervised data to the training set. We think that this result is due to the more complex nature of the IAM task.

6. CONCLUDING REMARKS

The adaptation of image and language models from partially supervised data has been studied in the context of computer-assisted handwritten text transcription. A system prototype called GIDOC has been described in which confidence measures estimated from word graphs are used to assist the user in locating system errors. A simple yet realistic user interaction model has been proposed to simulate user actions at different degrees of supervision. Empirical results have been reported on two difficult, real tasks.

7. 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] F. L. Bourgeois and H. Emptoz. DEBORA: Digital AccEss to BOoks of the RenAissance. *IJDAR*, 9:193–221, 2007.

[3] http://prhlt.iti.es/projects/handwritten/idoc/content.php?page=gidoc.php.

[4] L. Likforman-Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *IJDAR*, 9:123–138, 2007.

[5] D. Pérez et al. The GERMANA database. In *Proc. of ICDAR*, pages 301–305, Barcelona (Spain), 2009.

[6] L. Tarazón et al. Confidence Measures for Error Correction in Interactive Transcription of Handwritten Text. In *Proc. of ICIAP*, Vietri sul Mare (Italy), 2009.

[7] A. H. Toselli et al. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *IJPRAI*, 18(4):519–539, 2004.

[8] F. Wessel and H. Ney. Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition. *IEEE Trans. on Speech and Audio Processing*, 13(1):23–31, 2005.