Confidence Measures for Error Correction in Interactive Transcription Handwritten Text*

Lionel Tarazón, Daniel Pérez, Nicolás Serrano, Vicent Alabau,
Oriol Ramos Terrades, Alberto Sanchis, and Alfons Juan

DSIC/ITI, Universitat Politècnica de València
Camí de Vera, s/n, 46022 València, Spain
{lionel,nserrano,dperez,valabau,oriolrt,asanchis,ajuan}@iti.upv.es

Abstract. An 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. In this paper, we focus on a particular system prototype called GIDOC, which can be seen as a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. More specifically, we focus on the handwriting recognition part of GIDOC, for which we propose the use of confidence measures to guide the human supervisor in locating possible system errors and deciding how to proceed. Empirical results are reported on two datasets showing that a word error rate not larger than a 10% can be achieved by only checking the 32% of words that are recognised with less confidence.

Keywords: Computer-assisted Transcription of Handwritten Text, User Interfaces, Confidence Measures.

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 [1,2,3], 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

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

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 567–574, 2009.

transcription task as efficiently as possible. This computer-assisted transcription (CAT) approach has been successfully followed in the DEBORA [4] and iDoc [5] 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 [5].

Here we will focus on the handwriting recognition part of GIDOC. As in the most advanced handwriting recognisers 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 (hopefully minor) system errors. In [6], for instance, a transcription task is considered in which GIDOC achieves around 37% of (test) word error rate (WER) after transcribing 140 document pages out of a total of 764 (18%). Although a WER of 37% is not too bad for effective CAT, it goes without saying that considerable human effort has to be put into *locating* and *editing* systems errors, and this is true for handwritten text transcription tasks in general.

In this paper, we again resort to standard speech technology and, in particular, to *confidence measures (at word level)* [7,8], which are proposed for error (location and) correction in *interactive* handwritten text transcription. Although the use of confidence measures for offline handwritten text line recognition is not new (see [9] and the references therein), here we go a step further and confidence measures are proposed to guide the human supervisor in locating possible system errors and deciding how to proceed. For instance, if a small number of transcription errors can be tolerated for the sake of efficiency, then he/she might validate the system output after only checking those (few) words, if any, for which the system is not highly confident. On the contrary, if at a first glance no significant portion of the text line seems to be correctly recognised, then he/she might ignore system output and transcribe the whole text line manually. On the other hand and by contrast to previous works [9], here confidence measures are based on *posterior word probabilities* estimated from *word graphs* since, at least in the case of speech recognition, experimental evidence clearly shows that they outperform alternative confidence measures, and even posterior word probabilities estimated from *N-best lists* [7,8].

The paper is organised as follows. After a brief overview of GIDOC in Section 2, estimation of posterior word probabilities from word graphs is described in Section 3. Experiments are reported in Section 4, while conclusions and future work are discussed in Section 5.

2 GIDOC Overview

As indicated in the introduction, GIDOC is a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text

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

line detection and handwritten text transcription [5]. 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 [5] for more details on 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 emphasised (in blue colour) 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, it is possible for the user to 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 emphasised (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 emphasised in red. He/she may accept, edit or discard the current line transcription given by the system.

3 Word Posterior Confidence Estimation

In this section we briefly explain the estimation of word-level confidence measures. Taking advantage of the use of standard speech technology by GIDOC, we have adopted a method that has been proved to be very useful for confidence estimation in speech recognition. This method was proposed in [7] and uses posterior word probabilities computed from word graphs as confidence measures.

A word graph G is a directed, acyclic, weighted graph. The nodes correspond to discrete points in space. The edges are triplets $[w, s, e]$, where w is the hypothesized word from node s to node e. The weights are the recognition scores associated to the word graph edges. Any path from the initial to the final node forms a hypothesis $\boldsymbol{f}_1^J$.

Given the observations $\boldsymbol{x}_1^T$, the posterior probability for a specific word (edge) $[w, s, e]$ can be computed by summing up the posterior probabilities of all hypotheses of the word graph containing the edge $[w, s, e]$:

$$P([w,s,e] \mid \boldsymbol{x}_1^T) = \frac{1}{P(\boldsymbol{x}_1^T)} \sum_{\substack{\boldsymbol{f}_1^J \in G : \\ \exists [w',s',e'] : \\ w'=w, s'=s, e'=e}} P(\boldsymbol{f}_1^J,\, [w,s,e],\, \boldsymbol{x}_1^T) \tag{1}$$

The probability of the sequence of observations $P(\boldsymbol{x}_1^T)$ can be computed by summing up the posterior probabilities of all word graph hypothesis:

$$P(\boldsymbol{x}_1^T) = \sum_{\boldsymbol{f}_1^J \in G} P(\boldsymbol{f}_1^J,\, \boldsymbol{x}_1^T)$$

The posterior probability defined in Eq. 1 does not perform well because a word w can occur in slightly different starting and ending points. This effect is represented in the word graph by different word edges and the posterior probability mass of the word is scattered among the different word segmentations (see Fig. 2).

To deal with this problem, we have considered a solution proposed in [7]. Given a specific word (edge) $[w, s, e]$ and a specific point in time $t \in [s, e]$, we compute the posterior probability of the word w at time t by summing up the posterior probabilities of the word graph edges $[w, s', e']$ with identical word w

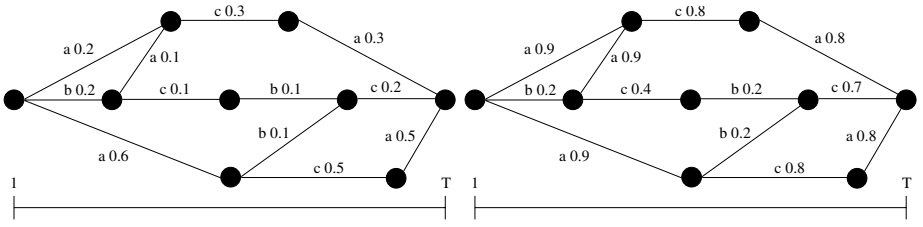

Fig. 2. Word graph with the word posterior probabilities computed as Eq. 1 (left) and as Eq. 3 (right)

and for which t is within the interval time $[s', e']$:

$$P_t([w, s, e] \mid \boldsymbol{x}_1^T) = \sum_{t \in [s', e']} P([w, s', e'] \mid \boldsymbol{x}_1^T) \tag{2}$$

Based on Eq. 2, the posterior probability for a specific word $[w, s, e]$ is computed as the maximum of the frame time posterior probabilities:

$$P([w, s, e] \mid \boldsymbol{x}_1^T) = \max_{s \leq t \leq e} P_t([w, s, e] \mid \boldsymbol{x}_1^T) \tag{3}$$

The probability computed on Eq. 3 is in the interval $[0, 1]$ since, by definition, the sum of the word posterior probabilities for a specific point in time must sum to one (see Fig. 2, left). The posterior probabilities calculated as Eq. 3 are used as word confidence measures (see Fig. 2).

Using these posterior probabilities, a word is proposed to the human supervisor if $P([w, s, e] \mid \boldsymbol{x}_1^T)$ is lower that a certain threshold τ (cf. section 4.2).

4 Experiments

4.1 Databases

The IAM-DB 3.0 dataset [10] contains forms of handwritten English text, scanned at a 300dpi resolution and saved as PNG images with 256 gray levels. Feature extraction has been performed using the geometric-based method. HMMs have lineal topology composed of 7 states with a mixture of 16 gaussians per state. We have achieved a WER of 35.5% for IAM test corpus.

The GERMANA database is the result of digitising and annotating a 764-page Spanish manuscript written in 1891. Most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. GERMANA is a single-author book on a limited-domain topic. It has typical characteristics of historical documents that make things difficult: spots, writing from the verso appearing on the recto, unusual characters and words, etc. Also, the manuscript includes many notes and appended documents that are written in languages different from Spanish, namely Catalan, French and Latin. The manuscript was carefully

Table 1. Basic statistics of IAM and GERMANA

	IAM 20K Voc.			GERMANA 9.4K Voc.		
	Train	Val.	Test	Train	Val.	Test
Pages	747	116	336	94	36	38
Lines	6.161	920	2.781	2.131	811	811
Run.Words (K)	53.8	8.7	25.4	23.7	9.4	9.1
out-of-voc (%)	–	6.6	6.4	–	17.5	18.6

scanned by experts from the Valencian Library at 300dpi in true colours. The experiments have been performed using only the first 179 pages, which correspond to well structured pages written only in spanish. It is also worth noting that 68% of language model words occur once (singletons), and abbreviations appear in many different ways. Furthermore, 33% of words are incomplete since they are at the beginning or the end of lines. HMMs have lineal topology composed of 6 states with a mixture of 64 gaussians per state. We have achieved 42% of WER on the test set. See [6] for a full description.

4.2 Evaluation Measures

Let us assume that, after the recognition output is obtained, the system produces C correctly recognised words and I incorrectly recognised words. Using confidence measures, only words with confidence below on the decision threshold (see section 3) are suggested to the human supervisor for correction. We can distinguish two outcomes in this interactive paradigm:

- *True Rejection* (TR): words incorrectly recognised are suggested for correction.
- *False Rejection* (FR): words correctly recognised are suggested for correction.

When the human supervisor completes the revision of the suggested corrections, we are interested to evaluate the human effort along with the improvement achieved. For this purpose, we compute the ratio of words supervised by the human and the improvement on the system accuracy as a result of the interactive correction process.

$$Supervised = \frac{TR + FR}{I + C} \qquad Accuracy = \frac{TR + C}{I + C}$$

To provide an adequate overall estimation of these two measures, we need to compute both values for all possible decision threshold τ. This can be easily achieved based on a *Receiver Operating Characteristic* (ROC) curve. ROC curves are typically used to evaluate the performance of confidence measures. A ROC curve represents the *True Rejection Rate* (TRR) against the *False Rejection Rate* (FRR) for all possible values of τ. TRR and FRR are computed as:

$$TRR = \frac{TR}{I} \qquad FRR = \frac{FR}{C}$$

Let (frr,trr) be a point of the ROC curve, we can compute the supervision and accuracy measures for this decision threshold, as:

$$Supervised(frr, trr) = \frac{trr \cdot I + frr \cdot C}{I + C} \qquad Accuracy(trr) = \frac{trr \cdot I + C}{I + C}$$

Computing the Accuracy and Supervision as a function of the ROC curve allows to evaluate the impact of confidence measures over the trade-off accuracy-effort.

4.3 Results

The proposed approach has been tested using GIDOC toolkit along with the IAM and GERMANA corpora (described in Sec. 4.1).

For both corpus, a bigram language model and character-level HMMs have been obtained using the training set. Upper and lower case words were distinguished and punctuation marks were modelled as separate words. The validation set has been used to adjust the Grammar Scale Factor (GSF) and Word Insertion Penalty (WIP) recognition parameters. For confidence estimation, a parameter to scale the language model probabilities has been also optimized using the validation set. This scaling has an important impact on the performance of word posterior probabilities as confidence measures [7]. The optimized parameters have been used in the test phase.

The improvements on the transcription accuracy as a function of the ROC curve are shown in Figure 3. We have emphasised the supervision needed to achieve 80%, 90% and 95% of transcription accuracy.

The transcription accuracy baseline (without supervision) for the IAM corpus is about 69%. Confidence estimation allows us to improve it up to an 80% by supervising only 15% of recognised words. This figure increases to a nearly optimal 99% by supervising 69% of recognised words. In absolute terms, this implies a saving of $7k$ words to be supervised. Another view is that, when a small

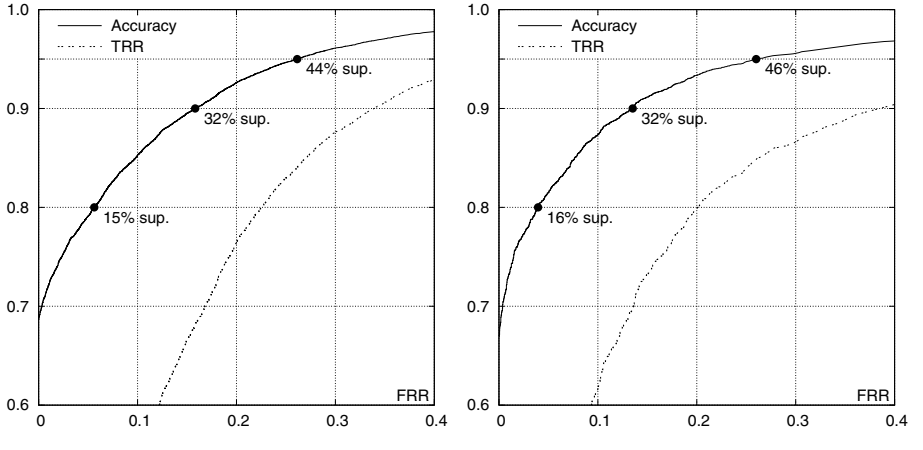

Fig. 3. ROC curve and Accuracy on IAM (left) and Germana (right) databases

number of transcription errors can be tolerated for the sake of efficiency, the use of confidence measures can help to reduce drastically the supervision effort. For the IAM, a 97% of accuracy is achieved by supervising half of the words.

Similar results have been obtained on the GERMANA corpus. The accuracy baseline (67%) is improved to an 80% by supervising only 16% of recognised words. Also, an accuracy of 96% is achieved by supervising half of the words.

5 Conclusions and Future Work

We have presented confidence estimation to reduce the supervision effort in interactive transcription of handwritten text. Posterior probabilities computed from word graphs have been used as confidence measures. The approach proposed have been tested using the GIDOC toolkit along with the IAM and GERMANA databases. We have shown how the use of confidence measures can help to reduce drastically the supervision effort improving the transcription accuracy. Experimental results show that the transcription accuracy can be higher than 95% while the user effort is reduced to the half. Future work should be explore new ways of using confidence measures in the interactive paradigm. Different criteria can be used to suggest for validation the words that are likely to be recognition errors. We plan to study the impact of these strategies over the supervisor effort.

References

1. Toselli, A.H., Juan, A., Keysers, D., et al.: Integrated handwriting recognition and interpretation using finite-state models. IJPRAI 18(4), 519–539 (2004)
2. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. IJDAR 9, 123–138 (2007)
3. Bertolami, R., Bunke, H.: Hidden markov model-based ensemble methods for offline handwritten text line recognition. Patter Recog. 41, 3452–3460 (2008)
4. Bourgeois, F.L., Emptoz, H.: DEBORA: Digital AccEss to BOoks of the RenAissance. IJDAR 9, 193–221 (2007)
5. Juan, A., et al.: iDoc research project (2009), `http://prhlt.iti.es/projects/handwritten/idoc/content.php?page=idoc.php`
6. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos, O., Juan, A.: The GERMANA database. In: Proc. of ICDAR 2009 (2009)
7. Wessel, F., Schlüter, R., Macherey, K., Ney, H.: Conf. measures for large vocabulary speech recognition. IEEE Trans. on Speech and Audio Proc. 9(3), 288–298 (2001)
8. Sanchis, A.: Estimación y aplicación de medidas de confianza en reconocimiento automático del habla. PhD thesis, Univ. Politécnica de Valencia, Spain (2004)
9. Bertolami, R., Zimmermann, M., Bunke, H.: Rejection strategies for offline handwritten text recognition. Pattern Recognition Letter 27, 2005–2012 (2006)
10. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. IJDAR, 39–46 (2002)