Document downloaded from:

http://hdl.handle.net/10251/50537

This paper must be cited as:

Serrano Martinez Santos, N.; Giménez Pastor, A.; Civera Saiz, J.; Sanchis Navarro, JA.; Juan Císcar, A. (2014). Interactive handwriting recognition with limited user effort. International Journal on Document Analysis and Recognition. 17(1):47-59. doi:10.1007/s10032-013-0204-5.

The final publication is available at

http://link.springer.com/article/10.1007%2Fs10032-013-0204-5

Copyright Springer Verlag (Germany)

**International Journal on Document Analysis and Recognition manuscript No.**
(will be inserted by the editor)

# Interactive Handwriting Recognition with Limited User Effort

**Nicolás Serrano · Adrià Giménez · Jorge Civera · Alberto Sanchis · Alfons Juan**

Received: date / Accepted: date

**Abstract** Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. Although post-editing automatic recognition of handwritten text is feasible, it is not clearly better than simply ignoring it and transcribing the document from scratch. A more effective approach is to follow an interactive approach 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. Nevertheless, in some applications, the user effort available to transcribe documents is limited and fully supervision of the system output is not realistic. To circumvent these problems, we propose a novel interactive approach which efficiently employs user effort to transcribe a document by improving three different aspects. Firstly, the system employs a limited amount of effort to solely supervise recognised words that are likely to be incorrect. Thus, user effort is efficiently focused on the supervision of words for which the system is not confident enough. Secondly, it refines the initial transcription provided to the user by recomputing it constrained to user supervisions. In this way, incorrect words in unsupervised parts can be automatically amended without user supervision. Finally, it improves the underlying system models by retraining the system from partially supervised transcriptions. In order to prove these statements, empirical results are presented on two real databases showing that the proposed approach can notably reduce user effort in the transcription of handwritten text in (old) documents.

**Keywords** Handwriting Recognition · Computer-assisted Text Transcription · Active Learning · Semi-supervised Learning · Confidence measures · constrained Viterbi search

N. Serrano, A. Giménez, J. Civera, A. Sanchis, A. Juan
DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
E-mail: {nserrano,agimenez,jcivera,josanna,ajuan}@dsic.upv.es

## 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 automatically transcribing all document images off-line, and then manually supervising system transcriptions to edit incorrect parts. However, state-of-the-art technologies for Handwritten Text Recognition (HTR) are still far from perfect both in, unconstrained domains [4, 10, 13, 28], and in old text documents [8]. Thus, post-editing machine-generated output is not clearly better than simply ignoring it and transcribing the document from scratch.

To circumvent this problem, HTR systems can be used within a Computer Assisted Transcription (CAT) framework, 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. In CAT systems, the main aim is to employ user effort efficiently since it is expensive and limited.

In this work, we describe a novel CAT approach to transcribe (old) text documents in which user effort is considered to be limited. The aim is to build a system, which employs the limited user effort to generate the best possible transcriptions as efficiently as possible. The system employs the limited effort by supervising only hypothesised words that are likely to be misrecognised [27]. Thus, limited user effort is efficiently focused only on the supervision of the output parts for which the system is not confident enough. Low confidence words are presented to the user in isolated boxes, in a similar way as in [2], focusing user attention and preventing them from wasting effort in reading their context. Once user supervisions are performed, the system recomputes the transcription subjected to user supervised words by means of a constrained-Viterbi search [23]. In this way, output errors in the unsupervised parts can be automatically amended without user supervision. At the end of the process, partially supervised transcriptions are used to improve the current system performance by means of adaptation techniques [21]. These techniques improve the underlying system models by retraining from correctly transcribed words and high confidence parts within the transcriptions.

This paper provides a comprehensive description of models and techniques that have been studied and reported in separate works [21, 23, 24, 27]. However, significantly improved baseline experimental results are reported for the first time in this work as a result of updating our feature extraction algorithm. Also, as a novelty, we further extend the constrained Viterbi-based search [23] to deal with the case of word deletions. Moreover, a new experimental study with a complete analysis of diverse aspects of the proposed approach is carried out on two HTR databases: GERMANA [17] and RODRIGO [22].

The remainder of the paper is organised as follows. Firstly, a brief review of related work about automatic transcription of handwritten text documents is detailed in Section 2. Secondly, Section 3 presents a detailed explanation of our interactive HTR approach. Finally, Section 4 is devoted to report empirical results whereas conclusions drawn and future work are summarised in Section 5.

## 2 Related Work

State-of-art HTR systems cannot guarantee fully-automatic high-quality transcription of handwritten text documents [10]. However, they can be integrated in a computer-assisted application to boost transcriptor performance, as it was successfully achieved in OCR recognition systems [2, 14].

State-of-art HTR systems are grounded on the statistical framework [18]. This framework also constitutes a successful approach for CAT in HTR [21, 23, 29]. Traditionally, as stated in [5], the task of HTR can be introduced from a statistical point of view as follows. Given a sequence of feature vectors $\mathbf{x} = x_1, \cdots, x_T = x_1^T$ representing a text line image, the recognition task can be understood as the search for the sequence of words $\mathbf{w}$ that maximises the posterior probability:

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{x}) = \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w}) \tag{1}$$

where $p(\mathbf{x} \mid \mathbf{w})$ corresponds to the image models and $p(\mathbf{w})$ corresponds to the language model. On the one hand, $p(\mathbf{x} \mid \mathbf{w})$ is the probability of a sequence of words $\mathbf{w}$ to correspond to a text line image $\mathbf{x}$. This probability is typically modelled using Hidden Markov Models (HMMs). On the other hand, $p(\mathbf{w})$ is the probability of a sentence $\mathbf{w}$ and it is usually modelled using a smoothed $n$-gram language model. This technology is commonly adopted in current state-of-the-art HTR systems [18]. However, even the best systems do not produce an acceptable automatic transcription of these documents [10], and although post-editing is possible, it may not be better than to manually transcribe documents. Alternatively, it is more effective to interactively transcribe the document with the aid of a CAT system.

### 2.1 Computer Assisted Transcription of Text Images

A first approach of CAT of text images was proposed in [29] following previous ideas applied to machine translation and speech recognition [3, 20]. In this work the authors proposed a prefix-based interactive-predictive approach in which the user reads from left to right both, the system output, and its corresponding text image, correcting the first incorrect word. Then, a valid prefix $\mathbf{p}$ is defined including all words up to the one corrected. Next, the system recomputes its hypothesis constrained to this (fully supervised) prefix, which may improve the unsupervised words. This process continues until all words have been supervised.

This supervision protocol updates the current hypothesis by searching for the most probable suffix $\hat{\mathbf{s}}$ that better completes the validated prefix $\mathbf{p}$. This is achieved by conveniently introducing the prefix dependency on Eq. (1)

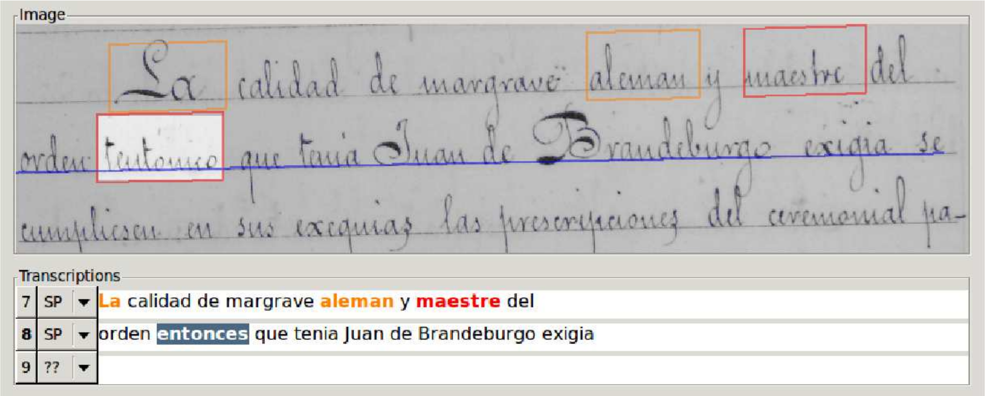
$$\hat{\mathbf{s}} = \operatorname*{argmax}_{\mathbf{s}} p(\mathbf{s} \mid \mathbf{x}, \mathbf{p}) = \operatorname*{argmax}_{\mathbf{s}} p(\mathbf{x} \mid \mathbf{s}, \mathbf{p})\, p(\mathbf{s} \mid \mathbf{p}) \tag{2}$$

In order to perform this search, the sequence of feature vectors is split into two fragments $x_1^b$ and $x_{b+1}^T$, which depends only on $\mathbf{p}$ and $\mathbf{s}$, respectively. The boundary $b$ is unknown, and considered a hidden variable, the estimation of which is approximated in the search process

$$\begin{aligned} \hat{\mathbf{s}} &\approx \operatorname*{argmax}_{\mathbf{s}} \sum_{1 \leq b \leq T} p(x_1^b \mid \mathbf{p})\, p(x_{b+1}^T \mid \mathbf{s})\, p(\mathbf{s} \mid \mathbf{p}) \\ &\approx \operatorname*{argmax}_{\mathbf{s}} \max_{b}\ p(x_1^b \mid \mathbf{p})\, p(x_{b+1}^T \mid \mathbf{s})\, p(\mathbf{s} \mid \mathbf{p}) \end{aligned} \tag{3}$$

This two-step interactive-predictive search defined in Eq. (3) is repeated until the transcription has been completely validated. As a result, error-free transcriptions are obtained.

However, the prefix-based approach presents three main limitations. Firstly, the user needs to supervise all recognised words. Thus, this approach is not applicable when user effort is limited. There are many applications in which user effort is limited or expensive. For instance, some applications need to build competent systems from scarce annotated data [9, 11, 21] in order to be used as soon as possible. Alternatively, in other applications complete annotation of documents is not required to convey the meaning, or to be used as source for other application, such as information retrieval [12]. Secondly, supervision must be performed from left to right, and an important user effort has to be devoted to locate output errors. Thirdly, underlying models remain the



| | | |
|---|---|---|
| 7 | SP | La calidad de margrave aleman y maestre del |
| 8 | SP | orden entonces que tenia Juan de Brandeburgo exigia |
| 9 | ?? | |

**Fig. 1** Interactive transcription of the word "entonces" in GIDOC. Its corresponding reference word "teutonico" is highlighted darkening the rest.

same over the whole transcription process, not taking advantage of the new data that becomes available through the interactive process.

These three limitations were overcome in previous work presented by the authors [21, 23, 24, 27]. In our approach, we deal with the interactive transcription of handwritten text documents when user effort is limited. The system's first objective is to wisely select which words' supervision most improves the system and the resulting transcriptions. Next, supervised words may help to further improve the current transcription as their supervision reduces the uncertainty. Finally, the system is adapted from the partially supervised transcription produced. In the following section, we summarise our previous work to provide a complete picture of our research lines and highlight the contributions of the current work.

## 3 The Interactive Transcription Approach

As mentioned, we deal with the interactive transcription of (old) text documents in which user effort is limited. In our proposed approach, user effort is employed in supervising low confidence hypothesised words. For the sake of clarity, we detail the supervision of a recognised word from the user point of view. Figure 1 shows the transcription dialog of GIDOC [25], which is a set of tools that implements the proposed interactive transcription approach.

In this figure, it can be observed that a text line image, whose baseline is underlined in blue, has been automatically recognised and the result is presented in line number eight. In this moment, the system asks the user for supervision of a recognised word, which may be possibly incorrect. The word to be supervised is highlighted both in the image by darkening all but the corresponding word, and in the editable line by selecting it. It must be noted that word highlighting helps to focus user attention and prevents him from reading the context whenever unnecessary, saving user effort. In this case, the recognised word to be supervised is "entonces" instead of the correct "teutonico", which can be corrected without looking at the context. The user will simply input the correct word and move to the next supervision.

It must be noted that the snapshot shown in Figure 1 is a simple user supervision. More complex supervisions involving incorrectly segmented words in the image are also common and will be analysed in Section 3.2.

### 3.1 Confidence measures

In order to ensure error-free transcriptions, the user needs to supervise the whole transcription. However, in tasks in which the system error rate is acceptable, only few words are incorrectly recognised. A more effective interaction is to ask the user to supervise only those words about which the system is less confident. To this purpose, active learning techniques can be used [26]. In our approach, we adopt a strategy based on the use of confidence measures [19, 31] in order to select which words should be supervised [27].

Word-level confidence measures are calculated as word posterior probabilities estimated from word graphs. Generally speaking, word graphs are used to represent, in a compact form, large sets of transcription hypotheses with relatively high probability of being correct. Consider the example in Fig. 2, where a small (pruned) word graph is aligned with its corresponding text line image. In this figure, recognised and true transcriptions are shown above and below the image, respectively. 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 the word graph of Fig. 2, the word "sus" has a posterior probability of 0.69 of ocurring between "estaba" and "un", and 0.03 of occurring between "estaba" and "con". Note that all word posteriors add up 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 within its recog-

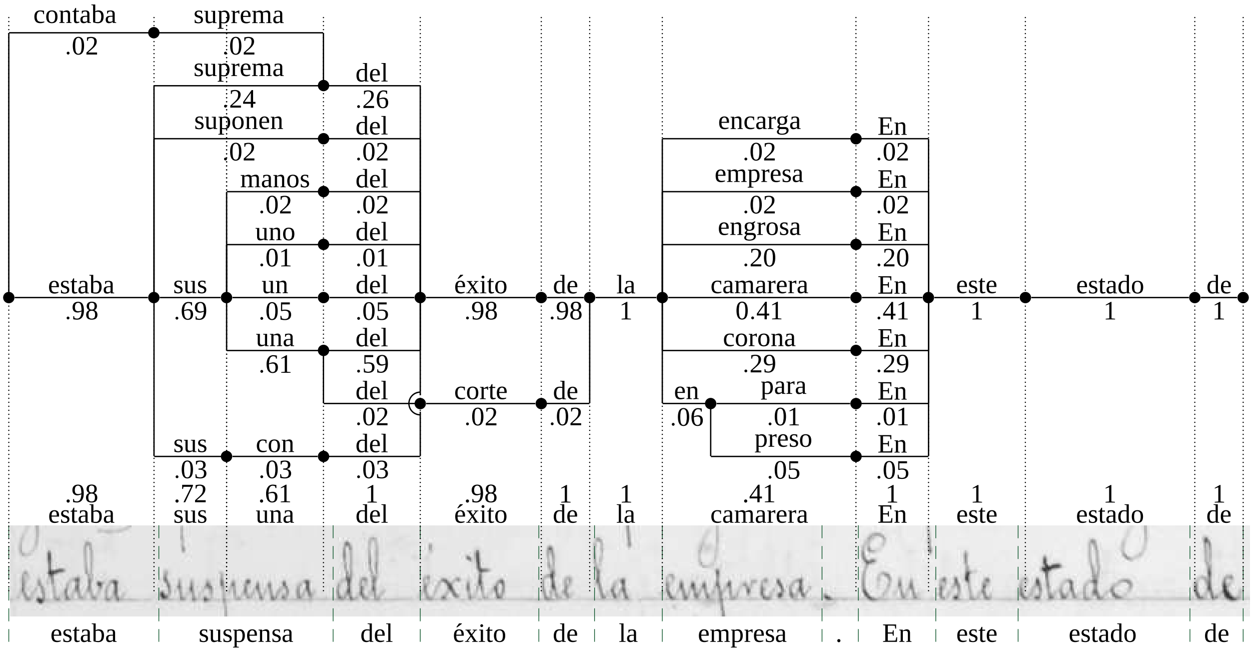


**Fig. 2** Word graph example aligned with its corresponding text line image and its recognised and true transcriptions. Each recognised word is labelled (above) with its associated confidence measure.

nition boundaries since the two edges labelled with “sus” occurs simultaneously. The word-level confidence measure is calculated from these point-dependent posteriors by simply computing the maximum posterior probability over all points within the word recognition boundaries (Viterbi-aligned). As an example, each recognised word in Fig. 2 is labelled (above) with its associated confidence measure.

Confidence measures were tested in [27] on two real handwritten databases being GERMANA one of both. In this work, confidence measures are employed to automatically detect words to be supervised by a fictitious user knowing the reference transcription. A predefined percentage of words are supervised in increasing confidence order and error is computed after supervision. Experimental results shown that the use of confidence measures can help to reduce drastically the supervision effort improving the transcription accuracy. The interested reader is referred to [27] for more details.

## 3.2 User supervision

Confidence measures help the system to select actively which words need supervision. However, supervision of recognised words is not a straightforward process.

As it has been presented in Fig. 1, when the system asks the user to supervise a recognised word, the text line image segment corresponding to this word is presented to the user. But, it might be the case that image segmentation and recognised word alignment are not perfect. For this reason, we need to consider the following four supervision cases:

1) The text line image segment contains a word that has been correctly recognised.
2) The text line image segment contains a word that has been incorrectly recognised.
3) The text line image segment contains more than one word.
4) The text line image segment corresponds to a portion of a word.

The first two cases simply ask the user to supervise the content of a correctly segmented word, which corresponds to the case detailed in Fig. 1. In this situation, the user simply amends or accept the recognised word depending whether it has been misrecognised or not. An example of the third case is shown in Fig. 2, where the supervision of the recognised word “camarera” would result in two user edition operations: the substitution of this word by “empresa” and the insertion of “.”. Lastly, an example of the fourth case occurs when supervising the word “una” in the same figure. In this case, the image segment cannot be correctly identified as a single word, and consequently, the user would delete the current hypothesised word “una”. Later on, if the user is asked to supervise the preceding or next image segment corresponding to a previously deleted word, such as “sus” in the figure, the system would show to the user the image segment associated with the word “sus” plus the deleted word “una”, as they could correspond to a whole word “suspensa”.

### 3.3 Constrained Viterbi-based search

As we already pointed out, the easiest way to improve the system transcription is to simply ask the user to supervise some (hopefully misrecognised) words. This simple strategy will be referred to from here on as *conventional*, and considered to be the interactive baseline system with respect to the other interactive approaches. However, user supervisions can be used to further improve the transcription beyond basic correcting. Following this idea, we proposed an extension to the conventional approach, in which given the supervision of an image segment, the system recomputes a new transcription subject to user supervisions [23]. As it has been said, this approach has also been followed by Toselli et al [29], but as observed in Eq. 2, it is constrained to a left-to-right supervision protocol. On the contrary, in our approach any word can be supervised independently from their context. This is due to the migration from lattice-based search [29] to constrained Viterbi-based search [12]. The constrained Viterbi-search allows for the definition of words that must be necessarily recognised for a given image segment during the search process. These words narrow the expansion of the search trellis at them, reducing the amount of hypothesis that are explored.

In [23], the user performs the supervision according to the first three supervision cases previously described. As a result, the user defines a constraint $\mathbf{c} = (c_1, c_2, c_3)$ by which a word $c_3$ must be recognised from segment $x_{c_1}^{c_2}$ of the text line image. This constraint can be included in the general search problem (Eq. 1) as follows:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{c}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{x} \mid \mathbf{w}, \mathbf{c})\, p(\mathbf{w}) \tag{4}$$

where the language model $p(\mathbf{w})$ is assumed to be independent of the user constraint $\mathbf{c}$. At this point, it is convenient to split the image model in accordance with $\mathbf{c}$:

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{c}) = p(x_1^{c_1-1} \mid w_1^{s-1})\, p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})\, p(x_{c_2+1}^{T} \mid w_{s+1}^{|w|}) \tag{5}$$

where $p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})$ is the only part of the image model in which the constraint $\mathbf{c} = (c_1, c_2, c_3)$ takes effect. As $c_3$ is the only word that can be recognised from the image segment $x_{c_1}^{c_2}$, $p(x_{c_1}^{c_2} \mid w_s, \mathbf{c})$ will be computed as:

$$p(x_{c_1}^{c_2} \mid w_s, \mathbf{c}) = \begin{cases} p(x_{c_1}^{c_2} \mid w_s) & c_3 = w_s \qquad (6) \\ 0 & c_3 \neq w_s \qquad (7) \end{cases}$$

for each hypothesis $\mathbf{w}$ and any position $s$ in which $w_s$ is to be considered as the word written by hand in the image segment $x_{c_1}^{c_2}$. On the other hand, the image models for the prefix and suffix, $p(x_1^{c_1-1} \mid w_1^{s-1})$ and $p(x_{c_2+1}^{T} \mid w_{s+1}^{|w|})$, are assumed to only depend on the given word sequences.

As a novelty, we further extend in this work the approach presented in [23] to include the supervision of words that need to be deleted, i.e. the fourth supervision case described above (e.g. deletion of “sus” or “una” in Fig. 2). Now, the user defines a constraint $\mathbf{c} = (c_1, c_2, \bar{c}_3)$ by which word $c_3$ should not appear in any segment $\left(x_{k_1}^{k_2}\right)$, totally or partially, within segment $x_{c_1}^{c_2}$. Formally, Eqs. 5-7 can be extended to include the four supervision cases as follows:

$$p(\mathbf{x} \mid \mathbf{w}, \mathbf{c}) = \max_{0<k_1<k_2<T+1} p(x_1^{k_1-1}, x_{k_1}^{k_2}, x_{k_2+1}^{T} \mid \mathbf{w}, \mathbf{c}) \tag{8}$$

where

$$p(x_1^{k_1-1}, x_{k_1}^{k_2}, x_{k_2+1}^{T} \mid \mathbf{w}, \mathbf{c}) = p(x_1^{k_1-1} \mid w_1^{s-1})\, p(x_{k_1}^{k_2} \mid w_s, \mathbf{c})\, p(x_{k_2+1}^{T} \mid w_{s+1}^{|w|}) \tag{9}$$

with

$$p(x_{k_1}^{k_2} \mid w_s, \mathbf{c}) = \begin{cases} p(x_{k_1}^{k_2} \mid w_s) & {\scriptstyle [k_1,k_2]=[c_1,c_2] \atop c_3=w_s} \qquad (10) \\ 0 & {\scriptstyle [k_1,k_2]=[c_1,c_2] \atop c_3\neq w_s} \qquad (11) \\ 0 & {\scriptstyle [k_1,k_2]\cap[c_{\underline{1},}c_2]\neq\emptyset \atop c_3=w_s} \qquad (12) \\ p(x_{k_1}^{k_2} \mid w_s) & \text{otherwise} \qquad (13) \end{cases}$$

Note that Eq. 8 reduces to Eq. 5 when $[k_1, k_2] = [c_1, c_2]$ and, in this case, Eqs. 10-11 equal to Eqs. 6-7. The new deletion case is covered in Eqs. 12 and 13.

As explained above, constrained search generates a new hypothesis subject to user supervisions. However, as the user may ask for more than one supervision per text line image, the system could consider at least two alternative strategies regarding when a new hypothesis is recomputed . The first strategy, known as *delayed*, consists in recomputing the most probable hypothesis after all supervisions are done. To put it formally, let us assume that $M$ constraints $\{\mathbf{c}^{(m)}\}$ ($m = 1, \ldots, M$) must be satisfied for each hypothesis $\mathbf{w}$ and positions $\{s^{(m)}\}$ (with $s^{(1)} < \cdots < s^{(M)}$) in which their corresponding words $w_s^{(m)}$ are considered to be written by hand in segments $\{(k_1^{(m)}, k_2^{(m)})\}$ (with $0 < k_1^{(1)} < k_2^{(1)} < \cdots < k_2^{(M)} < T + 1$). Then, our single-constraint model in Eq, 8 can be extended to multiple constraints as follows:

$$p(\mathbf{x} \mid \mathbf{w}, \{\mathbf{c}^{(m)}\}) = \max_{\{(k_1^{(m)}, k_2^{(m)})\}} p(x_1^{{k_1}^{(1)}-1} \mid w_1^{s^{(1)}-1})\, p(x_{{k_1}^{(1)}}^{T} \mid w_{s^{(1)}}^{|w|}, \{\mathbf{c}^{(m)}\}) \tag{14}$$

with

$$p(x_{{k_1}^{(1)}}^{T} \mid w_{s^{(1)}}^{|w|}, \{\mathbf{c}^{(m)}\}) = \prod_{m=1}^{M} p(x_{{k_1}^{(m)}}^{{k_2}^{(m)}} \mid w_{s^{(m)}}, \mathbf{c}^{(m)})\, p(x_{{k_2}^{(m)}+1}^{k_1^{(m+1)}-1} \mid w_{s^{(m)}+1}^{s^{(m+1)}-1}) \tag{15}$$

where each constraint-conditioned model $p(x_{{k_1}^{(m)}}^{{k_2}^{(m)}} \mid w_{s^{(m)}}, \mathbf{c}^{(m)})$ is computed as in the single-constraint case (Eqs. 10–13).

In Eq. 15, it is also assumed that $k_1^{(M+1)} - 1 = T$ and $s^{(M+1)} - 1 = |w|$ (corresponding to the final image segment).

The second strategy, referred to as *iterative*, consists in recomputing a new hypothesis after each user supervision is committed. Figure 3 compares the conventional, delayed and iterative strategies regarding the behaviour of the system in a real example from the GERMANA database [17] carrying out three user supervisions.

At the top of Fig. 3, the reference transcription is aligned with the text line image. Below the image, the figure is divided into four sections from top to bottom: Initial, Conventional, Delayed and Iterative. The initial section displays the most probable hypotheses provided by the system before any user supervision is performed. On the other hand, conventional, delayed and iterative sections show the resulting hypotheses once the user has interacted with the system following the corresponding strategy. It should be noted that the iterative section presents the result of three consecutive user supervision, since the system recomputes the most probable hypotheses after each supervision. Most probable hypotheses are displayed as a list of words for each image segment defined by the system. Below the most probable hypothesis, alternative words are shown in grey, which give us an idea of the uncertainty in that segment. Words are accompanied by their corresponding confidence measure value. Additionally, for each best hypothesis, incorrect words are underlined using a wavy line and user supervised words are shown in bold face.

First, the Conventional strategy simply presents the result of supervising the three least confident words from the initial recognition: “ratificacion” and the last occurrences of “la” and “este”. As no hypothesis recomputation is performed, non-supervised segments remain unchanged and they contain the initial errors. Secondly, the Delayed approach presents the final transcription after supervising the same three recognised words and hypothesis recomputation based on constrained search has been carried out. As a result, non-supervised segments are modified to satisfy user supervisions and the remaining errors are automatically corrected. Specifically, the initial incorrectly recognised words “cetro” and “este” are replaced by the correct words “cuatro” and “esto”, respectively. Lastly, the Iterative strategy resulting into three supervision steps is shown. In this case, we can observe that the supervised words are different from previous strategies, since non-supervised words change every time a recomputation hypothesis is carried out. First, the user replaces “la” by “20”, which causes the word previously recognised as “ratificacion” to change to “estimacion”. Then the user corrects “cetro” with “cuatro”, which causes the misrecognised word “este” to be replaced by the correct word “esto”. Finally, the user substitutes the word “este” by “Octubre”. Even though, an error would remain at the end of the supervision based on interactive strategy.

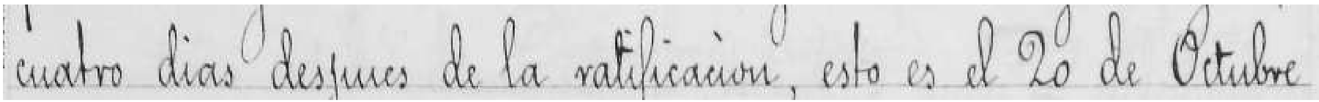
### 3.4 Semi-supervised learning

Up to this point, we have described how to select possibly incorrect recognised words, supervise them, and use this supervision to improve the system hypothesis. At the end of this procedure, the system returns a final transcription constituted by supervised and unsupervised words. In Fig. 3 supervised words were marked in bold face. The supervised words, which have been annotated by the user, can be extracted and directly added to the training set to improve the underlying system models. In addition, as the least confident words have been supervised, the remaining unsupervised words would correspond to high confidence words of the text line image, and therefore, they could also be used in improving the system. For instance, in Fig. 3, we would like to add the whole sample produced by the Delayed approach, which is completely correct. However, as there are unsupervised words, the system needs to select which words may be correct. So, we resort again to confidence measures to successfully adapt from high confidence unsupervised words by means of semisupervised learning [9].

Nevertheless, supervised and high confidence words may define non-continuous image segments and may not cover the entire text line image. In order to split and extract text line image fragments along with their corresponding words as new training data, we use the forced Viterbi alignment as suggested in [31]. In the end, supervised and high confidence words are incorporated as new fresh training data to improve system performance. Image and language models are retrained incorporating this fresh training data in batch model, although we plan to incorporate on-line training techniques to update models in real time [16]. Being that as it may, we successfully adopted and tested semi-supervised learning techniques in HTR [21], corroborating previous results in the area of speech recognition [11]. It must be noted that, to our knowledge, this is the first work that combines active and semisupervised learning at the word level in HTR.

## 4 Experiments

Experiments have been carried out on two recently compiled datasets: GERMANA [17] and RODRIGO [22]. GERMANA is the result of digitising and annotating a 764-page Spanish manuscript from 1891, in which most pages only contain nearly calligraphy text written on ruled sheets of well-separated lines. The example shown in Fig. 1 corresponds to page 144. GERMANA is solely written in Spanish up to page 180, but then it includes many parts written in languages other than Spanish. RODRIGO is similar to GERMANA both, in size and page layout. However, it comes from a much older manuscript, from 1545, the writing style has clear Gothic influences, and it is completely



| | cuatro | dias | despues | de | la | ratificacion, | esto | es | el | 20 | de | Octubre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Initial | cetro .92<br>nuestro .04<br>datos .02 | dias .92 | despues 1 | de 1 | la 1 | ratificacion .73<br>situación .22<br>ratificación , .04 .04<br>situación , .01 .01 | este .90<br>esta .10 | es .98<br>es .02 | el 1 | la .36<br>de .57<br>a .04<br>no .02<br>rio .01 | de 1 | este .61<br>estaba .23<br>titulo .12<br>esta .03<br>la .01 |
| Conventional | cetro .92<br>nuestro .04<br>datos .02 | dias .92 | despues 1 | de 1 | la 1 | **ratificacion,** 1 | este .90<br>esta .10 | es .98 | el 1 | **20** 1 | de 1 | **Octubre** 1 |
| Delayed | cuatro 1 | dias 1 | despues 1 | de 1 | la 1 | **ratificacion,** 1 | esto .98<br>este .02 | es 1 | el 1 | **20** 1 | de 1 | **Octubre** 1 |
| Iterative | cetro .50<br>cuales .45<br>nuestro .03<br>datos .02 | dias 1 | despues 1 | de 1 | la 1 | estimacion .88<br>ratificación .10<br>situación .03 | este .90<br>esta .10 | es .98<br>en .02 | el 1 | **20** 1 | de 1 | este .66<br>estaba .20<br>titulo .10<br>esta .03<br>la .02 |
| | **cuatro** 1 | dias 1 | despues 1 | de 1 | la 1 | estimacion .89<br>ratificación .11 | esto .98<br>esta .11 | es .98<br>en .03 | el 1 | **20** 1 | de 1 | este .69<br>estaba .19<br>titulo .08<br>esta .03<br>la .02 |
| | **cuatro** 1 | dias 1 | despues 1 | de 1 | la 1 | estimacion .89<br>ratificación .11 | esto .98<br>este .02 | es .98<br>en .03 | el 1 | **20** 1 | de 1 | **Octubre** 1 |

**Fig. 3** Comparative of the conventional, delayed, and iterative strategies when supervising a given recognised sentence. At the top, the reference is aligned with its corresponding text line image. The initial hypothesis is displayed after the image, in which each word is accompanied by its confidence. Misrecognised words are underlined using a wavy line, and alternative hypotheses for each word are shown in grayscale. The most probable hypotheses after user supervision of three words for the presented strategies are shown. The three supervised words are highlighted in bold face.

written in Spanish. Some basic statistics of GERMANA and RODRIGO are provided in Table 1.

Figures in Table 1 reflect that GERMANA is more complex than RODRIGO. The vocabulary size and the number of out-of-vocabulary (OOV) words are larger in GERMANA. OOV words constitute a major source of errors since they represent the percentage of running words in the test set that do not appear in the training set. Moreover, GERMANA also has greater perplexity, which is a clear indication of a more difficult task. This difference between the perplexity of both documents is due to the multilingual nature and document layout variability in GERMANA.

| | GERMANA | RODRIGO |
|---|---|---|
| Pages | 764 | 853 |
| Lines | 20529 | 20357 |
| Running words (K) | 217 | 232 |
| Vocabulary size (K) | 27.1 | 17.3 |
| Out-Of-Vocabulary(%) | 25.7 | 11.9 |
| Perplexity | 274.1 | 177.1 |

**Table 1** Statistics of GERMANA and RODRIGO. Out-of-vocabulary words correspond to the percentage of running words in the test set, which do not appear in the training set. Perplexity can be considered as the average number of words which can follow any word sequence, and has been calculated using a ten-fold validation on the whole document.

We simulated the interactive transcription of these two handwritten text documents. Due to their sequential book structure, the task is to transcribe them from the beginning to the end. Previous works [23, 24] only focused on the Spanish part of the GERMANA database, however here we consider the complete transcription of both the GERMANA and RODRIGO databases. Each database was divided into 7 consecutive blocks of 3200 lines, except for the first block, which only contains 1000 lines, and the last block, which also includes the last remnant of the lines. The experimental setting for each database is as follows. The first block is devoted to train an initial system, and tune the preprocessing and recognition parameters. These optimised parameters remain the same for the rest of experiments. Next, starting from block two to the last block, each new block is recognised and evaluated in terms of Word Error Rate (WER). WER is calculated as the number of edit operations (i.e. insertions, deletions and substitutions) needed to convert the recognised transcription into the reference divided by the number of reference words. Next, the recognised block is processed to select new candidate training segments (if necessary), and lastly, added to the training set. Finally, the system is fully re-trained each time a new block is added to the training set. It must be noted that complete re-training of models cannot be performed in real time since it takes several days in a single core.

In the remainder of the section, first, in Section 4.1, we establish our baseline system comparing two feature extraction algorithms. Then, in Section 4.2, we present a user supervision model to assess our interactive HTR system. Finally, experimental results are reported in Section 4.3.

### 4.1 Baseline experiments

In order to establish a strong baseline system that guarantees high recognition accuracy and hence, improved system usability, two feature extraction algorithms were compared on both databases. Our previous works [21, 23] applied a derivative-based algorithm as feature extraction method, however the PCA window-based algorithm has proved to obtain competitive results in other tasks [6]. As a novelty, our first results using the latter feature extraction algorithm are reported in this work.

Table 2 shows comparative WER results for both feature extraction algorithms on GERMANA and RODRIGO databases. As observed, the PCA window-based clearly supersedes the derivative-based algorithm in terms of recognition performance, since it captures a broader image context to compute each feature vector, and PCA considers the complete database to discard nuisance dimensions. Also, feature vectors obtained by PCA present lower dimensionality than those obtained with the derivative-based algorithm, providing faster training and recognition times. For these reasons, the PCA window-based algorithm is adopted in the rest of experiments presented in this work.

| | GERMANA | RODRIGO |
|---|---|---|
| Derivative-based | 50.7 | 42.7 |
| PCA windowed-based | 40.5 | 28.0 |

**Table 2** Comparative WER results for both feature extraction algorithms, derivative-based and PCA window-based, on the GERMANA and RODRIGO databases.

As reported in Table 2, results on GERMANA are significantly poorer than those on RODRIGO. This is explained by GERMANA multilinguality and great variety of document layouts. A posterior error analysis on GERMANA showed that OOV words were recognised as the concatenation of shorter words separated by blanks, since the blank symbol was always inserted between each pair of words. However, letting the system decide whether a blank should be inserted or not, improved the baseline results by 4.9 points, resulting in 35.6 of WER. This improvement was incorporated into the baseline system for GERMANA. However, this same idea provided worse results in RODRIGO, so it was not considered in its baseline system.

### 4.2 User Interaction Model

In order to evaluate the actual performance of the interactive HTR system proposed, we should carry out an evaluation campaign with real users. However, human evaluation is an expensive and time-consuming task. Alternatively, an automatic evaluation allows us to rapidly assess and compare different interactive strategies at very low cost. To this purpose, a user interaction model is defined to simulate the interaction of a real user with our interactive HTR system.

Here, we consider an interaction model in which the user is asked to supervise *n* recognised words of each image line in increasing confidence order. To this purpose, recognised words are first delimited in the image text line as a byproduct of the Viterbi-based search. Next, image lines are divided into segments which are monotonically aligned to words in the most probable hypothesis. When a word requires supervision, its corresponding image segment is presented in closed widgets to a fictitious user, as in Fig. 1. Then, the user corrects it according to one of the supervision cases described in Section 3.2. Each of these cases implies a different kind of supervision that can be represented by one or more Levenshtein edit operations [15]. For example, the second case corresponds to a substitution , while the third case corresponds to a substitution plus one or more insertions, and finally, the fourth case represents a deletion.

### 4.3 Interactive Experiments

In this section, we study the interactive transcription of GERMANA and RODRIGO. In the experiments, a simulated user interactively transcribes the whole document considering that the amount of effort is limited. At the end of the process, the quality of the resulting transcriptions is evaluated based on WER.

Two alternative interaction protocols have been evaluated. In both protocols, words are supervised in order of confidence from lowest to highest. The difference is that in the first interaction protocol supervision is carried out line-per-line, whereas in the second protocol supervision is performed at block level. Thus, for a given supervision effort of $X\%$, the difference is to supervise $X\%$ of the least confidence words at the line level or $X\%$ of the least confident words at the block level. When supervising line by line, errors are assumed to be uniformly distributed over lines. Obviously, this is an unrealistic assumption, but it is compatible with the order in which documents are usually transcribed.

All the interactive learning strategies described in Section 3.3: conventional (C), iterative (I), and delayed (D) have been evaluated following the line-level interaction protocol. Additionally, only the delayed strategy has also been evaluated following the block-level interaction protocol. We will denote this strategy as delayed block-level (DB). From the user point of view, the iterative strategy fits better in a line-by-line supervision, while a block-level supervision seems more reasonable to be applied for the delayed strategy. In any case, all these interactive strategies have been compared with the non-interactive supervision strategy called supervised (S). In this latter strategy, the supervision effort of $X\%$ is employed in the manual transcription of the first $X\%$ words of the document and the rest of the document is automatically transcribed using models trained from the manual transcriptions.

When evaluating interactive strategies, user effort is initially devoted to fully supervise the first block (the first 1000 lines). This block is used to train and tune the initial system. In the line-level experiments, user efforts of 14%, 22%, 31% and 40% have been considered. These percentages correspond with the supervision of one, two, three or four words per line, respectively. Note that, in both corpora, the average number of words per line is 11. Same values have been used in block-level experiments. In the case of the supervised strategy, the user effort is measured stepwise as the transcription of 2000-line blocks. It must be noted that block size in supervised experiments have been adjusted to simulate similar user efforts to those of the interactive experiments.

For all interactive strategies, each block is automatically transcribed and partially supervised according to each strategy. Once the supervision of one block is finished, supervised and high confidence parts of the resulting transcriptions are added as new training material to built new models to recognise the next block.

Fig. 4 shows the result of the performed experiments for both corpora. The X axis measures the user effort employed, which is calculated as the percentage of reference words that have been supervised. Word supervision is considered under the cases detailed in Sec. 3.2, even when it corresponds to the supervision of a correct word. In the Y axis, the quality of the transcribed document is evaluated in terms of WER.

The second point of the curves, around 56% and 50% of WER for GERMANA and RODRIGO, respectively, corresponds to the first fully-annotated block (1000 lines) used to tune all necessary parameters for interactive strategies. Even though this system was trained from little annotated data, its evaluation provides a glimpse of the task difficulty. Both corpus have a relatively big vocabulary containing a large number of singletons. Since these words appear only once in the whole document, recognition error increases due to these out-of-vocabulary (OOV) words. This effect is greater in GERMANA, where there are six different languages and multiple document layout structures, such as list, letters, and notes.

The objective of the interactive strategies is to produce the best transcriptions with the lowest user effort. This best case would correspond to a curve passing as close to the XY axis as possible. On the other hand, the worst case corresponds to a diagonal line connecting the top left point, which represents a void transcription, with the bottom right point, which represents the manual annotation of the whole document. In this worst case, user effort would be devoted to manually transcribe a part of the document leaving the rest untranscribed. As observed in Fig. 4, all the strategies achieve to reduce user effort over manual transcription, since all curves are below the worst-case diagonal. Indeed, the same transcription quality can be achieved with lesser user effort depending on which interactive strategy is employed.

Regarding comparison between the strategies proposed, all of them present a similar behaviour. Transcription accuracy is directly related to the available user effort. However, this improvement greatly decreases when 20% of the document is supervised. This effect is caused because the initial system is not be able to deal with image character variability and language complexity. Once sufficient training data is supervised, image models are well estimated since they correspond to a unique author with a uniform script. However, the language complexity remains mostly due to OOV words. This latter effect can be directly observed in the supervised approach which improves uniformly as more data is supervised. Despite the fact that correct data improves the system as is added to the training set, the improvement from correct data is limited [11, 21]. However, this improvement

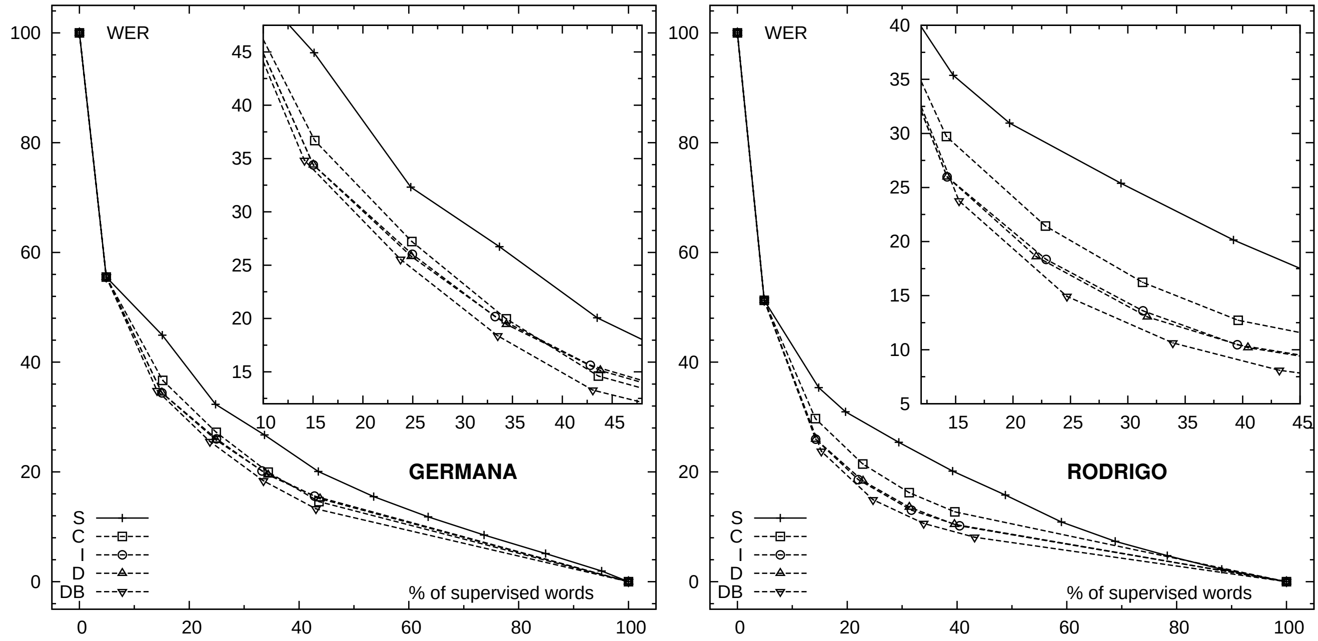


**Fig. 4** WER results from the interactive transcription experiments performed. Word Error Rate (WER) of the final transcriptions is shown for each approach using a limited user effort. A close-up is shown in the upper right corner depicting interactive approaches.

is also true in the case of interactive strategies in which data is added to the training set based on confidence measures.

All interactive transcription strategies outperform the supervised strategy. Indeed, for a similar user effort, there is an important improvement in the transcription quality of 8, and 15 points of WER on average for GERMANA and RODRIGO, respectively. This is mainly caused because user effort is used more efficiently. Interactive strategies employ user effort to supervise likely incorrect words based on confidence measures. Consequently, user corrections directly reduce the error. On the contrary, the supervised approach supervise all words independently of their confidence which is a waste of user effort.

Performance behaviour of line level interactive approaches is slightly different from the supervised approach. There is a greater improvement in the transcription quality when the user supervises one or two words per line, with respect to the case in which three of four supervisions per line are performed. The reason behind this behaviour is an erroneous detection of incorrect words based on confidence measures. Confidence measures correctly identify the first word in need of supervision 80% of the times. However the second word to be supervised is actually incorrect 60% of the times. The explanation of this difference is that, as expected, not all errors are uniformly distributed over lines. Also, small errors, such as one character mismatch, are likely to go unnoticed to the confidence measures.

Fig. 4 zooms the interactive results for each corpus. In RODRIGO, both constrained search strategies, iterative (I) and delayed (D), clearly outperform the conventional (C) in all the experiments. As said, the constrained Viterbi technique, described in Sec. 3.3, recomputes the system hypothesis constrained to user supervisions. This recomputation improves the initial transcription reducing the uncertainty in the search. For example, when only one word is supervised per line, the constrained search improves the results by 5 WER points, decreasing down to 2.5 WER points when four words are supervised. This fact is directly related to the mentioned effect of the confidence measures detecting incorrect words beyond the third and fourth supervised words. On the contrary, in GERMANA, the constrained strategies only outperform the conventional strategy in 5 and 2.5 points of WER when supervising one or two words per line, respectively. A posterior analysis of the results showed that the special treatment of blank symbol described in 4.1 harms the constrained recomputation.

We can also observed that there is no significant difference between the iterative and delayed strategies in both corpora when supervisions are performed on the line level, as corroborated by a bootstrap evaluation [7]. The iterative strategy was expected to be the best one since transcriptions are automatically modified based on user supervisions. However, a detailed analysis showed that the confidence of unsupervised words increases as more words are supervised and, consequently, the system recomputation does not replace them independently of their correctness. The delayed strategy can be considered as the better performance strategy because recomputation cannot be performed in real time. Long waiting times are needed in the interactive approach to recompute hypotheses. Specifically, each recomputation took 30 seconds on average in an Intel i7 with 2.80 GHz.

Regarding comparison between the two different interaction protocols, delayed block-level slightly improved all previous approaches for all user efforts considered. Concretely, results are improved by 1.25 points of WER on average. This is mainly due to a better usage of user effort which is used to supervise more erroneous words than the previous experiments. However, the improvement is not significant in all cases and it would be expected to be higher. For instance, on the second point of GERMANA, which corresponds to a 15% of user effort on average, all approaches that include the constrained-Viterbi recomputation achieved the same result independently of the interaction protocol applied. A deep analysis of the results indicates that a uniform distribution of the error seems adequate when the available quantity of user effort is small. The reason is because the least confidence words in the lines almost correspond to the least confident words in the block. On the contrary, when supervision effort is high, uniform distribution of the error per line is unrealistic and, consequently, the block-level approach is more effective in the aim of supervising the words which are more likely to be incorrect.

In the experiments discussed above user effort has been measured in terms of the percentage of supervised words. This metric has been used for two reasons. Firstly, in order to establish a fair comparison between all the strategies independently from the specific words which are supervised. Note that supervised words can be different depending on the interactive strategy applied. Secondly, the difficulty to assess user effort. Actual supervision cost can only be obtained in a real experiment with real users. This is a very cost and time consuming task and alternative metrics are needed to perform faster evaluation of the techniques. As alternative, we have considered that the percentage of supervised words is a straightforward metric which gives us an acceptable approximation to the actual cost of supervision. However, this metric has the drawback of considering the same cost for the four supervision cases detailed in Sec. 3.2. To circumvent this limitation, we have also used a new metric that compute the percentage of characters typed by a user in the supervision process. As a difference, this metric considers that the equal and deletion operations have a lower edit cost than the other edit operations. Thus, equal and deletion operations only require to type one character whereas in the other supervision cases the cost is the number of characters typed by the user.

Fig. 5 shows the results in terms of percentage of typed characters for the baseline Supervised (S) and the best interactive approach, i.e. Delayed block-level (DB). For comparison purposes, results using percentage of supervised words are also plotted for both approaches in the same figure. As observed, the supervised approach shows the same behaviour because the user effort is employed in completely annotating the first part of the document. On the other hand, the interactive approach shows a reduction of user effort in terms of typed characters when applying a high quantity of user effort. However, the improvement achieved by using a higher user effort decreases faster than in terms of supervised words. This is mainly caused by the previously mentioned problem about the effectiveness of confidence measures. As said, the first words to be supervised are likely to be incorrect and, thus, the user has to type a higher quantity of characters. On the contrary, when more words have been supervised, supervision of correct words increases and a simple key interaction is needed for supervision. As observed in Fig. 5, this effect greatly depended on the recognition performance. In GERMANA, in which there are more errors than in RODRIGO, the percentage of typed characters decreases more slowly.

## 5 Conclusions and future work

In this work, we have described an interactive approach to handwriting text transcription when user effort is limited. The main goal is to efficiently employ the available user supervisions to generate the best transcriptions. Three different interactive transcription strategies have been described and their performance compared with that of a fully supervised baseline system in two real databases. All interactive approaches have improved the baseline supervised approach.

In future work, we plan to improve interactive strategies in different manners. First, active learning techniques other than those used in this work will be considered to further improve system accuracy from user supervisions. Second, we will study how to improve language modelling by applying ideas from our recent work on using external resources [30] and character-based language models [1]. Third, we plan to perform real user evaluations to develop more realistic automatic metrics. Finally, we will consider different, complementary approaches to reduce the time needed to recompute hypotheses and model retraining. In order to speedup hypothesis recomputation, the search space will be reduced to a lattice representation of the most probable hypotheses. On the other hand, to reduce the computational cost of model retraining, online and incremental learning techniques will be tried.

**Acknowledgements** The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures). Also supported by the Spanish Government (MICINN, MITyC, ”Plan E”, under grants MIPRCV ”Consolider Ingenio 2010”, MITTRAL (TIN2009 - 14633-C03-01), erudito.com (TSI-020110-2009-439), iTrans2 (TIN2009-14511), and FPU (AP2007-02867), and the Generalitat Valenciana (grants Prometeo/2009/014 and GV/2010/067).

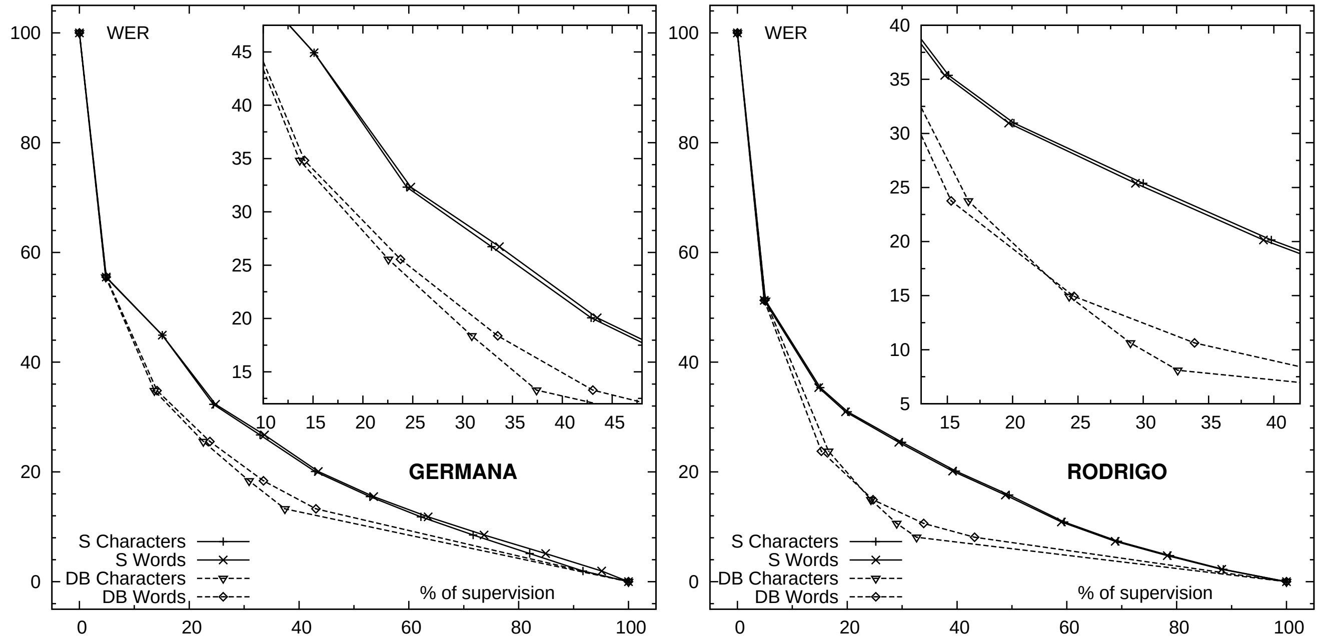


**Fig. 5** WER results from the interactive transcription experiments performed for supervised and the best interactive approaches. Supervision effort is measured in terms of percentage of typed characters and supervised words.

## References

1. Agua M, Serrano N, Civera J, Juan A (2012) Character-based handwritten text recognition of multilingual documents. In: Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012), Madrid (Spain), pp 187–196
2. Ahn LV, Maurer B, Mcmillen C, Abraham D, Blum M (2008) reCAPTCHA: Human-based character recognition via web security measures. Science 321:1465–1468
3. Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lagarda AL, Ney H, Tomás J, Vidal E (2009) Statistical approaches to computer-assisted translation. Computational Linguistics 35(1):3–28
4. Bertolami R, Bunke H (2008) Hidden markov model-based ensemble methods for offline handwritten text line recognition. Pattern Recognition 41:3452–3460
5. Bunke H, Bengio S, Vinciarelli A (2004) Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6):709 – 720
6. Dreuw P, Jonas S, Ney H (2008) White-space models for offline arabic handwriting recognition. In: Proceedings of the 19th International Conference on Pattern Recognition, pp 1–4
7. Efron B, Tibshirani RJ (1994) An Introduction to Bootstrap. Chapman & Hall/CRC
8. Fischer A, Wuthrich M, Liwicki M, Frinken V, Bunke H, Viehhauser G, Stolz M (2009) Automatic transcription of handwritten medieval documents. In: Proceedings of the 15th International Conference on Virtual Systems and Multimedia, pp 137 –142
9. Frinken V, Bunke H (2009) Evaluating retraining rules for semi-supervised learning in neural network based cursive word recognition. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona (Spain), pp 31–35
10. Graves A, Liwicki M, Fernandez S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(5):855 –868
11. Hakkani-Tür D, Riccardi G, Tur G (2006) An active approach to spoken language processing. ACM Transactions on Speech and Language Processing 3:1–31
12. Kristjannson T, Culotta A, Viola P, McCallum A (2004) Interactive information extraction with constrained conditional random fields. In: Proceedings of the 19th Natural Conference on Artificial Intelligence, San Jose, CA (USA), pp 412–418
13. Laurence Likforman-Sulem AZ, Taconet B (2007) Text line segmentation of historical documents: a survey. International Journal on Document Analysis and Recognition 9:123–138
14. Le Bourgeois F, Emptoz H (2007) Debora: Digital access to books of the renaissance. International Journal on Document Analysis and Recognition 9:193–221
15. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics

Doklady 10(8):707–710

16. Neal RM, Hinton GE (1999) Learning in graphical models. MIT Press, Cambridge, MA, USA, chap A view of the EM algorithm that justifies incremental, sparse, and other variants, pp 355–368
17. Pérez D, Tarazón L, Serrano N, Ramos-Terrades O, Juan A (2009) The GERMANA database. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona (Spain), pp 301–305
18. Plötz T, Fink GA (2009) Markov models for offline handwriting recognition: a survey. International Journal of Document Analysis and Recognition 12(4):269–298
19. Quiniou S, Cheriet M, Anquetil E (2012) Error handling approach using characterization and correction steps for handwritten document analysis. International Journal on Document Analysis and Recognition 15(2):125–141
20. Rodríguez L, García-Varea I, Vidal E (2010) Multimodal computer assisted speech transcription. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM, New York, NY, USA, pp 30:1–30:7
21. Serrano N, Pérez D, Sanchis A, Juan A (2009) Adaptation from partially supervised handwritten text transcriptions. In: Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction, Cambridge, MA (USA), pp 289–292
22. Serrano N, Castro F, Juan A (2010) The RODRIGO database. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valleta (Malta), pp 2709–2712
23. Serrano N, Giménez A, Sanchis A, Juan A (2010) Active learning strategies for handwritten text transcription. In: Proceedings of the 12th International Conference on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction, Beijing (China)
24. Serrano N, Sanchis A, Juan A (2010) Balancing error and supervision effort in interactive-predictive handwriting recognition. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, Hong Kong (China), pp 373–376
25. Serrano N, Tarazón L, Pérez D, Ramos-Terrades O, Juan A (2010) The GIDOC prototype. In: Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems, Funchal (Portugal), pp 82–89
26. Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison
27. Tarazón L, Pérez D, Serrano N, Alabau V, Ramos-Terrades O, Sanchis A, Juan A (2009) Confidence measures for error correction in interactive transcription of handwritten text. In: Proceedings of the 15th International Conference on Image Analysis, Processing, Vietri sul Mare (Italy)
28. Toselli A, Juan A, Keysers D, González J, Salvador I, Ney H, Vidal E, Casacuberta F (2004) Integrated handwriting recognition and interpretation using finite-state models. International Journal of Pattern Recognition and Artificial Intelligence 18(4):519–539
29. Toselli A, Romero V, Rodríguez L, Vidal E (2007) Computer assisted transcription of handwritten text. In: Proceedings of the 9th International Conference on Document Analysis and Recognition, Curitiba (Brazil), pp 944–948
30. Valor J, Pérez A, Civera J, Juan A (2012) Integrating a state-of-the-art asr system into the opencast Matterhorn platform. In: Proc. of Advances in Speech and Language Technologies for Iberian Languages (IBERSPEECH 2012), Madrid (Spain), pp 237–246
31. Wessel F, Ney H (2005) Unsupervised training of acoustic models for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 13(1):23 – 31