2009 10th International Conference on Document Analysis and Recognition

Using Mouse Feedback in Computer Assisted Transcription of Handwritten Text Images*

Verónica Romero, Alejandro H. Toselli, Enrique Vidal
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, Spain
{vromero, ahector, evidal}@iti.upv.es

Abstract

To date, automatic handwring recognition systems are far from being perfect and heavy human intervention is often required to check and correct the results of such systems. In order to achieve correct transcriptions, human knowledge can be integrated into the transcription process, following an Interactive Predictive paradigm. We have recently proposed Mouse Actions as a significant feedback information source for the underlying interactive system to improve the poductivity of the human transcriptor. In this paper we review this way to interact with the system and report comparative results using the publicly available IAMDB dataset.

1. Introduction

Many documents used every day include handwritten text and, in many cases, it would be interesting to recognize these text images automatically. However, state-of-the-art handwritten text recognition systems (HTR) can not suppress the need of human work when high quality transcriptions are needed. HTR systems can achieve fairly high accuracy for restricted applications with rather limited vocabulary (reading of postal addresses or bank checks) and/or form-constrained handwriting. However, in the case of unconstrained transcription applications, the current HTR technology typically only achieves results which do not meet the quality requeriments of practical applications. Therefore, once the full recognition process of one document has finished, heavy human expert revision is required to really produce a transcription of standard quality. Such a *post-editing* solution is rather inefficient and uncomfortable for the human corrector.

*Work supported by the EC (FEDER), the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme (CSD2007-00018) and the grant TIN2006-15694-C02-01 by the Universitat Politècnica de València (FPI fellowship 2006-04)

A way of taking advantage of the HTR systems is to combine them with the knowledge of a human transcriptor, constituting the so-called "Computer Assisted Transcription of Text Images" (CATTI) scenario [8, 5]. In this scenario, the system uses the text image and a previously validated part (prefix) of its transcription to propose a suitable continuation. Then the user finds and corrects the next system error, thereby providing a longer prefix which the system uses to suggest a new, hopefully better continuation. The results obtained show that this system can save significant amounts of human effort.

Mouse Actions (MA) can be used as an additional information source: as soon as the user points to the next system error, the system proposes a new, hopefully more correct, continuation. This way, many explicit user corrections are avoided. Preliminary results of this new kind of interaction were reported on tasks such as the transcription of old documents and spontaneous sentences extracted from survey forms [4]. In this paper we review this way to interact with the system using the MA and report comparative results using the publicly available IAMDB dataset.

2. CATTI Framework

In the CATTI framework, the user is directly involved in the transcription process since he/she is responsible of validating and/or correctiong the HTR output. The process starts when the HTR system proposes a full transcription of the input image. Then, the user reads this transcription until finding a mistake; i.e, he or she validates a prefix of the transcription which is error-free. Now, the user enters a word to correct the erroneous text that follows the validated prefix. This generates a new, extended prefix (the previous validated prefix, plus the user amendments), which is used by the HTR system to attempt a new prediction hypothesis, thereby starting a new cycle that is repeated until a final correct transcription is achieved [8, 5].

The traditional handwritten text recognition problem can

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.139

be formulated as the problem of finding a most likely word sequence, $\hat{w}$, for a given handwritten sentence image represented by a feature vector sequence x, that is:

$$\hat{w} = \arg\max_{w} Pr(w|x) = \arg\max_{w} Pr(x|w) \cdot Pr(w) \quad (1)$$

$Pr(x|w)$ is typically approximated by concatenated character Hidden Markov Models (HMMs) [3] and $Pr(w)$ is usually approximated by a n-gram word language model [2].

In the CATTI framework, in addition to x, a user-validated prefix p of the transcription is available. The HTR should try to complete this prefix by searching for a most likely suffix $\hat{s}$ as:

$$\hat{s} = \arg\max_{s} Pr(s|x,p) = \arg\max_{s} Pr(x|s,p) \cdot Pr(s|p) \quad (2)$$

Accordingly, the search must be performed over all possible suffixes s of p and the language model probability $Pr(s|p)$ must account for the words that can be written after p. As in Eq. (1), $Pr(x|s,p)$ can be modelled by HMMs. On the other hand, to implement the language model constraints involved in $Pr(s|p)$, we can take advantage of the information coming from p, as discussed in [8]:

$$Pr(s|p) \simeq \prod_{i=k+1}^{k+n-1} Pr(w_i|w_{i-n+1}^{i-1}) \cdot \prod_{i=k+n}^{l} Pr(w_i|w_{i-n+1}^{i-1}) \quad (3)$$

where the consolidated prefix is $w_1^k = p$ and $w_{k+1}^l = s$ is a possible suffix. The first term of Eq. (3) accounts for the probability of the $n-1$ words of the suffix conditioned by words from the prefix, and the second one is the usual n-gram probability for the rest of the words in the suffix.

2.1. CATTI using word graphs

The search problem corresponding to Eq. 2 and 3 can be solved using search techniques based on *word-graphs*.

A word graph represents the transcriptions with higher $Pr(w|x)$ of the image sentence. In this case, the word graph is just (a pruned version of) the Viterbi search trellis [2] obtained when transcripting the whole image sentence. During the CATTI process the system makes use of this word graph to complete the prefixes accepted by the user.

A word graph can be represented as a weighted directed acyclic graph, where each edge (e) is labeled with a word (w_e) and a score ($score(e)$), and each node (n) is labeled with a point (horizontal position) of the handwritten image (t_n). For each edge, we denote S_e and E_e its start and end node. The graph has a single start node, that points to the start of the text image, and a single end node.

The score of an edge is computed by multiplying the morplhological-lexical probability of the image between its start and end node points $Pr(x_{t_{S_e}}^{t_{E_e}}|w_e)$, by the language model probability of the given word at the edge $Pr(w_e)$. However, in practice, the simple multiplication is modified to balance the absolute values of both probabilities. The most comon modification is to use the *language weight* α and the *insertion penalty* β [1]:

$$score(e) = \log Pr(x_{t_{S_e}}^{t_{E_e}}|w_e) + \alpha \log Pr(w_e) + \beta \quad (4)$$

The word labels of any path from the start node to the end node form a transcription hypothesis, whose score is computed as the sum of the scores of the edges along the path.

As the word graph is a representation of a *subset* of the possible transcriptions for a source handwritten text image, it may happen that some prefixes given by the user can not be exactly found in the word graph. To circumvent this problem some heuristics need to be implemented. In this work, we modified the score associated to each edge in order to cope with the differences between the words in the prefix and the words in the path that best match the given prefix. This heuristic can be implemented as an error-correcting parsing dynamic programming algorithm. Moreover, this algorithm takes advantage of the incremental way in which the user prefix is generated, parsing only the new suffix appended by the user in the last interaction (see [4]).

The computational cost of this approach is much lower than use the naïve Viterbi adaptation we had used in previous works. Therefore, using word-graph techniques the system is able to interact with the human transcriptor in a time efficient way. However, a drawback of this implementation is that some accuracy can be lost.

3. Enriching the CATTI Interaction Process

In CATTI applications the user is repeatedly interacting with the system. Hence, making the interaction process easy is crucial for the success of the system. In conventional CATTI, before typing a new word in order to correct a hypothesis, the user needs to position the cursor in the place where he wants to type the word. This is done by performing a MA (or equivalent pointer-positioning keystrokes). By doing so, the user is already providing some very useful information to the system: he is validating a prefix up to the position where he positioned the cursor, and, in addition, he is signalling that the following word located after the cursor is incorrect. Hence, the system can already capture this fact and directly propose a new suitable suffix in which the first word is different to the first word of the previous suffix. In fig. 1 we can see an example of such behaviour. As in conventional CATTI, the process starts when the HTR system proposes a full transcription $\hat{s}$ of the input image x. Then,

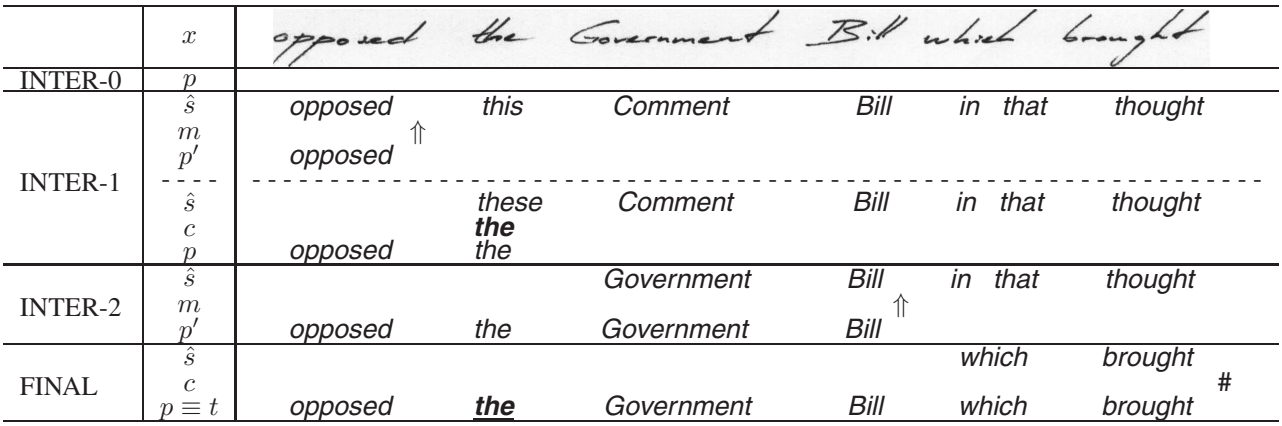

Figure 1. Example of CATTI operation using MA.

the user reads this prediction until a transcription error is found (e) and makes a MA (m) to position the cursor at this point. This way, the user validates an error-free transcription prefix p'. Now, before the user introduces a word to correct the erroneous one, the HTR system, taking into account the new prefix p' and the wrong word that follows p', suggests a suitable continuation (i.e., a new $\hat{s}$). If the new $\hat{s}$ corrects the erroneous word (e) a new cycle starts. However, if the new $\hat{s}$ has an error in the same position that the previous one, the user can enter a word, c, to correct the erroneous text e. This action produces a new prefix p (the previously validated prefix, p', followed by c). Then, the HTR system takes into account the new prefix to suggest a new suffix and a new cycle starts. This process is repeated until a correct transcription of x is accepted by the user. In fig. 1 the underlined boldface word in the final transcription is the only one which was phisically corrected by user. Note that in the iteration 1 a (single) MA does not succeds and the correct word needs to be phisically typed. However, the iteration 2 only needs a MA.

This new kind of interaction needs not be restricted to a single pointer-positioning MA. Several scenarios arise, depending on the number of times the user performs a MA. In the simplest one, the user only makes a MA when it is necessary to displace the cursor (single-MA). In this case the MA does not involve any extra human effort, because it is the same action that the user should make in the conventional CATTI to position the cursor before typing the correct word. Another scenario that can be considered consists in performing a MA systematically before writing, although the cursor is in the correct position. In this case, however, there is a cost associated to this kind of MAs, since the user does need to perform additional actions, which may or may not be beneficial. This scenario can be easily extended allowing to the user to make several MA before having to write a correct word him/herself.

Since we have already dealt, in the section 2, with the problem of finding a suitable suffix $\hat{s}$ when the user validates a prefix p' and introduce a correct word c, we focus now on the problem in which the user only makes a MA. In this case the decoder has to cope with the input image x, the validated prefix p' and the erroneous word that follows the validated prefix e, in order to search for a transcription suffix $\hat{s}$:

$$\begin{aligned}\hat{s} &= \arg\max_s Pr(s|x, p', e) \\ &= \arg\max_s Pr(x|p', s, e) \cdot Pr(s|p', e) \end{aligned} \quad (5)$$

As in Eq. (1), $Pr(x|p', s, e)$ can be modelled using HMMs. On the other hand, $Pr(s|p', e)$ can be approached by adapting an n-gram language model so as to cope with the validated prefix p' and with the erroneous word e. The language model presented in section 2 would provide a model for the probability $Pr(s|p')$, but now the first word of s is conditioned by e. Therefore, some changes are needed.

Let $p' = w_1^k$ be the validated prefix and $s = w_{k+1}^l$ be a possible suffix and considering that the wrong-recognized word e only affects the first word of the suffix w_{k+1}, $Pr(s|p', e)$ can be computed as:

$$\begin{aligned}&Pr(s|p', e) \simeq Pr(w_{k+1}|w_{k+2-n}^k, e) \cdot \\ &\prod_{i=k+2}^{k+n-1} Pr(w_i|w_{i-n+1}^{i-1}) \cdot \prod_{i=k+n}^{l} Pr(w_i|w_{i-n+1}^{i-1}) \end{aligned} \quad (6)$$

Now, taking into account that the first word of the possible suffix w_{k+1} has to be different to the erroneous word e, $Pr(w_{k+1}|w_{k+2-n}^k, e)$ can be formulated as follows:

$$Pr(w_{k+1}|w_{k+2-n}^k, e) = \frac{\bar{\delta}(w_{k+1}, e) \cdot Pr(w_{k+1}|w_{k+2-n}^k)}{\sum_{w'} \bar{\delta}(w', e) \cdot Pr(w'|w_{k+2-n}^k)} \quad (7)$$

where $\bar{\delta}(i, j)$ is 0 when $i = j$ and 1 otherwise.

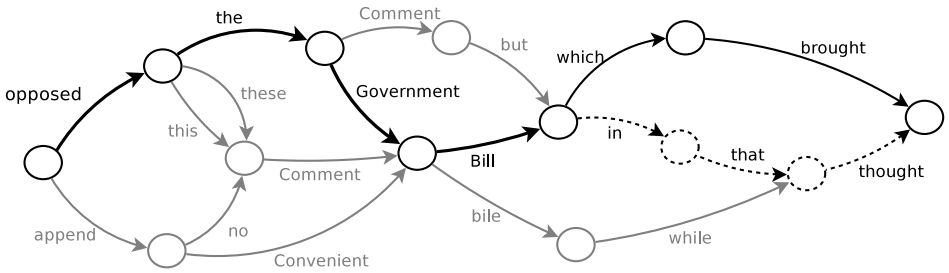

Figure 2. Example of word-graph generated after the user validates the prefix "opposed the Government Bill". The edge corresponding to the wrong-recognized word "in" was disabled.

As in the conventional CATTI, this decoding can be implemented using word-graphs. The restrictions entailed by (7) can be easily implemented by deleting the edge labeled with the word e after the prefix has been matched. An example is shown in fig. 2.

4. HTR System Overview

The HTR system used here follows a classical architecture composed of three modules: preprocessing, feature extraction and recognition. The first one entail different well-known standard techniques such as skew and slant corrections and size normalization. On the other hand, the feature extraction process transforms a preprocessed text line image into a sequence of 60-dimensional feature vectors (see [6]).

The recognition process is based on HMMs. Characters are modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. On the other hand, each lexical word is modelled by a Stochastic Finite-State automaton, and text sentences are modelled using bi-grams. All these finite-state models can be integrated into a single global model in which the decoding process is performed using the word-graphs obtained by the Viterbi algorithm [2].

5. Experimental Results

In order to test the effectiveness of using MA in the CATTI system, different experiments were carried out. The corpus, the different measures and the obtained experimental results are explained below.

5.1. IAMDB Corpus

This publicly accessible corpus was compiled by the Research Group on Computer Vision and Artificial Intelligence (FKI) at Institute of Computer Science an Applied Mathematics (IAM). The acquisition was based on the Lancaster-Oslo/Bergen Corpus (LOB). Fig. 1 shows an example of a handwritten sentence images from this corpus.

The last released version (3.0) is composed of 1,539 scanned text pages, handwritten by 657 different writers. The database is provided at different segmentation levels: words, lines, sentences and page images. In our case, the sentence segmentation level is considered (see [9]).

The corpus was partitioned into training and test sets. The former is composed of 5,799 text lines which add up to 2,124 sentences, handwritten by 448 different writers, whereas the latter comprises 200 sentences, written by 100 different writers. Table 1 summarizes all this information.

Table 1. Basic statistics of the IAMDB corpus

Number of:	Training	Test	Total	Lex.
writers	448	100	548	–
sentences	2,124	200	2,324	–
words	42,832	3,957	46,789	8,938
characters	216,774	20,726	237,500	78

5.2. Assessment Measures

Different evaluation measures have been adopted. On the one hand, the quality of the transcription without any system-user interactivity is given by the well known *word error rate* (WER). On the other hand, *the word stroke ratio* (WSR) can be defined as the number of (word level) user interactions that are necessary to produce correct transcriptions using the CATTI system, divided by the total number of reference words. Finally, the *word click rate* (WCR) can be defined as the number of additional MA per word that the user has to do using the new interaction with respect to using the conventional CATTI system.

The relative difference between WER and WSR (called Estimated Effort-Reduction) gives us an estimation of the reduction in human effort achieved, in terms of words to be corrected, by using CATTI with respect to using a conventional HTR system followed by human postediting.

Note that the additional human effort needed for the verification of the transcription and positioning the cursor in the appropriate place is the same in both conventional CATTI and new single-MA user-CATTI interaction system. In both cases the user should read the transcription proposed by the system until he or she finds an error and then position the cursor in the place where the new word has to be typed.

5.3. Results

Table 2 shows the obtained results. In the first part (left), we can see an estimation of the reduction in human effort (E-R) achieved by using the conventional CATTI system with respect to the classic HTR post editing. In the second part (right), the results obtained with the new single-MA interaction mode are shown. It is important to notice

Table 2. Results obtained with the corpus IAMDB using the conventional CATTI (left) and the new single-MA interaction (right)

Conventional CATTI			single-MA interaction	
WER	WSR	E-R	WSR	E-R
25.6%	23.4%	8.6%	19.6%	23.4%

that some of the results in table 2 do not correspond with those reported in [7]. The differences are because in [7] full Viterbi search was used, while in this work a much faster search technique based on (pruned) word graphs is adopted.

According to table 2, the estimated human effort to produce error-free transcriptions using MAs is significantly reduced with respect to using a conventional HTR system or the conventional CATTI. The new interaction mode can save about 23% of the overall effort. Fig. 3 shows the WSR, the Estimated Effort-Reduction (E-R) and the word click rate (WCR) as a function of the maximal number of MA allowed by the user before writing the correct word. The first point (0) corresponds to the results of the conventional CATTI, and the point "S" corresponds to the the single-MA interaction considered in the previous table. A good trade-off is obtained when the maximum number of clicks is around 3, because a significant amount of expected human effort is saved with a fairly low number of extra clicks per word.

6. Remarks and Conclusions

In this paper, we have considered new user feedback sources for CATTI. By considering MAs, we have shown that a significant benefit can be obtained, in terms of word-stroke reductions. A simple implementation using word-graphs has been described and some experiments have been carried out.

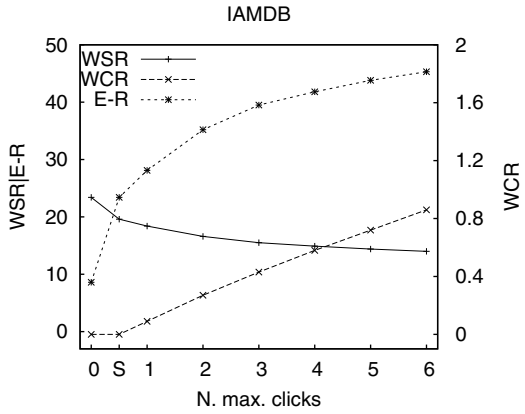

Figure 3. WSR, E-R and WCR as a function of the maximal number of MA allowed by the user before writing the correct word.

It is worth noting that alternative (n-best) suffixes could also be obtained with the conventional CATTI system. However, by considering the rejected words to propose the alternative suffixes, the interaction methods here studied are more effective and more comfortable for the user. Moreover, using the single-MA interaction method, a second alternative suffix is obtained without extra human effort.

References

[1] K. T. A. Ogawa and F. Itakura. Balancing acoustic and linguistic probabilites. *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, pages 181–184, 1998.

[2] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[3] L. Rabiner. A Tutorial of Hidden Markov Models and Selected Application in Speech Recognition. *Proc. IEEE*, 77:257–286, 1989.

[4] V. Romero et al. Improvements in the computer assisted transciption system of handwritten text images. In *Proc. of the PRIS 2008*, pages 103–112, June 2008.

[5] V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal. Computer assisted transcription for ancient text images. *In Proc. of ICIAR 2007*, Vol. 4633:1182–1193, 2007.

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

[7] A. H. Toselli et al. Computer assisted transcription of text images and multimodal interaction. In *Proc. of the MLMI*, volume 5237 of *LNCS*, pages 296–308. 2008.

[8] A. H. Toselli, V. Romero, L. Rodríguez, and E. Vidal. Computer assisted transcription of handwritten text. In *Proc. of ICDAR 2007*, pages 944–948. IEEE Computer Society, 2007.

[9] M. Zimmermann, J.-C. Chappelier, and H. Bunke. Offline grammar-based recognition of handwritten sentences,. *IEEE TPAMI*, 28(5):818–821, May 2006.