Handwritten Text Recognition for Marriage Register Books

Verónica Romero, Joan Andreu Sánchez, Nicolás Serrano, Enrique Vidal
Instituto Tecnológico de Informática
Universitat Politècnica de València
Valencia, Spain.
{*vromero,jandreu,nserrano,evidal*}@*iti.upv.es*

***Abstract*—Marriage register books are documents that were used for centuries by ecclesiastical institutions to register marriages. Most of these books were handwritten. These documents have interesting information, useful for demography studies. The information in these books is usually collected by expert demographers that devote a lot of time to transcribe them. The automatic transcription of these documents by using Handwritten Text Recognition techniques is difficult since the vocabulary is large, given that it is composed mainly of proper names. In this work, interactive Handwritten Text Recognition techniques were studied for the assisted transcription of these documents.**

***Keywords*-handwritten text recognition; marriage register books; interactive framework, assisted transcription**

I. Introduction

In the last years, huge amounts of handwritten historical documents residing in libraries, museums and archives have been digitalized and have been made available to scholars and to the general public through specialized web portals. However, there still exist a lot of documents residing in other institutions that have not been made public and that contain very interesting information. One example of this kind of handwritten documents are ecclesiastical archives, such as the marriage register books considered in this work.

The automatic transcription of these ancient handwritten documents is still an incipient research field that in recent years has been started to be explored. Currently available OCR text recognition technologies are very far from offering useful solutions to the transcription of this sort of documents, since usually characters can by no means be isolated automatically. Therefore, the required technology should be able to recognize all text elements (sentences, words and characters) as a whole, without any prior segmentation. This technology is generally referred to as *"off-line Handwritten Text Recognition"* (HTR) [1].

Several approaches have been proposed in the literature for HTR that resemble the noisy channel approach that is currently used in Automatic Speech Recognition. Thus, the HTR systems are based on Hidden Markov Models (HMM) [2] or hybrid HMM and Artificial Neural Networks (ANN) [3]. The transcription process can be made both significantly faster and more precise by using Language Models. For tasks in which the vocabulary size is very large, the language model plays a fundamental role in this noisy channel approach by restricting notably the search space.

For unconstrained text images, current HTR state-of-the-art prototypes provide accuracy levels that roughly range from 20 to 60% [1], [2]. If full correct transcriptions are needed, a human-expert revision is required. Recently, an interactive-predictive HTR approach was introduced [4] in which the recognition system and the human expert collaborate in order to obtain the correct transcription. In this approach, the system takes profit of the corrective feedback provided by the user as soon as it is available in order to provide a better solution in the next step. The transcribed text can be used to adapt the system models and/or to restrict the search space. This approach has demonstrated to be effective in order to reduce the human effort needed [5].

In this work we study the above mentioned interactive-predictive HTR technology for a task in which the vocabulary size is large and the language model hardly contributes to restrict the search space. The task consists in the transcription of indexes of marriage register books. This task is characterized by a large vocabulary size with regard to the small number of running words. Each book has an index at the beginning in order to register the page in which each marriage record is located. In this work we focused in the transcription of the index of a marriage register book. This sort of documents is being used for migratory studies, and therefore their transcription is considered an interesting problem [6].

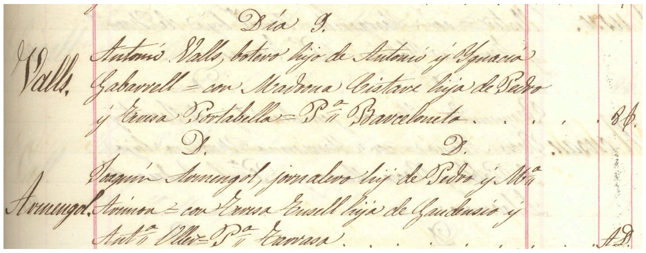
This task is described in detail in the following section, and the main problems that arise in these documents are explained. Then, the HTR technology used for tackling this problem and the experiments are reported in Sections III and IV. Finally, conclusions are draw in Section V.

II. Task description

Marriage register books are documents that were used for centuries to register marriages in ecclesiastical institutions (sometimes associated to a church). Most of these books were handwritten documents, with a structure analogous to an accounting book. In the books considered in this work, each record typically has the marriage day, the names of people to be married, their parent's names, the name of the church where the two people were baptised, and

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.113

the marriage stipends. Figure 1 shows two handwritten consecutive marriage records. The record language depended on the place where the church was located. This language was changing along the centuries. The records in Figure 1 are in Spanish. As can be noted, most of the vocabulary is composed by proper nouns.



Día 9.

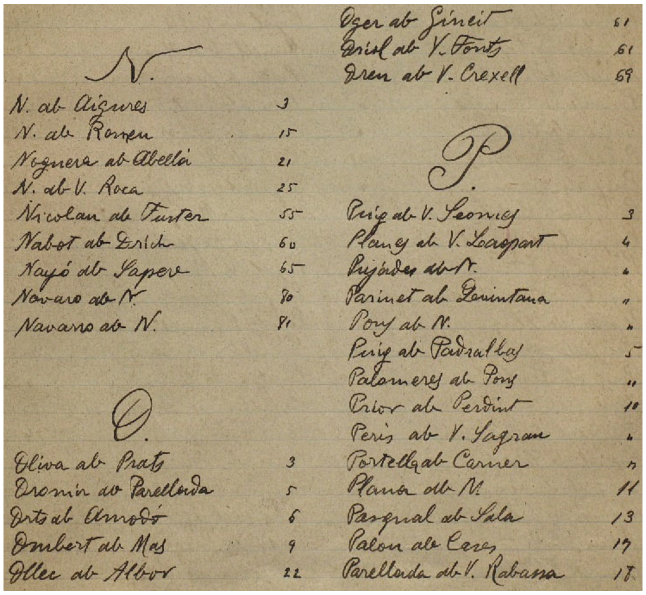
Valls. Antonio Valls, botero hijo de Antonio y Ygnacia
Gabarrell = con Madrona Cristave hija de Pedro
y Teresa Portabella = P.ª Barceloneta . . . . 86.

D. D.

Armengol. Joaquin Armengol, jornalero hij de Pedro y M.ª
Ximon = con Teresa Ensell hija de Gaudencio y
Ant.ª Oller = P.ª Tarrasa . . . . 42.

Figure 1. Example of two marriage records in Spanish.

Most books had an index at the beginning, and each index line had the surnames of the two persons being married and a page number of the record. Figure 2 shows an example of an index page that this time was in Catalan. In this work, we deal with the transcription of this kind of index pages. The surnames that appear in these indexes are being directly used for demography and migratory movement studies.



N.

N. ab Aigues	3
N. ab Romeu	15
Noguera ab Abellà	21
N. ab V. Roca	25
Nicolau ab Furter	55
Nabot ab Dorich	60
Nayó ab Sapere	65
Navarro ab N.	80
Navarro ab N.	81

O.

Oliva ab Prats	3
Oromir ab Parellada	5
Orts ab Armodó	6
Ombert ab Mas	9
Ollec ab Albor	22
Oger ab Simeit	61
Oriol ab V. Fonts	61
Oren ab V. Crexell	69

P.

Puig ab V. Seomies	3
Planes ab V. Llaopart	4
Pujades ab N.	〃
Parinet ab Quintana	〃
Pou ab N.	〃
Puig ab Padralbas	5
Palomeres ab Pons	〃
Prior ab Perdiut	10
Peris ab V. Sagran	〃
Portella ab Carner	〃
Plana ab M.	11
Pasqual ab Sala	13
Palou ab Ceres	17
Parellada ab V. Rabassa	18

Figure 2. Example of an index page of a marriage register book.

It is important to note that OCR approaches are useless for solving the transcription of these documents given that character symbols can not be automatically segmented with sufficient reliability. On the other hand, current state of the art HTR techniques are only able to obtain error-prone transcriptions. Therefore, interactive-predictive (IP) HTR techniques [4] were considered here in order to produce correct transcriptions of these pages. This technology takes profit of the language model in order to interactively restrict the search space. However, in this index, no language model can provide enough information to restrict this search space in an effective way.

Nevertheless, the indexes to be transcribed have some regularities that can be exploited in order to speed up the transcription process in the IP HTR approach. It can be noted that the syntactic structure of these lines is very simple. Each line has first a man surname, the word "ab" (that in old Catalan meant "with"), then a woman surname and the page. When one surname was unknown, then it was substituted by the word "N.". The index is organized alphabetically according the first surname letter. But lines starting with the same letter are not alphabetically ordered. Page numbers are in increasing order of surnames starting with the same letter, although not always. These characteristics were taken into account in the IP HTR process in order to reduce the transcription effort.

III. Interactive-Predictive Handwritten Text Recognition

The HTR system used in this paper follows the classical architecture composed of three main modules: document image preprocessing, line image feature extraction and Hidden Markov Model (HMM) training/decoding [7], [2].

The following steps take place in the preprocessing module: first, the skew of each page is corrected. Then, a conventional noise reduction method is applied on the whole document image, whose output is then fed to the text line extraction process which divides it into separate text lines images. Finally, slant correction and size normalization are applied on each separate line. More detailed description of this preprocessing can be found in [2], [8].

As our HTR system is based on HMMs, each preprocessed line image is represented as a sequence of feature vectors. To do this, the feature extraction module applies a grid to divide the text line image into $N \times M$ squared cells. In this work, $N = 20$ was adopted. This value has been empirically chosen after tuning the system. From each cell, three features are calculated: normalized gray level, horizontal gray level derivative and vertical gray level derivative. The way these three features are determined is described in [2]. Columns of cells or *frames* are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in its constituent cells. Hence, at the end of this process, a sequence of M 3×N-dimensional feature vectors is obtained.

Given a handwritten sentence image represented by a feature vector sequence, $\mathbf{x} = x_1\ x_2\ \dots\ x_m$, the HTR problem can now be formulated as the problem of finding a most likely word sequence, $\mathbf{w} = w_1\ w_2\ \dots\ w_n$, i.e.,

$\mathbf{w} = \text{argmax}_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x})$. Using the Bayes' rule we can decompose this probability into two probabilities, $P(\mathbf{x} \mid \mathbf{w})$ and $P(\mathbf{w})$, representing morphological-lexical knowledge and syntactic knowledge, respectively:

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) \approx \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \qquad (1)$$

$P(\mathbf{x} \mid \mathbf{w})$ is typically approximated by concatenated character models, usually HMMs [9], while $P(\mathbf{w})$ is approximated by a word language model, usually n-grams [9].

The characters are modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These models are estimated from training text images represented as feature vector sequences using the Baum Welch algorithm. Gaussian mixture serve as a probabilistic law to model the emission of feature vectors of each HMM state. The optimum number of HMM states and Gaussian densities per state are tuned empirically. On the other hand, each lexical word is modeled by a stochastic finite-state automaton, which represents all possible concatenations of characters to compose the word.

Finally, the concatenation of words into text line sentences is usually modeled by bi-grams, estimated from the training transcriptions of the text line images. However, in this work, given the kind of documents we are dealing with, we have used a very simple language model that strictly accounts for the simple syntactic structure of the lines studied. Note that an n-gram language model would not be useful for this task given that the second surname is not conditioned by the first surname. Figure 3 shows a graphical representation of this language model. First a surname must be recognized, then the word "ab", and then another surname that can be preceded by the word "V." (this letter means that the woman was widow and she was using her previous husband surname). Finally a page number or the quotation marks symbol must be recognized.

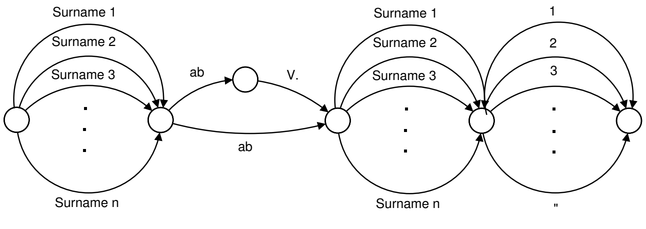


Figure 3. Language model for the index task.

All these finite-state (character, word and sentence) models can be *integrated* into a single *global* model on which a search process is performed for decoding the feature vectors sequence $\mathbf{x}$ into the words sequence $\mathbf{w}$. This search is optimally carried out by using the Viterbi algorithm [9].

A. Dynamic Language Modeling

To reduce the human effort needed to obtain the perfect transcription, we use the information derived during the interactive transcription process and the prior knowledge that we had about the task.

As we already know, the index is organized alphabetically according to the first surname letter, and it is expected to be transcribed alphabetically. Let α be the first letter of the man surname from the previous recognized and validated line. We can take advantage of the knowledge of α to improve the recognition of the man surname in the current line. Most specifically, we dynamically modify the language model to force the recognition of surnames beginning with α or with β, where $\beta > \alpha$ in alphabetical order (just in case the previous line was the last one corresponding to the letter α). For example, if the user validates a sentence where the man surname is "Oliva", we dynamically modified the language model to force the recognition of a man surname beginning with the character "O" or the character "P" in the upcoming sentence. If the prior probability of the letter β is very low, we added more letters in alphabetical order until the prior of the letters exceed a threshold.

B. Word Autocompleting

Furthering the goal of reducing the human effort to obtain the perfect transcription, we introduce an "autocompleting" approach. In this scenario, when the user enters a character to correct some incorrect surname transcription, the system automatically proposes the most probable surname that begins with the given surname prefix [11]. This process is iterated until a perfect transcription is obtained.

IV. Experimental Setup

In order to assess the use of the interactive predictive HTR technology and the autocompleting approach for transcribing an index document, different experiments were carried out. In the next subsections the information of the corpus, the assessment measures and the obtained results are explained.

A. Corpus

The corpus used on the experiments was compiled from the index at the beginning of a legacy handwritten marriage register book. This book was kindly provided by the *Centre d'Estudis Demogràfics* (CED) of the *Universitat Autònoma de Barcelona*. Figure 2 shows an example of an index page.

The "INDEX" corpus was written by only one writer and scanned at 300dpi. As a legacy document, it suffered the typical degradation problems of this kind of documents [12]: smear, significant background variations and uneven illumination, spots due to the humidity, and marks resulting from the ink that went through the paper (called "bleed-through").

INDEX was composed of 29 text pages. For each page, we used the GIDOC [13] prototype for: text block layout analysis, line segmentation, and transcription. Concretely, each text block layout was manually marked, next, a simple histogram-based line segmentation method was applied to each block, finally, text line images were automatically extracted and manually transcribed. The results were visually inspected and the few line-separation errors were manually corrected, resulting in a data-set of $1,563$ text line images, containing $6,534$ running words from a lexicon of $1,725$ different words.

Four different partitions were defined for cross-validation testing. All the information related with the different partitions is shown in Table I. The number of running words of each partition that does not appear in the other three partitions is shown in the OOV row (out of vocabulary).

Table I
BASIC STATISTICS OF THE DIFFERENT PARTITIONS FOR THE DATABASE "INDEX" .

	P0	P1	P2	P3	Total
Text lines	390	391	391	391	1,563
Words	1,629	1,640	1,632	1,633	6,534
Characters	7,629	7,817	7,554	7,809	30,809
OOV	326	346	298	350	1,320

B. Assessment Measures

Different evaluation measures were adopted to assess the HTR performance. The quality of the transcription at whole-word level is given by the well known *Word Error Rate* (WER). It is defined as the minimum number of words that need to be substituted, deleted or inserted to convert the sentences recognized by the system into the reference transcriptions, divided by the total number of words in these transcriptions. The WER is a good estimate of the non-interactive user effort. Since the predictive system presented only affects the first surname, the conventional classification error rate (ER) will be used to assess the accuracy of the predictions.

On the other hand, to assess the quality of the transcription at character level we used the well known *Character error rate* (CER), which is defined as the minimum number of character edit operations needed to edit an automatic transcription into the correct text, relative to the total number correct characters. Therefore, the CER is also a rough estimate of character-level user effort. To assess the "autocompleting" approach we have defined the "Key Stroke Ratio" (KSR) as the number of keystrokes that the user must enter to achieve the reference transcription, divided by the total number of reference characters.

These definitions make KSR and CER comparable.The relative difference between them gives us a good estimate of the reduction in human effort that can be achieved by using the autocompleting approach with respect to using a conventional post-editing system. The "Estimated Effort Reduction" will be denoted as "EFR".

C. Results

Two experiments were performed: one with closed vocabulary and other with open vocabulary. After some parameter tunning, WER of 28.6% and 40.2% were obtained in the closed and open vocabulary experiments respectively. In the first case, 9.2% of the errors corresponded to bad recognized man (first) surnames, a 9.3% to bad recognized woman (second) surnames. The remainder 10.1% corresponded to bad recognized numbers. On the other hand, in the open vocabulary case, 15,5% of the errors corresponded to the first surname, the 14.6% to the second surname and the 10.1% to number errors.

After a preliminary error analysis, we observed that around 250 errors from the 602 committed errors in the first surname using closed vocabulary were due to recognized surnames which began with a letter that was not correct. Similarly, using open vocabulary, 312 errors from the $1,006$ committed errors were due to the same problem.

Given that we know the index is organized alphabetically according the first surname letter, we consider the information provided by the user on the previous recognized and validated line to dynamically modified the language model as explained in Section III. Table II reports results for the recognition of the first surname of the $1,563$ text line images. In the first column we can see the error obtained without using the information of the previous line (ERc), whereas in the second column the result modifying dynamically the language model after each user interaction is shown (ERp). The last column of the table shows the relative difference of the two approaches. According to the results, the number of surnames that the user had to correct modifying the language model after each user interaction was reduced with respect to using the conventional HTR system. Therefore, the estimated human effort needed to obtain perfect transcription was reduced. The interactive predictive method can save around 9% and 3% using closed and open vocabulary, respectively.

Table II
ERROR OBTAINED WITH THE CONVENTIONAL SYSTEM (ERC) AND WITH THE PREDICTIVE SYSTEM (ERP). THE ESTIMATED EFFORT REDUCTION (EFR) FOR ERC WITH RESPECT TO ERP. ALL RESULTS ARE PERCENTAGES.

	ERc	ERa	EFR
Open voc.	9.2	8.4	8.7
Closed voc.	15.5	15.1	2.6

Tables III and IV show the results at character level obtained with closed and open vocabulary, respectively. The row labeled "Effort-Estimation" shows both plain and autocompleting results using both a conventional HTR system (without dynamically modifying the language model after each user interaction - CERc and KSRc) and a predictive HTR system (CERp and KSRp). An estimation of the reduction in human effort (EFR) achieved by using the predictive HTR system with autocompleting (KSRp) with respect to the plain HTR (vsCERc) is also shown. According to the results, the estimated human effort to produce error free transcription using an interactive-predictive HTR system with autocompleting is significantly reduced with respect to using the conventional HTR system. The overall estimated effort reductions are about 29% and 12% using closed and open vocabulary respectively.

Table III
CLOSED VOCABULARY CER AND KSR OBTAINED WITH THE CONVENTIONAL HTR APPROACH (CERC AND KSRC) AND WITH THE PREDICTIVE APPROACH (CERP AND KSRP). EFR FOR KSRC AND KSRP WITH RESPECT TO CERC AND CERP ARE ALSO SHOWN. ALL RESULTS ARE PERCENTAGES.

		CERc	CERp	KSRc	KSRp
Effort Estimation		17.4	16.4	13.3	12.3
EFR	vsCERc	0	5.7	23.6	**29.3**
	vsCERp	-	0	18.9	25.0

Table IV
OPEN VOCABULARY CER AND KSR OBTAINED WITH THE CONVENTIONAL HTR APPROACH (CERC AND KSRC) AND WITH THE PREDICTIVE APPROACH (CERP AND PKSRP). EFR FOR KSRC AND KSRP WITH RESPECT TO CERC AND CERP ARE ALSO SHOWN. ALL RESULTS ARE PERCENTAGES.

		CERc	CERp	KSRc	KSRp
Effort Estimation		25.5	23.4	24.7	22.5
EFR	vsCERc	0	8.2	3.1	**11.8**
	vsCERp	-	0	-	3.8

Note that, for real-world operations, the lower performance in the open vocabulary case can be improved using a large lexicon of surnames, instead of using only the surnames appearing on the training set.

V. CONCLUSIONS

In this paper, we have studied how to reduce the human effort needed to obtain the perfect transcriptions of the indexes of handwritten marriage register books. These documents were used for centuries by ecclesiastical institutions to register marriages and have interesting information that is being used by demographers that devote a lot of time to transcribe them. The automatic transcription of these documents by using Handwritten Text Recognition techniques is difficult since the vocabulary is large because it is composed mainly of proper names. Considering the results obtained in the experiments, we can conclude that, using the information derived during the interactive transcription process and the prior knowledge that we have of the task, significant amounts of human effort can be saved.

ACKNOWLEDGMENTS

Work supported by the Spanish Government (MICINN and “Plan E”) under the MITTRAL (TIN2009-14633-C03-01) research project and under the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), the Generalitat Valenciana under grant Prometeo/2009/014 and FPU AP2007-02867.

REFERENCES

[1] U.-V. Marti and H. Bunke, “Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System,” *IJPRAI*, vol. 15, no. 1, pp. 65–90, 2001.

[2] A. H. Toselli *et al.*, “Integrated Handwriting Recognition and Interpretation using Finite-State Models,” *IJPRAI*, vol. 18, no. 4, pp. 519–539, 2004.

[3] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez, “Improving offline handwriting text recognition with hybrid hmm/ann models,” *IEEE Transactions on PAMI*, vol. 33, no. 4, pp. 767–779, 2011.

[4] A. Toselli, V. Romero, M. Pastor, and E. Vidal, “Multimodal interactive transcription of text images,” *Pattern Recognition*, vol. 43, no. 5, pp. 1824–1825, 2009.

[5] A. H. Toselli, E. Vidal, and F. Casacuberta, Eds., *Multimodal Interactive Pattern Recognition and Applications*, 1st ed. Springer, Jun 2011, http://www.springer.com/computer/hci/book/978-0-85729-478-4.

[6] A. Esteve, C. Cortina, and A. Cabré, “Long term trends in marital age homogamy patterns: Spain, 1992-2006,” *Population*, vol. 64, no. 1, pp. 173–202, 2009.

[7] I. Bazzi, R. Schwartz, and J. Makhoul, “An Omnifont Open-Vocabulary OCR System for English and Arabic,” *IEEE Transactions on PAMI*, vol. 21, no. 6, pp. 495–504, 1999.

[8] V. Romero, M. Pastor, A. H. Toselli, and E. Vidal, “Criteria for handwritten off-line text size normalization,” in *Proceedings of the VIIP 06*, Palma de Mallorca, Spain, August 2006.

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

[10] A. Ogawa, K. Takeda, and F. Itakura, “Balancing acoustic and linguistic probabilites,” in *Proceeding IEEE CASSP*, vol. 1, Seattle, WA, USAR, May 1998, pp. 181–184.

[11] V. Romero, A. H. Toselli, and E. Vidal, “Using mouse feedback in computer assisted transcription of handwritten text images,” in *Proceedings of the 10th ICDAR*, ser. IEEE Computer Society, Barcelona, Spain, July 2009.

[12] F. Drida, “Towards restoring historic documents degraded over time,” in *Proceedings of the DIAL’06*, ser. IEEE Computer Society, Washington, DC, USA, 2006, pp. 350–357.

[13] N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades, and A. Juan, “The GIDOC prototype,” in *Proceedings of the 10th PRIS 2010*, Funchal (Portugal), pp. 82–89.