Handwritten Text Recognition for Historical Documents

Verónica Romero, Nicolás Serrano, Alejandro H. Toselli,
Joan Andreu Sánchez and **Enrique Vidal**
ITI, Universitat Politècnica de València, Spain
{vromero,nserrano,jandreu,atoselli,evidal}@iti.upv.es

Abstract

The amount of digitized legacy documents has been rising dramatically over the last years due mainly to the increasing number of on-line digital libraries publishing this kind of documents. The vast majority of them remain waiting to be transcribed into a textual electronic format (such as ASCII or PDF) that would provide historians and other researchers new ways of indexing, consulting and querying them. In this work, the state-of-the-art Handwritten Text Recognition techniques are applied for the automatic transcription of these historical documents. We report results for several ancient documents.

1 Introduction

In the last years, huge amount of handwritten historical documents residing in libraries, museums and archives have been digitalized and have been made available to the general public through specialized web portals. The vast majority of these documents, hundreds of terabytes worth of digital image data, remain waiting to be transcribed into a textual electronic format that would provide historians and other researchers new ways of indexing, consulting and querying them.

The automatic transcription of these ancient handwritten documents is still an incipient research field that has been started to be explored in recent years. For some time in the past decades, the interest in Off-line Handwritten Text Recognition (HTR) was diminishing, under the assumption that modern computer technologies will soon make paper-based documents useless. However, the increasing number of on-line digital libraries publishing large quantities of digitized legacy documents has turned HTR up in an important research topic.

HTR should not be confused with Optical Character Recognition (OCR). Nowadays, OCR systems are capable to recognizing text with a very good accuracy (Breuel, 2008; Ratzlaff, 2003). However, OCR products are very far from offering useful solutions to the HTR problem. They are simply not usable, since in the vast majority of the handwritten documents, characters can by no means be isolated automatically. HTR, specially for historical documents, is a very difficult task. To some extent HTR is comparable with the task of recognizing continuous speech in a significantly degraded audio file. And, in fact, the nowadays prevalent technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition (ASR) (Rabiner, 1989) as Hidden Markov Models (HMMs) (Bazzi et al., 1999) and n-Gram (Jelinek, 1998). The most important difference is that the input feature vector sequence of the HTR system represents a handwritten text line image, rather than an acoustic speech signal.

In this sense, the required technology should be able to recognize all the text elements (sentences, words and characters) as a whole, without any prior segmentation of the image into these elements. This technology is generally referred to as segmentation-free *off-line Handwritten Text Recognition* (HTR) (Marti and Bunke, 2001; Toselli and others, 2004; España-Boquera et al., 2011).

Given that historical documents suffered from the typical degradation problems of this kind of documents and, in order to obtain accurately transcription of them, different methods and techniques of the document analysis and recognition field are needed. Among them are the layout analysis and text line extraction methods, image preprocessing techniques, lexical and language modeling and HMMs. In this paper we study the adaptation/application of the above mentioned tech-

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 90–96,
Hissar, Bulgaria, 16 September 2011.

niques on historical documents, testing the system on four sort of different ancient documents characterized, among other things, by different handwritten styles from diverse places and time periods.

This paper is divided as follows. First, the HTR framework is introduced in section 2. Then, the different corpora used in the experiments are described in subsection 3.1. The experiments and results are commented in subsection 3.2. Finally, some conclusions are drawn in the section 4.

2 Handwritten Text Recognition

The handwritten text recognition (HTR) problem can be formulated in a similar way to ASR, as the problem of finding the most likely word sequence, $\mathbf{w} = (w_1\ w_2\ \ldots\ w_n)$, for a given handwritten sentence image represented by an observation sequence, $\mathbf{x} = (x_1\ x_1\ \ldots\ x_m)$, 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 $P(\mathbf{w})$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmax}}\, P(\mathbf{w} \mid \mathbf{x}) \approx \underset{\mathbf{w}}{\text{argmax}}\, P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \tag{1}$$

$P(\mathbf{x} \mid \mathbf{w})$ can be seen as a morphological-lexical knowledge. It is the probability of the observation sequence $\mathbf{x}$ given the word sequence $\mathbf{w}$ and is typically approximated by concatenated character HMMs (Jelinek, 1998). On the other hand, $P(\mathbf{w})$ represents a syntactic knowledge. It is the prior probability of the word sequence $\mathbf{w}$ and is approximated by a word language model, usually n-grams (Jelinek, 1998).

In practice, the simple multiplication of $P(\mathbf{x} \mid \mathbf{w})$ and $P(\mathbf{w})$ needs to be modified in order to balance the absolute values of both probabilities. To this end a language model weight α (Grammar Scale Factor, GSF), which weights the influence of the language model on the recognition result, and an insertion penalty β (Word Insertion Penalty, WIP), which helps to control the word insertion rate of the recognizer (Ogawa et al., 1998) are used. In addition, log-probabilities are usually used to avoid the numeric underflow problems that can appear using probabilities. So, Equation (1) can be rewritten as:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmax}}\ \log\ P(\mathbf{x} \mid \mathbf{w}) + \alpha\ \log\ P(\mathbf{w}) + l\beta \tag{2}$$

where l is the word length of the sequence $\mathbf{w}$ and α and β are optimized for all the training sentences of the corpus.

The HTR system used here follows the classical architecture composed of three main modules: a document image preprocessing module, in charge to filter out noise, recover handwritten strokes from degraded images and reduce variability of text styles; a line image feature extraction module, where a feature vector sequence is obtained as the representation of a handwritten text line image; and finally a HMM training/decoding module, which obtains the most likely word sequence for the sequence of feature vectors (Bazzi et al., 1999; Toselli and others, 2004).

2.1 Preprocessing

It is quite common for ancient documents to suffer from degradation problems (Drida, 2006). Among these are the presence of smear, background of big variations and uneven illumination, spots due to the humidity or marks resulting from the ink that goes through the paper (generally called bleed-through). In addition, there are other kinds of difficulties appearing in these pages as different font types and sizes in the words, underlined and/or crossed-out words, etc. The combination of all these problems contributes to make the recognition process difficult, and hence, the preprocessing module quite essential.

The following steps take place in the preprocessing module: first, the skew of each page is corrected. We understand as "skew" the angle between the horizontal direction and the direction of the lines on which the writer aligned the words. Then, a conventional noise reduction method is applied on the whole document image (Kavallieratou and Stamatatos, 2006), whose output is then fed to the text line extraction process which divides it into separate text lines images. The method used for the latter case is based on the horizontal projection profile of the input image. Local minimums in this projection are considered as potential cut-points located between consecutive text lines. When the minimum values are greater than zero, no clear separation is possible. This problem has been solved using a method based in connected components (Marti and Bunke, 2001). Finally, slant correction and size normalization are applied on each separate line. More detailed description can be found in (Toselli and others, 2004; Romero et al., 2006).

2.2 Feature Extraction

The feature extraction process approach used to obtain the feature vectors sequence follows similar ideas described in (Bazzi et al., 1999). First, a grid is applied to divide the text line image into $M \times N$ squared cells. M is chosen empirically and N is such that N/M equals the original line image aspect ratio. Each cell is characterized by the following features: *average gray level*, *horizontal gray level derivative* and *vertical gray level derivative*. To obtain smoothed values of these features, an $s \times s$ cell analysis window, centered at the current cell, is used in the computations (Toselli and others, 2004). The smoothed cell-averaged gray level is computed through convolution with two 1-d Gaussian filters. The smoothed horizontal derivative is calculated as the slope of the line which best fits the horizontal function of column-average gray level in the analysis window. The fitting criterion is the sum of squared errors weighted by a 1-d Gaussian filter which enhances the role of central pixels of the window under analysis. The vertical derivative is computed in a similar way.

Columns of cells (also called *frames*) are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in their constituent cells. Hence, at the end of this process, a sequence of M $3 \cdot N$-dimensional feature vectors (N normalized gray-level components and N horizontal and vertical derivatives components) is obtained. Figure 1 shows a representative visual example of the feature vectors sequence for the Spanish word "cuarenta" ("forty") and how a continuous density HMM models two feature vector subsequences corresponding to the character "a".

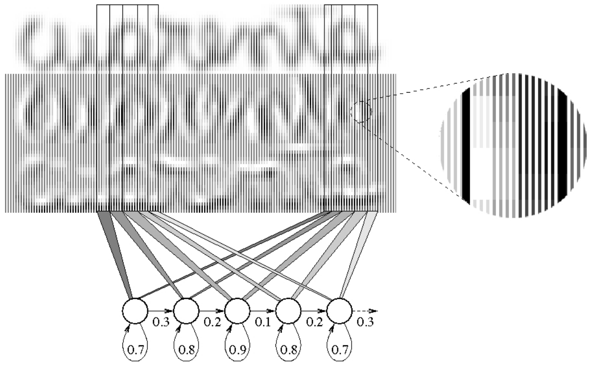

Figure 1: Example of feature-vector sequence and HMM modeling of instances of the character "a" within the Spanish word "cuarenta" ("forty"). The model is shared among all instances of characters of the same class. The zones modeled by each state show graphically subsequences of feature vectors (see details in the magnifying-glass view) compounded by stacking the normalized grey level and its both derivatives features.

2.3 Recognition

Characters (or graphemes) are considered here as the basic recognition units in the same way as phonemes in ASR, and therefore, they are modeled by left-to-right HMMs. Each HMM state generates feature vectors following and adequate parametric probabilistic law; typically, a Gaussian Mixture. Thereby, the total amount of parameters to be estimated depends on the number of states and their associated emission probability distributions, which need to be empirically tuned to optimize the overall performance on a given amount of available training samples. As in ASR, character HMMs are trained from images of continuously handwritten text (without any kind of segmentation and represented by their respective observation sequences) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called forward-backward or Baum-Welch re-estimation (Jelinek, 1998).

Each *lexical entry (word)* is modeled by a stochastic finite-state automaton which represents all possible concatenations of individual characters that may compose the word. By embedding the character HMMs into the edges of this automaton, a *lexical HMM* is obtained.

Finally, the concatenation of words into text lines or sentences is usually modeled by a bi-gram *language model*, with Kneser-Ney back-off smoothing (Katz, 1987; Kneser and Ney, 1995), which uses the previous $n-1$ words to predict the next one:

$$P(\mathbf{w}) \approx \prod_{i=1}^{N} P(w_i|\mathbf{w}_{i-n+1}^{i-1}) \quad (3)$$

This n-grams are estimated from the given transcriptions of the trained set.

However, there are tasks in which the relation of running words and vocabulary size is too low causing that bi-gram language models hardly contributes to restrict the search space. This is the case of one of the documents used in the experiments reported in section 3.2 called "Index". In the following subsection we describe the language model used for recognition in this specific task.

Once all the *character*, *word* and *language* models are available, the recognition of new test sentences can be performed. Thanks to the homogeneous finite-state (FS) nature of all these models, they can be easily *integrated* into a single *global* (huge) FS model. Given an input sequence of feature vectors, the output word sequence hypothesis corresponds to a path in the integrated network that produces the input sequence with highest probability. This optimal path search is very efficiently carried out by the well known Viterbi algorithm (Jelinek, 1998). This technique allows for the integration to be performed "on the fly" during the decoding process.

2.3.1 "Index" Language Model

The Index task (see section 3.2) is related to the transcription of a marriage register book and corresponds to the transcription of the index at the beginning of one of these books. This index registers the page in which each marriage record is located. These marriage register books were usually used for centuries to register marriages in ecclesiastical institutions and have been used recently for migratory studies. Their transcription is considered an interesting problem (Esteve et al., 2009). These index pages have some regularities and a very easy syntactic structure. The lines of the index pages used in this study have first a man surname, then the word "ab" (that in old Catalan means "with"), then a woman surname and finally the page number in which that marriage record was registered.

In this work, in order to improve the accuracy and speed up the transcription process of this document, we have defined a very simple language model that strictly accounts for the easy syntactic structure of the lines. Figure 2 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.

3 Experimental Results

In order to assess the effectiveness of the above-presented off-line HTR system on legacy documents, different experiments were carried out. The

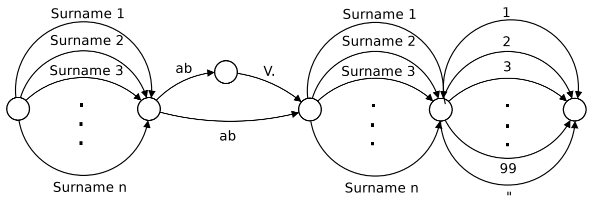

Figure 2: Language model for the Index task.

corpora used in the experiments, as well as the different measures and the obtained experimental results are summarized in the following subsections.

3.1 Corpora and Transcription Tasks

Four corpora with more or less similar HTR difficulty were employed in the experiments. The first three corpora, CS (Romero et al., 2007), Germana (Pérez et al., 2009) and Rodrigo (Serrano and Juan, 2010), consist of cursive handwritten page images in old Spanish from 16th and 19th century. The last corpus: Index (Romero et al., 2011), was compiled from the index at the beginning of a legacy handwritten marriage register book. Figure 3 shows examples of each of them.

Cristo-Salvador This corpus was compiled from the legacy handwriting document identified as *"Cristo-Salvador"*, which was kindly provided by the *Biblioteca Valenciana* (BIVAL)[1].

It is composed of 53 text page images, written by only one writer and scanned at 300dpi. As has been explained in section 2, the page images have been preprocessed and divided into lines, resulting in a data-set of $1,172$ text line images. The transcriptions corresponding to each line image are also available, containing $10,911$ running words with a vocabulary of $3,408$ different words.

Two different partitions were defined for this data-set. In this work we are going to use the partition called *hard* (Romero et al., 2007), where the test set is composed by 497 line samples belonging to the last 20 document pages, whereas the remaining 675 were assigned to the training set.

Germana The GERMANA corpus is the result of digitizing and annotating the Spanish manuscript *"Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragón"* written in 1891. It is a single-author book and a limited-domain topic, and the original manuscript

[1] http://bv2.gva.es

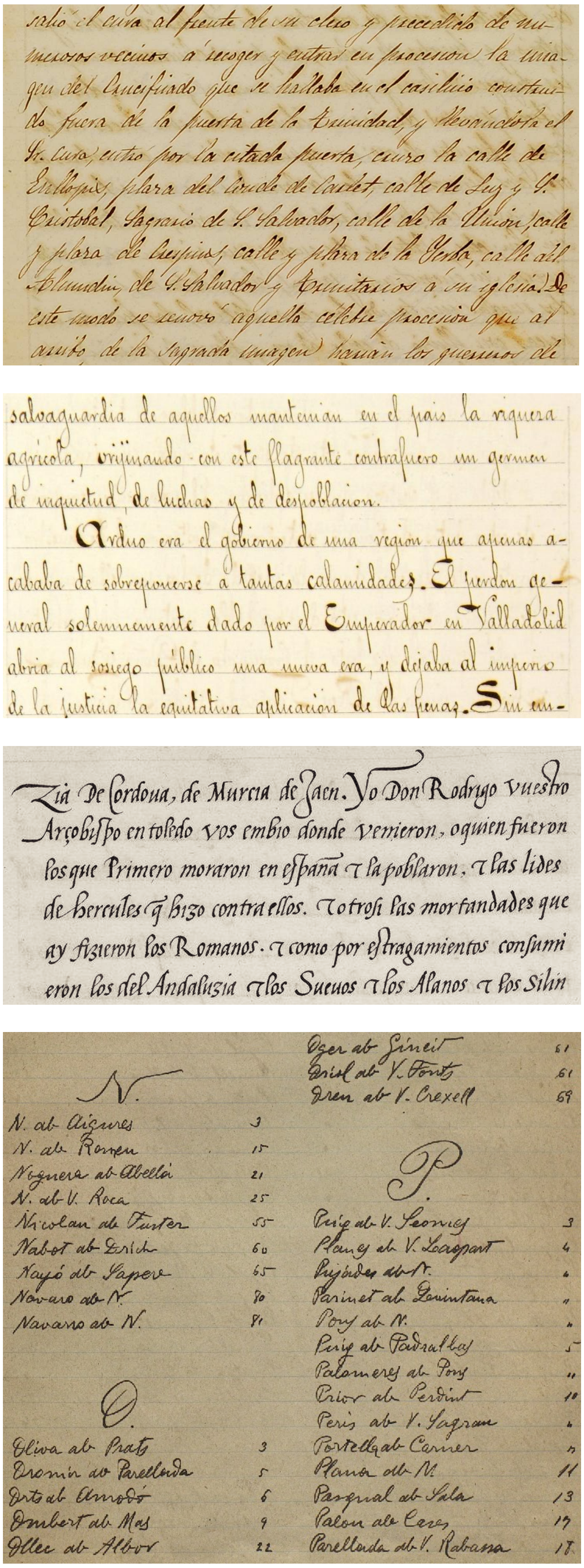

Figure 3: From top to bottom: Single-Writer Manuscripts from the XIX Century (CS and Germana), Single-Writer Spanish manuscript from XVI century (Rodrigo) and index page of a marriage register book (Index).

was well-preserved (Pérez et al., 2009). It is composed of 764 pages, with approximately $21k$ lines.

The page images were preprocessed and divided into lines 2. These lines have been transcribed by paleography experts, resulting in a data-set of $217k$ running words with a vocabulary of $30k$ words. To carry out the experiments, we have used the same partition described in (Pérez et al., 2009), that only uses the first 180 pages of the corpus.

Rodrigo The Rodrigo database corresponds to a manuscript from 1545 entitled *”Historia de España del arçobispo Don Rodrigo“*, and written in old Spanish by a single author. It is 853-page bound volume divided into 307 chapters.

The manuscript was carefully digitized by experts at 300 dpi and annotated in a procedure very similar to the one used for the Germana database. The complete annotation of Rodrigo comprises about $20K$ text lines and $231K$ running words form a lexicon of $17K$ words. In this work, the experiments have been carried out using the same partition described in (Serrano and Juan, 2010)

Index This corpus 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*. As previously said, the lines in these pages have some syntactic regularities that could be used to reduce the human effort needed to carry out the transcription (Romero et al., 2011).

The Index corpus was written by only one writer and scanned at 300 dpi. It was composed by 29 text pages. For each page, the GIDOC (Serrano et al.,) prototype was used to perform text block layout, line segmentation, and transcription. 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.

3.2 Results

The quality of the transcriptions obtained with the off-line HTR system is given by the well-known Word Error Rate (WER). It is widely used in HTR (Toselli and others, 2004; Toselli et al., 2010; España-Boquera et al., 2011) and in ASR (Jelinek, 1998). It is defined as the minimum number of words that need to be substituted, deleted or inserted to convert a sentence recognized by the system into the corresponding reference transcription, divided by the total number of words in the reference transcription.

The corresponding morphological (HMMs) and language models (the different *bi*-grams and the special language model for the Index task) associated with each corpus were trained from their respective training images and transcriptions. Besides, all results reported in Table 1 have been obtained after optimizing the parameters values corresponding to the preprocessing, feature extraction and modeling processes for each corpus.

Concerning to the CS corpus, the obtained WER (%) results was 33.5 using in this case a closed-vocabulary. For the Germana corpus, the best WER achieved were around 8.9% and 26.9% using closed-vocabulary and open-vocabulary respectively. Regarding the out-of-vocabulary (OOV) words, it becomes clear that a considerable fraction of transcription errors is due to the occurrence of unseen words in the test partition. More precisely, unseen words account here for approximately 50% of transcription errors. Although comparable in size to GERMANA, RODRIGO comes from a much older manuscript (from 1545), where the typical difficult characteristics of historical documents are more evident. The best WER figure achieved in this corpus until the moment is around 36.5%, where most of the errors are also caused by the occurrence of OOV words. Respect to the Index corpus, in which the transcription process used a specific language model, WER of 28.6% and 40.3% were obtained for closed-vocabulary and open-vocabulary respectively.

From the results we can see that current state-of-the-art segmentation-free "off-line HTR" approach produces word error rates as high as 9-40% with handwritten old documents, depending whether open or closed vocabulary is used. These results are still far from offering perfect solutions to the transcription problem. However, this accuracy could be enough for indexing and searching tasks or even to derive adequate metadata to roughly describe the ancient document contents.

4 Conclusions

In this paper the nowadays technology of HTR, which borrows concepts and methods from the field of Automatic Speech Recognition technology, has been tested for historical documents. This HTR technology is based on Hidden Markov Models using Gaussians as state emission probability function. The HMM-based HTR has a hierarchical structure with character HMMs modelling the basic recognition units. These models are concatenated forming word models, and these in turn concatenated forming sentence models. The HMM used in this work was furthermore enhanced by a language model incorporating linguistic information beyond the word level.

Several tasks have been considered to assess this HTR approach. Considering all the difficulties involving the old handwritten documents used in the experiments, although the results achieved are not perfect they are really encouraging. In addition, as previously commented, this accuracy could be enough for tasks such as document indexing and searching or even could be used to derive adequate metadata that describes roughly the content of documents. Moreover, other applications such as word-spotting can be easily implemented using this segmentation-free HTR technology. In this sense, results are expected to be much more precise than using the popular approaches which do not take advantage of the context of spotted words.

Finally, to obtain perfect transcriptions, instead of the heavy human-expert "post-editing" work, that generally results inefficient and uncomfortable to the user and also it is hardly accepted by expert transcribers, computer assisted interactive predictive solutions (Toselli et al., 2010) can be used. These solutions offer promising significant improvements in practice and user acceptance. In these approaches, the user and the system work interactively in tight mutual cooperation to obtain the final perfect transcription of the given text images.

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.

Table 1: Basic statistics information from each corpus along with the WER(%) obtained using the segmentation-free off-line HTR system.

Corpus		CS	GERMANA		RODRIGO	INDEX	
Language		19th C Sp.	19th C Sp.		16th C Sp.	Old Catalan	
Lan. Model	Lexicon	2 277	7 477		17 300	1 725	
	Train. Ratio	2.8	4.5		12.5	3.8	
HMMs	Characters	78	82		115	68	
	Train. Ratio	460	2 309		7 930	453	
Open Vocabulary		N	N	**Y**	**Y**	N	**Y**
WER (%)		**33.5**	**8.9**	**26.9**	**36.5**	**28.6**	**40.3**

References

I. Bazzi, R. Schwartz, and J. Makhoul. 1999. An Omnifont Open-Vocabulary OCR System for English and Arabic. *IEEE Transactions on PAMI*, 21(6):495–504.

Thomas M. Breuel. 2008. The ocropus open source ocr system. In *DRR*, page 68150.

F. Drida. 2006. Towards restoring historic documents degraded over time. In *Proceedings of the DIAL'06*, IEEE Computer Society, pages 350–357. Washington, DC, USA.

S. España-Boquera, M.J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez. 2011. Improving offline handwriting text recognition with hybrid hmm/ann models. *IEEE Transactions on PAMI*, 33(4):767–779.

A. Esteve, C. Cortina, and A. Cabré. 2009. Long term trends in marital age homogamy patterns: Spain, 1992-2006. *Population*, 64(1):173–202.

F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press.

S. M. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35:400–401, March.

E. Kavallieratou and E. Stamatatos. 2006. Improving the quality of degraded document images. In *Proceedings of the DIAL '06*, pages 340–349, Washington, DC, USA. IEEE Computer Society.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. volume 1, pages 181–184, Los Alamitos, CA, USA. IEEE Computer Society.

U.-V. Marti and H. Bunke. 2001. Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. *IJPRAI*, 15(1):65–90.

A. Ogawa, K. Takeda, and F. Itakura. 1998. Balancing acoustic and linguistic probabilites. In *Proceeding IEEE CASSP*, volume 1, pages 181–184, Seattle, WA, USAR, May.

Daniel Pérez, Lionel Tarazón, Nicolás Serrano, Francisco-Manuel Castro, Oriol Ramos-Terrades, and Alfons Juan. 2009. The germana database. In *Proceedings of the ICDAR'09*, pages 301–305, Barcelona (Spain), July. IEEE Computer Society.

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

E. H. Ratzlaff. 2003. Methods, Report and Survey for the Comparison of Diverse Isolated Character Recognition Results on the UNIPEN Database. In *Proceedings of ICDAR '03*, volume 1, pages 623–628, Edinburgh, Scotland, August.

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

V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal. 2007. Computer Assisted Transcription for Ancient Text Images. In *Proocedings of the ICIAR 2007*, volume 4633 of *LNCS*, pages 1182–1193. Springer-Verlag, Montreal (Canada), August.

V. Romero, Joan Andreu Sánchez, Nicolás Serrano, and E. Vidal. 2011. Handwritten text recognition for marriage register books. In *Proceedings of the 11th ICDAR*, IEEE Computer Society. To be published, September.

Nicolás Serrano and Alfons Juan. 2010. The rodrigo database. In *Proceedings of the LREC 2010*, Malta, May 19-21.

N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades, and A. Juan. The GIDOC prototype. In *Proceedings of the 10th PRIS 2010*, pages 82–89, Funchal (Portugal).

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

A.H. Toselli, V. Romero, M. Pastor, and E. Vidal. 2010. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1824–1825.