Evaluating a Post-editing Approach for Handwriting Transcription

Verónica Romero, Joan Andreu Sánchez, Nicolás Serrano, Enrique Vidal
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València, Spain
{vromero,jandreu,nserrano,evidal}@dsic.upv.es

Abstract

Marriage license books are documents that were used for centuries by ecclesiastical institutions to register marriage licenses. These books, that were handwritten until the beginning of the 20th century, have interesting information, useful for demography studies and genealogical research. This information is usually collected by expert demographers that devote a lot of time to manually transcribe them. As the accuracy of automatic handwritten text recognizers improves, post-editing the output of these recognizers could be foreseen as a possible alternative. Unluckily, most handwriting recognition techniques require large amounts of annotated images to train the recognition engine. In this paper we carry out a study about how the handwritten recognition system accuracy improves with respect to the amount of training data, and how the human efficiency increases during the transcription of a marriage license book.

1 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In the last years, huge amounts of handwritten historical documents residing in libraries, museums and archives have been digitalized and have been made available to scholars and to the general public through specialized web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. Many of these documents are collections of historical documents containing very valuable information in the form of records of quotidian activities. One example of this kind of handwritten documents are the marriage license books considered in this paper. In many cases, it would be interesting to transcribe these document images, in order to provide new ways of indexing, consulting and querying them.

Transcribing handwritten images manually is a very laborious and expensive work. This work is usually carried out by experts in palaeography, who are specialized in reading ancient scripts, characterized, among other things, by different handwritten/printed styles from diverse places and time periods. How long experts take to make a transcription of one of these documents depends on their skills and experience.

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) (Marti and Bunke, 2001). 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) (Toselli and others, 2004), recurrent neural networks (Graves et al., 2009) or hybrid HMM and neural networks (España-Boquera et al., 2011). These systems have proven to be

suited for restricted applications with very limited vocabulary or constrained handwriting achieving in these kind of tasks relatively high recognition rates. However, in the case of transcription applications of unconstrained handwritten documents (as old manuscripts), the current HTR technology typically achieves results which are far from perfect.

Therefore, once a full recognition process of the document has finished, human intervention is required in order to produce a high quality transcription. The human transcriber is, therefore, responsible for verifying and correcting the mistakes made by the system. In this context, the HTR process is performed off-line: First, the HTR system returns a full transcription of all the text lines in the whole document. Then, the human transcriber reads them sequentially (while looking at their correspondence in the original page images) and corrects the possible mistakes made by the system. Whether the HTR system accuracy is good enough, this post-editing approach can be foreseen as a possible alternative to manually transcription.

In the above mentioned HTR approaches, the character and language models are stochastic models whose parameters are automatically learned from annotated data. One of the bottlenecks of these approaches is the need of annotated data in order to automatically train the models. These annotated data is not usually available at the beginning of the transcription of new document, but as the document is being transcribed, new transcribed material is available for training the HTR models.

In this paper we carry out an study about how the performance of an HTR system varies as the amount of data that is available to train the models increases. First, an HTR system provides automatic transcriptions for a few pages. Second, these transcriptions are post-edited by expert palaeographers, and the models are retrained with the post-edited transcriptions. Then, new pages are transcribed and manually reviewed, and the models are retrained. This process goes on until the complete document is transcribed. We also study how the improvements in the system accuracy produced by retraining the HTR models with the new transcribed material affect to the human efficiency following the post-editing approach. This study has been carried out during the real transcription of a historical book compiled from a collection of Spanish marriage license books.

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 in this work is shown in section 3. The evaluation methodology and the obtained results are reported in sections 4 and 5. Finally, conclusions are draw in Section 6.

2 Task description

Marriage license books are documents that were used for centuries to register marriages in ecclesiastical institutions. These demographic documents have already been proven to be useful for genealogical research and population investigations, which renders their complete transcription an interesting and relevant problem (Esteve et al., 2009). Most of these books are handwritten documents, with a structure analogous to an accounting book.

In this paper we have used a book from a collection of Spanish marriage license books conserved at the Archives of the Cathedral of Barcelona. In this book each page is divided horizontally into three blocks, the *husband surname's block*, the *main block*, and the *fee block* and vertically into individual license records. Fig. 2 shows a page of marriage licenses from the book used in this paper. Each marriage license (see Fig. 1) typically contains information about the marriage day, husband's and wife's names, the husband's occupation, the husband's and wife's former marital status, and the socio-economic position given by the amount of the fee. In some cases, additional information is given as well, viz. the father's names and their occupations, information about a deceased parent, place of residence, or geographical origin. The fiscal marker, as well as the exhaustive nature of the source and the variety of types of the parishes involved – from the city centre to the most rural villages – allows researching multiple aspects of demography, specially the chronology and the geography in the constitution of new social classes and groups.

Compared to modern printed documents, the

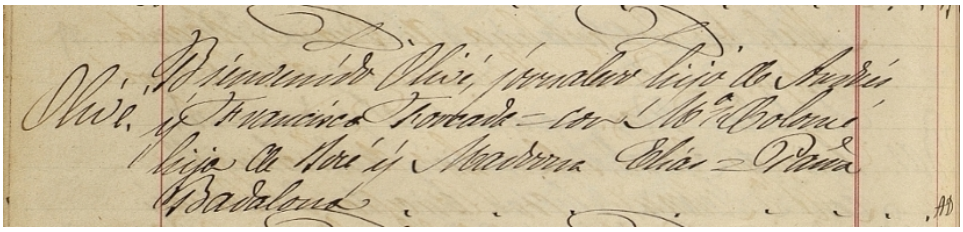

Figure 1: Example of a marriage licenses in Spanish.

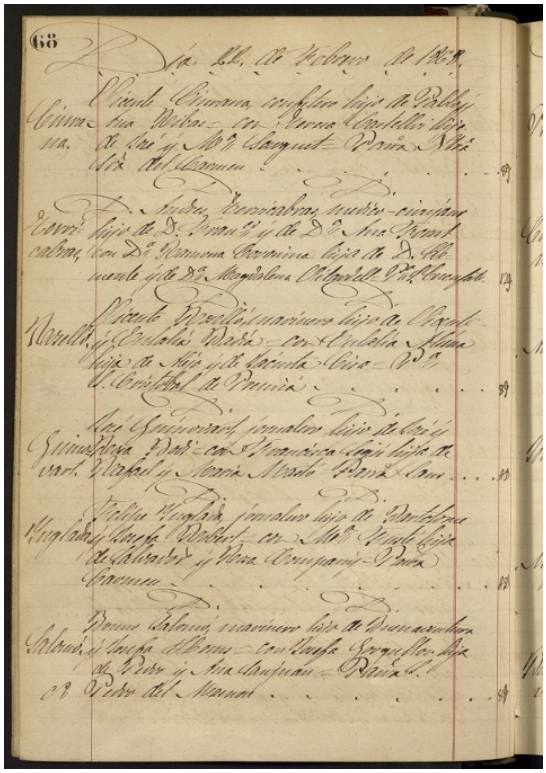

Figure 2: Example of a marriage licenses page.

analysis and recognition of these handwritten historical documents has many additional difficulties. Firstly, the typical paper degradation problems encountered in this kind of documents, such as presence of smear, significant background variation, uneven illumination, and dark spots, require specialized image-cleaning and enhancement algorithms. Secondly, show-through and bleed-through problems can render the distinction between background and foreground difficult (Drida, 2006). Thirdly, document collections spanning several centuries usually do not follow a strict standard notation, but differ from one century to another. The text contains a variety of special symbols and other recognition challenges. Among those are abbreviations and superscripts, crossed out words with inserted corrections, Roman numerical notation and added words written between the lines.

3 Handwritten text recognition system

The handwritten text recognition (HTR) problem can be formulated as the problem of finding the most likely word sequence, $\mathbf{w} = (w_1\ w_2\ \ldots\ w_l)$, for a given handwritten sentence image represented by a feature vector sequence $\mathbf{x} = (x_1\ x_1\ \ldots\ x_m)$, i.e., $\mathbf{w} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x})$. Using the Bayes' rule we can decompose this probability into two probabilities, $P(\mathbf{x} \mid \mathbf{w})$ and $P(\mathbf{w})$:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) \approx \arg\max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \quad (1)$$

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

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 model training/decoding module, which obtains the most likely word sequence for the sequence of feature vectors (Bazzi et al., 1999; Toselli and others, 2004).

3.1 Preprocessing

As previously said, it is quite common for handwritten documents, and particularly for ancient documents, to suffer from degradation problems (Drida, 2006). 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.

Concerning the preprocessing module used in this paper, the following steps take place: skew correction, background removal and noise reduction, line extraction, slant correction and size normalization. We understand as "skew" the angle between the horizontal direction and the direction of the lines on which the writer aligned the words. Skew correction is carried out on each document page image, by aligning their text lines with the horizontal direction. 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 is based on the horizontal projection profile of the input image. Finally, slant correction and size normalization are applied on each separate line. The slant is the clockwise angle between the vertical direction and the dominant direction of the written vertical strokes. This angle is determined using a method based on vertical projection profile, and used then by the slant correction process to put the written text strokes in an upright position. On the other hand, the size normalization process tries to make the system invariant to character size and to reduce the areas of background pixels which remain on the image because of the ascenders and descenders of some letters. More detailed description can be found in (Toselli and others, 2004; Romero et al., 2006).

3.2 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 $N \times M$ squared cells. In this work, $N = 20$ is chosen empirically and M must satisfy the condition that N/M is equal to the original line image aspect ratio. Each cell is characterized by the following features: *average gray level*, *horizontal component of the grey level gradient* and *vertical component of the grey level gradient*. To obtain smoothed values of these features, an 5×5 cell analysis window, centred 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 60-dimensional feature vectors (20 normalized gray-level components and 20 horizontal and vertical gradient components) is obtained.

3.3 Training and Recognition

Characters are considered here as the basic recognition units and they are modelled by left-to-right HMMs, with 6 states and a mixture of 64 Gaussian densities per state. This Gaussian mixture serves as a probabilistic law to the emission of feature vectors on each model state. The number of Gaussian densities as well as the number of states were empirically chosen after tuning the system. 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 modelled 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 modelled by a bigram *language model*, with Kneser-Ney back-off smoothing (Kneser and Ney, 1995), which uses the previous $n-1$ words to predict the next one:

$$P(\mathbf{w}) \approx \prod_{i=1}^{N} P(w_i|\mathbf{w}_{i-n+1}^{i-1}) \qquad (2)$$

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

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.

4 Evaluation methodology

The handwriting recognition techniques used here require annotated images to train the HMM and the language models. In order to assess how the system accuracy varies with respect to the amount of training data available for training the HTR models and how post-editing the output of the HTR system can save human effort, in this paper we have conducted a study during the real transcription of a marriage license book. In the next subsections the assessment measures, the information of the corpus and the procedure are explained.

4.1 Assessment Measures

Different evaluation measures were adopted to carry out the study. On the one hand, the quality of the automatic transcription can be properly assessed with the well known Word Error Rate (WER). The WER is also a reasonably good estimate of the human effort needed to *post-edit* the output of a HTR recognizer at the word level. It is defined as the minimum number of words that need to be substituted, deleted or inserted to convert a sentence recognized by the HTR system into the reference transcriptions, divided by the total number of words in these transcriptions. On the other hand, in our subjective test with a real user we measured the time needed to fully transcribe each license of the book following the post-editing approach.

4.2 Corpus

The corpus used on the experiments was compiled from a single book of the marriage license books collection conserved at the Archives of the Cathedral of Barcelona. Fig. 2 shows an example of a marriage license page.

The corpus was written by only one writer in 1868 and it was scanned at 300 dpi in true colours and saved in TIFF format. It contains 200 pages although only the firsts 100 have been used in this work. For each page, we used the GIDOC (Serrano et al.,) prototype for text block layout analysis and line segmentation. Concretely, a preliminary detection was performed by a fully automatic process using standard preprocessing techniques based on horizontal and vertical projection profiles. Then, the detected locations for each block and lines were verified by a human expert and corrected if necessary, resulting in a data-set of 2,926 text line images.

The main block of the whole manuscript was transcribed automatically and post-edited line by line by an expert palaeographer during the experimental process. This transcription has been carried out trying to obtain the most detailed transcription possible. That is, the words are transcribed in the same way as they appear on the text, without correcting orthographic mistakes. The book contains around 17k running words from a lexicon of around 3k different words. Table 1 summarizes the basic statistics of the corpus text transcriptions.

To carry out the study we grouped the pages of the document into 5 consecutive partitions of 20 pages each (1-20, 21-40, 41-60, 61-80, 81-100).

Table 2: Basic statistics of the different partitions for the database.

Number of:	P1	P2	P3	P4	P5
Pages	20	20	20	20	20
Lines	581	583	582	584	596
Run. words	3 560	3 523	3 560	3 533	3 615
Characters	19 539	19 234	19 544	19 644	19 818

Table 1: Basic statistics of the text transcriptions.

Number of:	Total
Pages	100
Lines	2,926
Running words	17,791
Lexicon size	2,210
Running characters	97,779
Character set size	84

All the information related with the different partitions is shown in Table 2.

4.3 Procedure

The study consisted in transcribing line by line, by a palaeographer expert, the first 100 pages of the marriage license book presented in the previous subsection. First, partition P1 was manually transcribed by the user without any help of the HTR system. Then, from partition P2 to P5, each partition was automatically transcribed by the system trained with all preceding partitions, which were previously post-edited by the user. This should help in improving the system accuracy.

The experimental process can be summarized in the following steps:

- Initially, the user manually transcribed the first 20 pages of the book (P1).
- The following block was automatically transcribed by the HTR system trained with all preceding transcriptions.
- Each automatically transcribed line was supervised and, if necessary, amended by the expert.
- After processing a block of pages, all supervised transcriptions were used to (re-)train the automatic transcription system.
- The previous 3 steps were iterated until all the blocks were perfectly transcribed.

The manual transcription process and the post-editing process were carried out by means of the GIDOC prototype (Serrano et al.,). In order to avoid possible biases due to human learnability to the user interface, the user became familiar with the engine.

5 Results and Discussion

Table 3 shows the results obtained in the transcription experiments for the different partitions. Column *Time* represents the minutes that the palaeographer needed to post-edit the HTR output of the 20 pages of each partition (except for the first block that was transcribed without any help). For each partition, the HTR system was trained with all the pages in the preceding partitions. The value in parentheses in column *Time* is the average number of minutes that the palaeographer needed to post-edit a marriage license in each partition. Column *WER* is the Word Error Rate for each partition. Note that this value can be interpreted as the percentage of corrections (deletions, insertions and substitutions) that the palaeographer needed to obtain the correct transcription. Column *% Running OOV* is the percentage of out of vocabulary words in each partition that were not observed in the training. The HTR system was not able to provide the correct transcription of these words because they were not included in the language model and therefore these errors were unavoidable for the HTR system. In other words, if we had available a lexicon, then the WER in the OOV words could be reduced at most in the amount represented in the *% Running OOV* column.

Regarding the results, it is important to remark the following issues. First, as expected, the

Table 3: Post-editing results.

	Time	WER	% Running OOV
P1 (1-20)	393 (3.4)	100	100
P2 (21-40)	305 (2.7)	57.7	15.9
P3 (41-60)	235 (2.1)	45.0	11.6
P4 (61-80)	193 (1.7)	35.7	11.6
P5 (81-100)	170 (1.5)	36.2	8.1

WER decreased as the amount of training data increased. In particular, the system achieved around 36% of WER for the last two partitions. This can be also observed in Figure 3, where the results are shown graphically.

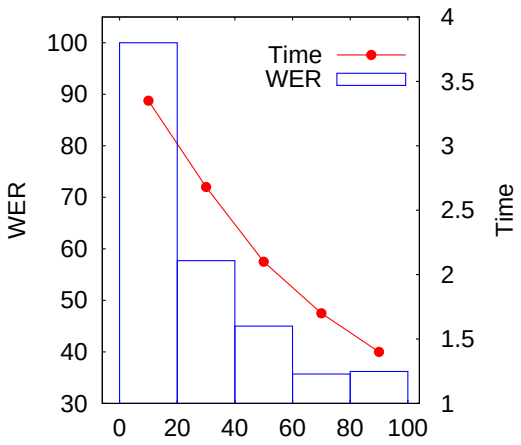

Figure 3: Transcription Word Error Rate (WER) and post-editing time (in minutes) as a function of the block of pages transcribed. For each block the HTR system is trained with all the pages in preceding blocks.

Second, regarding the time required by a human expert to post-edit the transcription proposed by the system, results showed that it became better with the number of partitions already processed. Most specifically, the relative difference between manually transcribe the first block with respect to post-edit the last block is 56%.

It is important to remark that during the transcription process the palaeographer learned to transcribe as new pages were processed. Therefore the time reduction for transcribing could be due to: i) the help of the HTR system, ii) the palaeographer's learning process, or iii) both of them. In order to clarify this issue, we plotted the average time that was needed to transcribe each license by page grouped by partitions, and then we fitted these times to a function. The gradient of this function may be interpreted as the "tendency" of the time needed to transcribe a page (see Figure 4). If the gradient is near to 0, then this could be interpreted as the palaeographer needed similar time to transcribe the initial pages of the partition and the final pages, and therefore the improvements in time are mainly due to the HTR system. This happened in the last two partitions.

In the first partition, the gradient was positive. This may be interpreted as the palaeographer was learning to transcribe and some pages were difficult. In partitions 2 and 3, the gradient was negative; that may be interpreted as the palaeographer was taking profit of the experience acquired in previous pages, and he was learning as new pages were post-edited.

Although we think that there is room for significant improvements, it must be noted that the results are reasonably good for effective post-editing.

6 Conclusions

In this paper we have studied how the accuracy of the HTR systems is reduced with respect to the amount of data used to train the models. In addition, we have also studied how the improvements in the system accuracy affect to the human efficiency following a post-editing approach. The experiments have been carried out with a marriage license book. These documents have interesting information that is being used by demographers that devote a lot of time to transcribe them.

Considering the results obtained in the field study, we can conclude that post-editing the output of an automatic transcription system, significant amounts of human effort can be saved.

Acknowledgements

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 and

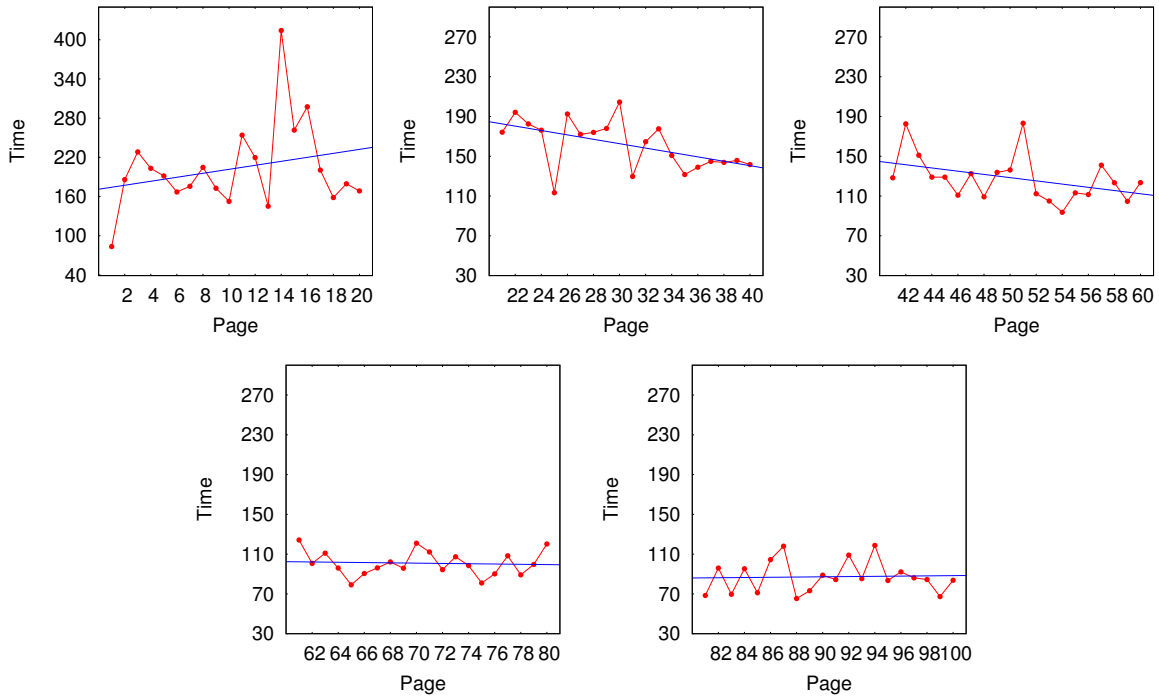

Figure 4: Average time (in seconds) that was needed to post-edit each license by page, and linear function fitted to this time.

by the Universitat Politècnica de València (PAID-05-11).

References

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

F. Drida. 2006. Towards restoring historic documents degraded over time. In *Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pages 350–357. Lyon, France.

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 Pattern Analysis and Machine Intelligence*, 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.

A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

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

E. Kavallieratou and E. Stamatatos. 2006. Improving the quality of degraded document images. In *Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pages 340–349, Washington DC, USA.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. volume 1, pages 181–184, Detroit, USA.

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

V. Romero, M. Pastor, A. H. Toselli, and E. Vidal. 2006. Criteria for handwritten off-line text size normalization. In *Proc. of the 5th Int. Conf. on Visualization, Imaging and Image (VIIP 2006)*, Palma de Mallorca, Spain, August.

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. *Int. Journal on Pattern Recognition and Artificial Intelligence*, 18(4):519–539.