Chapter 1

From Isolated Words to Unconstrained Documents: Bringing Handwriting Recognition to the Meeting Room[1]

1.1 Introduction

The earliest handwriting recognition approaches date back to the eighties, when the first attempts of automatically recognizing handwritten words were proposed, e.g., in Mori et al. (1984), Burr (1983), or Bozinovic and Srihari (1989). However, it is only in the mid nineties that the domain takes off thanks to two main factors (Vinciarelli, 2002): on one hand, the diffusion of cheap image acquisition and storage technologies that made it possible to perform experiments on large databases of handwritten material. On the other hand, the extensive use of handwriting recognition tasks (in particular the automatic transcription of handwritten digits) in the machine learning community (Le Cun et al., 1990; Cortes and Vapnik, 1995).

While not being aimed at the improvement of handwriting recognition technologies - digit recognition was adopted because it was a challenging task for pattern recognition techniques - machine learning works still contributed significantly in terms of methodology. Initially, the focus of handwriting research was on two application domains, namely the recognition of town names in handwritten postal addresses and the transcription of bank-check amounts written in letters (Plamondon and Srihari, 2000).

The two tasks above dominated handwriting recognition research for at

[1]This chapter was written by Alessandro Vinciarelli.

least a decade (1995-2005). The main reason was probably that involved lexica - the lists of possible transcriptions for the handwritten data - were small enough to allow satisfactory performances (10-1000 words). In other words, a-priori constraints on the possible transcriptions of the data were tight enough to make the tasks possible while still being challenging. Furthermore, it was possible to combine the recognition of the words with the recognition of associated handwritten digits (the zip codes for the postal addresses and the amount written in digits for the bank-checks). In this way, the performances could quickly achieve levels sufficient for real-world applications. Not surprisingly, it is in this period (1995-2000) that some of the most important companies selling handwriting recognition products were born like, e.g., Vision Objects[2], A2iA[3], Abbyy[4], etc.

IM2 handwriting efforts started at the birth of the NCCR (2002) as a continuation of pre-existing activities of two partners (Idiap and the University of Bern). Given the large amount of work done on the recognition of words in the previous years, it was difficult to achieve any significant progress on the two areas mentioned above (or any other application domain involving the recognition of isolated words). Hence, IM2 efforts targeted since the beginning the shift from the recognition of isolated words to the transcription of texts. The need for such a step was widely recognized in the handwriting community. Furthermore, the recognition of texts was a crucial need in a meeting scenario where participants take notes and minutes. However at least two problems were challenging for the state-of-the-art of the time, namely the adoption of lexica two order of magnitude larger than those used until that moment (from 100 to 50000 words), and the modeling of word sequences rather than isolated words.

IM2 efforts started with the collection of large databases of handwritten texts (no data of this type was available) and continued with the adaptation of continuous speech recognition technologies to handwritten data. Preliminary works (in collaboration between Idiap and the University of Bern) appeared in 2003 (Vinciarelli et al., 2003) and reached their maturity in the following years (Vinciarelli et al., 2004; Zimmermann et al., 2006). As a side product, it became possible to address problems that were simply not accessible before like, e.g., the application of indexing, retrieval and categorization approaches to handwritten texts (Vinciarelli, 2005a,b), the understanding of witheboard notes (Liwicki and Bunke, 2008) and, more in general, the automatic processing of handwritten documents (Grosicki et al., 2009).

The rest of this paper outlines the contributions of IM2 in more detail: Section 1.2 presents the offline handwriting recognition problem, Section 1.3 describes the main steps in the shift from word to text recognition,

[2] http://www.visionobjects.com

[3] http://www.a2ia.com/Web_Bao/HOMEPAGE-Eng.aspx

[4] http://www.abbyy.com

Section 1.4 shows the most important effects of such a step and the final Section 1.5 draws some conclusions.

1.2 Offline Word Recognition

Offline handwriting recognition is the automatic transcription of handwritten data available as static images (hence the name "*offline*"). Unlike the case of *online* handwriting recognition, any information about the trajectory of the pen is missing and the temporal order of the ink strokes is unknown (Plamondon and Srihari, 2000). The most important steps of the recognition process are shown in Figure 1.1. The goal of the *preprocessing* is to convert input data into a format suitable for further analysis, namely binary images where the handwritten words are the foreground. In some cases, input data is naturally available in such a format (e.g., literary manuscripts where the text was written on white paper). In other cases, it is necessary to perform operations like the *binarization* (automatic thresholding of pixel intensities so that background and foreground become white and black, respectively), the removal of background textures (particularly frequent in the case of bank-checks), the removal of spots or deteriorations (typical in historical documents), etc.

At the end of the preprocessing, the data undergoes the *normalization*, i.e. the removal of *slant*, inclination of the strokes supposed to be vertical, and *slope*, inclination of the word with respect to the horizontal direction (Vinciarelli and Luettin, 2001). A normalized word is expected to be horizontal and both *ascendents* and *descendents*, long strokes supposed to be vertical in letters like "*d*" or "*g*", actually appear to be vertical. While not being supported by any theoretic justification, the normalization was extensively shown to significantly improve the performance of any recognition system (Vinciarelli and Luettin, 2001). The main reason is probably that this step removes variability due to indvidual handwriting style and not to the words being written.

Even though the temporal order of the strokes is not available, it is still possible to assume that the spatial distribution of the foreground pixels accounts for the temporal sequence of the strokes. In other words, the more a stroke is on the right hand side of an image, the later it has probably been written. Hence, handwritten data are typically represented with sequences of feature vectors extracted from word fragments isolated during the *segmentation*. In some cases, the segments are isolated with an *explicit* approach, i.e. an attempt to identify atomic elements that, like phonemes in case of speech, compose any possible character (Bozinovic and Srihari, 1989; Mohamed and Gader, 1996). In other cases, the segmentation is *implict*, i.e. the observations are extracted at regular steps from an area of predefined width typically called *window* (Vinciarelli and Luettin, 2000).

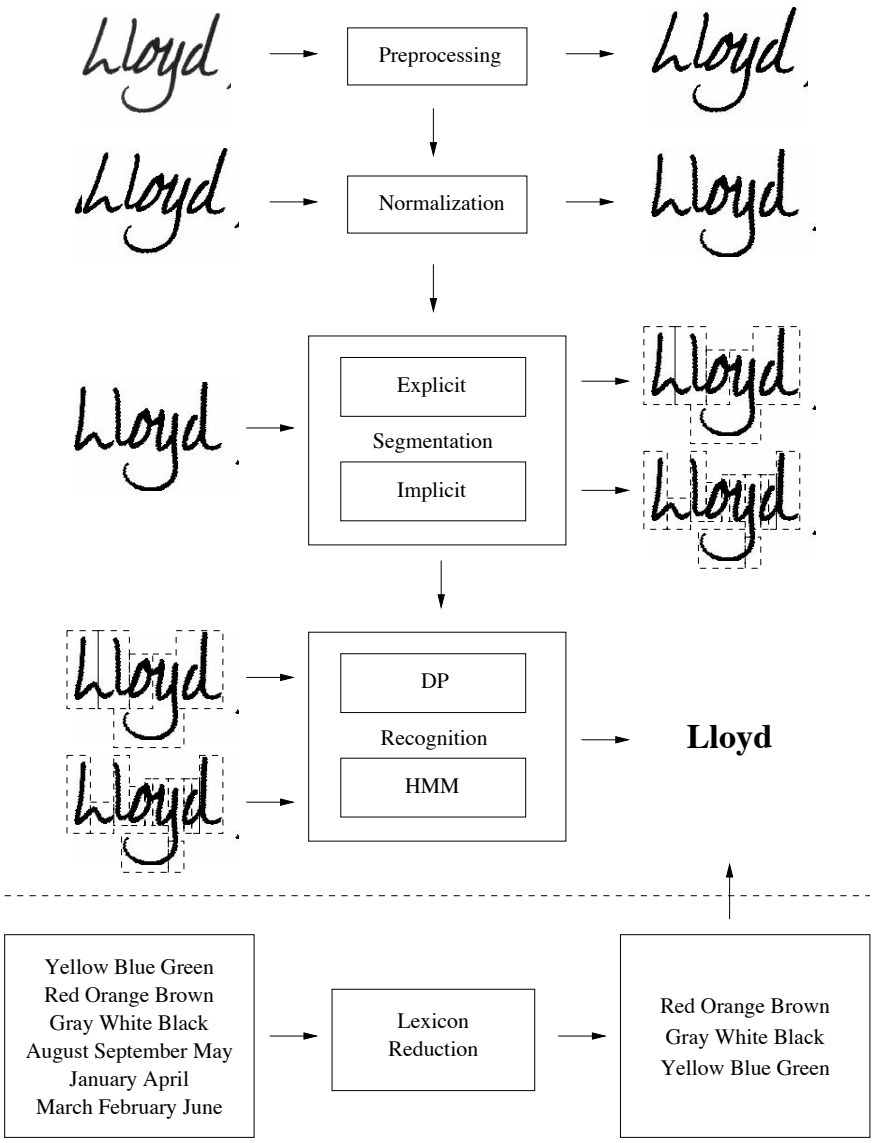

Figure 1.1: General scheme of an Offline Word Recognition system.

At the last step, the *recognition*, the observation sequence is matched with every entry of a *lexicon*, the list of allowed transcriptions. When the segmentation is implict, the matching is typically performed with Hidden Markov Models (or other probability distributions defined over sequential data), when the segmentation is explicit, the most frequent matching approach is Dynamic Programming. Some early approaches include a lexicon reduction step that aims at removing entries incompatible with the "shape" of the handwritten word. For example, the word "Lloyd" in Figure 1.1 has

three ascendents and one descendent. Hence, any word that does not include the same features should be removed from the lexicon (Zimmermann and Mao, 1999).

1.3 From Word to Text Recognition

At the moment IM2 starts, state-of-the-art systems have the structure presented in Section 1.2, with minor variations between one case and the other. The large majority of the works revolves around the recognition of town names in handwritten postal addresses or the transcription of bank-check amounts written in words. In both cases, there is no need for large lexica because a-priori information tightly constrains allowed transcriptions: in the case of postal addresses, the zip code (recognized automatically) limits the number of possible towns between 10 and 1000, with the latter value reached only rarely. In the case of bank-checks, 20-30 lexicon entries are sufficient, depending on the language. Furthermore, for security reasons only amounts up to a certain value are processed automatically and this further limits the size of the lexicon (e.g., no more than $10,000$ Dollars in the United States).

The earliest attempts to move towards the recognition of unconstrained texts were still based on the hypothesis that each word can be recognized separately (the difficulty of segmenting texts into words was never taken into account) like, e.g., in Senior and Robinson (1998). Moreover, since no language-models were adopted, it was not possible to use lexica larger than 1000 words, not enough for the recognition of realistic unconstrained texts. It is at this point that two crucial resources available in IM2 make it possible to move from the recognition of words to the recognition of texts, namely the speech recognition expertise available at Idiap, in particular for what concerns decoding techniques and statistical language modeling (Moore et al., 2006), and the first databases of unconstrained handwritten texts collected at Idiap and at the University of Bern (Marti and Bunke, 2002).

1.3.1 The Data

The first, publicly available database of handwritten texts - the "*Cambridge*" database - was collected at the in the mid nineties and it was the transcription of a document from the Brown Corpus, a collection of texts supposed to be representative of standard written English, Senior and Robinson (1998). In this respect, the corpus was a novelty with respect to the data used until that moment. However, it was relatively small (353 lines) and, most importantly, it included samples by one writer only.

Hence, the first, real corpus of handwritten texts was the "*IAM*" database, a collection of 928 lines written by roughly 400 persons (transcriptions of documents belonging to the "*Lancaster Oslo Bergen*" Corpus), Marti and Bunke (2002). The IAM database made it possible for the first time to work

on realistic, unconstrained texts collected with linguistically oriented criteria. In this way, it was possible for the first time, at least in the handwriting community, to include statistical language modeling in the recognition process.

Later, another corpus, the "*Reuters*" database, was collected at Idiap with the purpose of going beyond the symple recognition of texts and to perform content analysis as well. In particular, the data collected at Idiap (803 lines written by a single individual) were transcriptions of the Reuters Corpus, a collection of digital texts aimed at text categorization and information retrieval. This collection was important not only to develop handwriting recognition approaches, but also to develop handwritten document processing techniques (see Section 1.4), Vinciarelli (2005a).

1.3.2 Decoding Techniques and Language Modeling

The recognition of handwritten texts posed two challenges that were virtually unknown in the handwriting community, the adoption of statistical language models and the development of efficient decoding techniques capable of dealing with large search spaces. It is at this stage that the collaboration within IM2 gave the most important results for handwriting. In fact, IM2 speech researchers were familiar with both problems and did not hesitate to share their expertise and resources. This made it possible to complete the first recognizer of unconstrained texts in a relatively short time and to publish the first, important results in a mere two years after IM2 started (see below).

For what concerns the language models, the choice was to adopt N-grams, statistical models that estimate the probability of a word sequence $W = w_1, \ldots, w_T$ as follows (Rosenfeld, 1996; Katz, 1987):

$$p(W) = \prod_{k=1}^{T} p(w_k | w_{k-1}, \ldots, w_{k-N+1}), \tag{1.1}$$

where N, the order of the model, is usually 2 or 3 depending in the amount of available text training data. While not including any linguistic knowledge (N-grams can be applied to any sequence of symbols belonging to a finite set), N-grams were the most effective language models available and, most importantly, they were easy to plug in decoding techniques.

The decoding can be thought of as the search of the path that better matches the handwritten data (represented as a sequence of vectors as explained in Section 1.2) in a search space. In the handwriting case, the latter is the space of all sequences of word models, where each model is typically a HMM. This latter estimates the probability of a certain sequence of feature vectors being extracted from a certain word. The main challenge is that the search space can be very large: if the lexicon includes L entries and a

sentence contains T words, then the number of possible word sequences - hence, the number of possible paths - is V^T, a very high number even for limited values of V and T.

In the same period as handwriting researchers were moving from the recognition of words to the recognition of texts, the Idiap speech group was developing Juicer, a decoding approach based on Weighted Finite State Transducers, the state-of-the-art at that moment (Moore et al., 2006). The major advantage of Juicer was that it allowed transcriptions of the data based on units different from phonemes. This made it possible to apply the decoder to the handwriting problem and to perform the first experiments aimed at the recognition of handwritten texts.

1.3.3 Experiments

The experiments built upon technology developed at Idiap since 1999 and addressing all aspects of handwriting recognition, from normalization (Vinciarelli and Luettin, 2001), to feature extraction (Camastra and Vinciarelli, 2001, 2003; Vinciarelli and Bengio, 2002a), to statistical modeling of observation sequences (Vinciarelli and Luettin, 2000; Vinciarelli and Bengio, 2002b). The decoding technologies developed in the speech group made it possible to complete the recognition process in view of the automatic transcription of texts. The results were published between 2003 and 2004 in two papers that are still today cited frequently in the literature (more than 200 citations in total at the moment this paper is being written). All corpora mentioned in Section 1.3.1 were used for the tests resulting into one of the most extensive eperimentations performed until that moment in the handwriting community (Vinciarelli et al., 2003, 2004).

The experiments aimed at measuring the recognition rate as a function of three main parameters, namely the presence or the absence of a language model, the order ofthe language model (when used) and the size of the lexicon. The results are plotted in Figure 1.2 and show how the adoption of a language model increases, to a statistically significant extent, the performance of the recognizer. On the other hand, the difference between models of different order is not large. The reason is that handwritten texts were recognized, at least in the experiments presented in the paper, line by line. Since an average handwritten line contained between 6 and 9 words, only a limited fraction of the data (between 50 and 66%) could actually benefit from an increased order of the model.

One of the most impressive differences with respect to the state-of-the-art was the size of the lexica. Before the adoption of language models, it was hardly possible to go above 100 words while the experiments in Vinciarelli et al. (2004) were performed using lexica including up to $5 \cdot 10^4$ entries. Unconstained texts were way more challenging than any data considered before (addresses and bank-checks), but language models provided a-priori

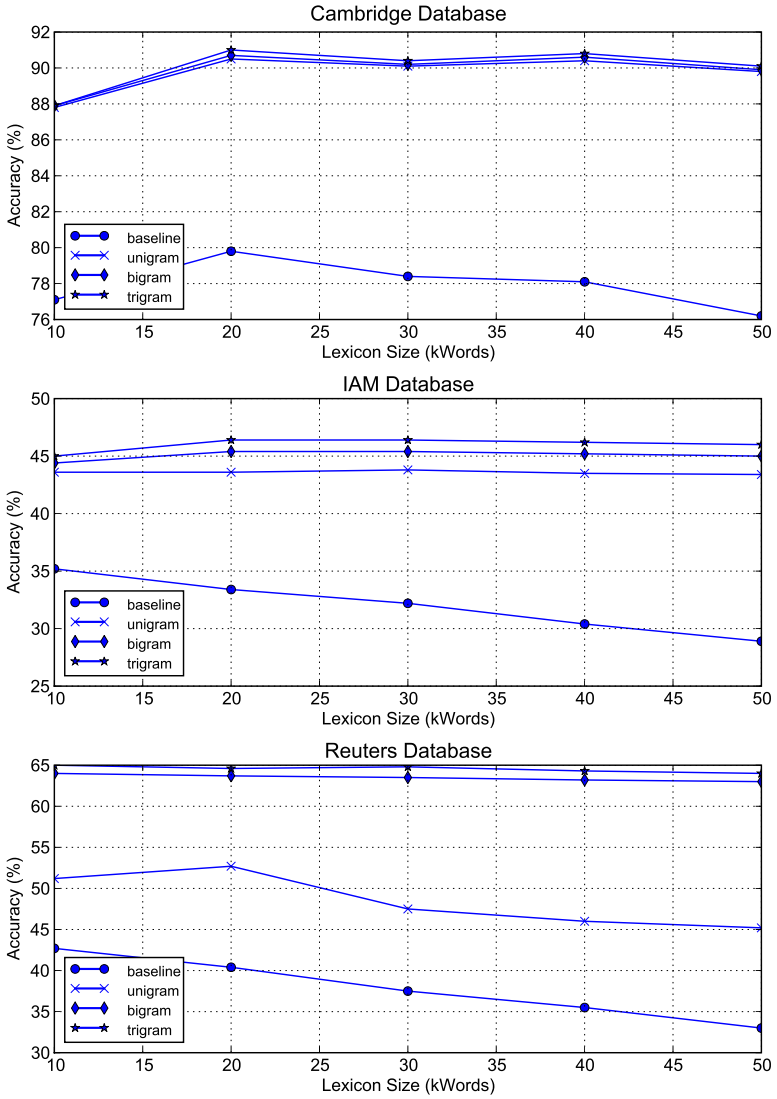

Figure 1.2: Accuracy over Cambridge, IAM abd Reuters database.

information sufficiently constraining to allow the use of very large lexica.

1.4 From Text to Documents

Recognizing texts rather than words opened the way towards a further improvement of the state-of-the-art, that is the processing of entire handwritten documents and, in particular, the application of indexing and retrieval techniques commonly applied to digital texts. The main question was whether text processing techniques were sufficiently robust to deal with handwriting

recognition errors.

Idiap efforts concentrated on two main problems: handwritten document retrieval and handwritten document categorization. In the former task, the goal is to identify documents relevant to a user query out of a database of available texts. In the latter case, the goal is to assign each document to one or more predefined categories. In this endeavour as well, research on handwriting benefited from expertise available in IM2. In particular, spoken document retrieval activities carried out at Idiap in a European project[5] helped to shape the first experiments in handwritten document retrieval. Results on speech were showing that high error rates (up to 50%) did not have major effect on the performance of retrieval technologies. Hence, the same result was likely to be observed for handwritten data as well.

The results along this line of research were published between 2004 and 2005 and substantially confirmed the experiments performed on spoken documents (Vinciarelli, 2005a,b). While the recognition rate was low (less than 50% in some cases), the decrease in categorization and retrieval performances with respect to a manual, error free transcription of the data was negligible. This activity this did not result only in papers and publications, but also in a workshop organized for several years by IBM[6] and a large French project aimed at collecting a database of several thousands of handwritten letters[7]. In both cases, the organizers explicitly recognized the influence of the Idiap contributions on the design of their respective initiatives.

1.5 Conclusions

This article has outlined the handwriting activities conducted at Idiap in the framework of IM2. The adoption of meeting based scenarios has helped to make a significant step with respect to the state-of-the-art of the time, namely the shift from the recognition of isolated words to the recognition of unconstrained texts. Such an important breakthrough has been made possible by the collaboration between different IM2 sites, in particular Idiap and the University of Bern. In this respect, the role of IM2 as a facilitator of interdiscplinary exchanges has been crucial. In the years that followed the latest IM2 contributions to handwriting (2005 to present), the community has moved towards the application of handwriting technologies to different real-world domains. Nowadays, experiments are performed over handwritten notes collected at meetings or in the classroom, data written on whiteboards, mathematical equations, musical scores, etc. In parallel, a few companies have made of handwriting recognition their core business and capitalize on the work done in the scientific community to endow machines with the ability

[5] http://spandh.dcs.shef.ac.uk/projects/thisl/overview-oct98/

[6] The AND workshop: https://sites.google.com/site/and2010workshop/.

[7] Rimes Project: http://rimes.it-sudparis.eu

of reading handwritten information (Vision Objects, Abbyy, A2ia, etc.).

Current research in handwriting recognition is still inspired by speech recognition like, e.g., the extensive works on the application of tri-phones (applied to letters rather than to phonemes) in Bianne-Bernard et al. (2011). In this respect as well, IM2 seems to have played a pioneering role. On the other hand, debates on whether handwriting is still a necessary skill when portable devices allow one to easily write with a keyboard are taking place[8]. Therefore, it is difficult to imagine for how long handwriting recognition will still be a topic of interest in the scientific community. From this point of view, it should be noted that many people predicted the end of books (and any form of printed information) when the web started to pervade our lifes. Roughly twenty years after, books are still being printed and produced in increasing quantities, then the end of handwriting might still be very far.

[8] http://www.wisegeek.com/should-people-still-use-cursive-writing.htm

Bibliography

Bianne-Bernard, A., Menasri, F., Mohamad, R., Mokbel, C., Kermorvant, C., and Likforman-Sulem, L. (2011). Dynamic and contextual information in hmm modeling for handwritten word recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2066–2080.

Bozinovic, R. and Srihari, S. (1989). Offline cursive script word recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(1):69–83.

Burr, D. (1983). Designing a handwriting reader. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(5):554–559.

Camastra, F. and Vinciarelli, A. (2001). Cursive character recognition by learning vector quantization. *Pattern Recognition Letters*, 22(6-7):625–629.

Camastra, F. and Vinciarelli, A. (2003). Combining neural gas and learning vector quantization for cursive character recognition. *Neurocomputing*, 51:147–159.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Grosicki, E., Carree, M., Brodin, J., and Geoffrois, E. (2009). Results of the rimes evaluation campaign for handwritten mail processing. In *Proceeedings of the International Conference on Document Analysis and Recognition*, pages 941–945.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

Le Cun, Y., Boser, B., Denker, J., Howard, R., Habbard, W., Jackel, L., and Henderson, D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404.

Liwicki, M. and Bunke, H. (2008). *Recognition of Whiteboard Notes: Online, Offline and Combination*. World Scientific Publishing.

Marti, U. and Bunke, H. (2002). The IAM-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Mohamed, M. and Gader, P. (1996). Handwritten word recognition using segmentation-free Hidden Markov modeling and segmentation-based dynamic programming techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):548–554.

Moore, D., Dines, J., Doss, M., Vepa, J., Cheng, O., and Hain, T. (2006). Juicer: A weighted finite-state transducer speech decoder. In *Proceedings of Machine Learning for Multimodal Interaction*, pages 285–296.

Mori, S., Yamamoto, K., and Yasuda, M. (1984). Research on machine recognition of handprinted characters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(4):386–405.

Plamondon, R. and Srihari, S. (2000). On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modelling. *Computer speech and language*, 10(3):187.

Senior, A. and Robinson, A. (1998). An off-line cursive handwriting recognition system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):309–321.

Vinciarelli, A. (2002). A survey of off-line cursive script recognition. *Pattern Recognition*, 35(7):1433–1446.

Vinciarelli, A. (2005a). Application of information retrieval techniques to single writer documents. *Pattern Recognition Letters*, 26(14-15):2262–2271.

Vinciarelli, A. (2005b). Noisy text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1882–1895.

Vinciarelli, A. and Bengio, S. (2002a). Offline cursive word recognition using continuous density HMMs trained with PCA or ICA features. In *Proceeedings of the International Conference on Pattern Recognition*, pages 81–84.

Vinciarelli, A. and Bengio, S. (2002b). Writer adaptation techniques in HMM based off-line cursive script recognition. *Pattern Recognition Letters*, 23(8):905–916.

Vinciarelli, A., Bengio, S., and Bunke, H. (2003). Offline recognition of large vocabulary cursive handwritten text. In *Proceedings of 7th IEEE International Conference on Document Analysis and Recognition*, pages 1101–1105.

Vinciarelli, A., Bengio, S., and Bunke, H. (2004). Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):709–720.

Vinciarelli, A. and Luettin, J. (2000). Off-line cursive script recognition based on continuous density HMMs. In *Proceedings of 7th International Workshop on Frontiers in Handwriting Recognition*, pages 493–498.

Vinciarelli, A. and Luettin, J. (2001). A new normalization technique for cursive handwritten words. *Pattern Recognition Letters*, 22(9):1043–1050.

Zimmermann, M., Chappelier, J., and Bunke, H. (2006). Offline grammar-based recognition of handwritten sentences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):818–821.

Zimmermann, M. and Mao, J. (1999). Lexicon reduction using key characters in cursive handwritten words. *Pattern Recognition Letters*, 20(11):1297–1304.