A Multimodal Approach to Dictation of Handwritten Historical Documents

Vicent Alabau, Verónica Romero, Antonio-L. Lagarda, Carlos-D. Martínez-Hinarejos

Institut Tecnològic d'Informàtica, Universitat Politècnica de València
Camino de Vera, s/n, 46022, Valencia, Spain
{valabau,vromero,alagarda,cmartine}@iti.upv.es

Abstract

Handwritten Text Recognition is a problem that has gained attention in the last years due to the interest in the transcription of historical documents. Handwritten Text Recognition employs models that are similar to those employed in Automatic Speech Recognition (Hidden Markov Models and n-grams). Dictation of the contents of the document is an alternative to text recognition. In this work, we explore the performance of a Handwritten Text Recognition system against that of two speech dictation systems: a non-multimodal system that only uses speech and a multimodal system that performs a text recognition which is used in the posterior speech recognition. Results show that the multimodal combination outperforms any of the other considered non-multimodal systems.

Index Terms: speech recognition, dictation, language modelling, handwritten text recognition

1. Introduction

In the last years, many on-line archives and digital libraries are publishing large quantities of digitised legacy documents. These documents must be transcribed into an appropriate textual electronic format in order to allow text-based search of their contents and provide historians and other researchers new ways of indexing, consulting and querying their contents. However, the vast majority of these documents (hundreds of terabytes of digital image data) remain waiting to be transcribed into a textual electronic format. Therefore, manual transcription of these documents is an important task for making available the contents of digital libraries.

These transcriptions are usually carried out by experts in paleography, who are specialised in reading ancient scripts. These scripts are characterised by different handwritten/printed styles from diverse places and time periods. The time that takes for an expert to make a transcription of one of these documents depends on their skills and experience. Most paleographs agree that each page needs several hours to be transcribed.

In this context, Handwritten Text Recognition (HTR) [1] has become an important research topic. HTR tries to obtain the word sequence contained in the image of a handwritten text line. This process needs a previous detection of lines of text in an image, as well as some preprocessing steps to make the handwritten text more regular. The final result is a sequence of words (transcription) of the text line, that may contain errors. When the rate of errors of the transcription is low enough, HTR can be a very useful tool to speed up the transcription of handwritten text documents.

However, when consulting paleographs on the most comfortable method to transcribe a handwritten text document, many of them claim that a dictation of the words is the best option. Consequently, Automatic Speech Recognition (ASR) systems are an important alternative to HTR systems. In addition, the current state-of-the-art ASR and HTR systems share many features: Hidden Markov Models (HMM) [2, 3] are used to model the basic elements of the signal (sounds for speech, strokes for handwritten text) and n-grams language models (LM) are used to model word sequences [2]. From this viewpoint, HTR systems fit in the Natural Language Processing paradigm. Therefore, many features that are usual to ASR systems (such as the use of training data for HMM and n-grams) are common to HTR systems as well.

The similarities between the two types of systems make possible to combine them easily into a multimodal system that may obtain a more reliable final hypothesis, since two different data sources (handwritten text and speech) can be used. In fact, previous attemps in combining handwritten input and speech input have been done [4], but most of them center in the use of on-line handwritten text. In this work, we compare the use of speech dictation to transcribe handwritten text documents against the direct use of text recognition. Speech dictation is developed in non-multimodal (when only an ASR system is available) and multimodal (when both HTR and ASR systems are available) scenarios. We will show that using an initial HTR recognition allows to restrict the set of ASR hypothesis and obtain better results than only using text recognition or plain speech dictation.

The paper is organised as follows: Section 2 describes the fundamentals of a HTR system, Section 3 explains the use of the HTR decoding to improve the ASR recognition, Section 4 summarises the experimental set-up, Section 5 shows the results, and Section 6 provides the main conclusions and future work lines in this field.

2. Handwritten text recognition

The HTR problem can be formulated as the problem of finding the most likely word sequence, $\boldsymbol{w} = (w_1, w_2, \ldots, w_{|\boldsymbol{w}|})$, for a given handwritten sentence image represented by a feature vector sequence, $\boldsymbol{x} = (x_1, x_2, \ldots, x_{|\boldsymbol{x}|})$:

$$\hat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{w}} P(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w}) \quad (1)$$

$P(\boldsymbol{x}|\boldsymbol{w})$ is typically approximated by concatenated character models, usually HMMs, and $P(\boldsymbol{w})$ is approximated by a word LM, usually n-grams [2]. HMMs are used in the same way as they are used in the current ASR systems [3]. The most important differences lay in the type of input sequences of feature vectors: while in the case of ASR they represent acoustic data, the input sequences for off-line HTR are line-image features. Figure 1 shows an example of how a HMM models two feature vector subsequences pertaining to the character "a".

28 – 31 August 2011, Florence, Italy

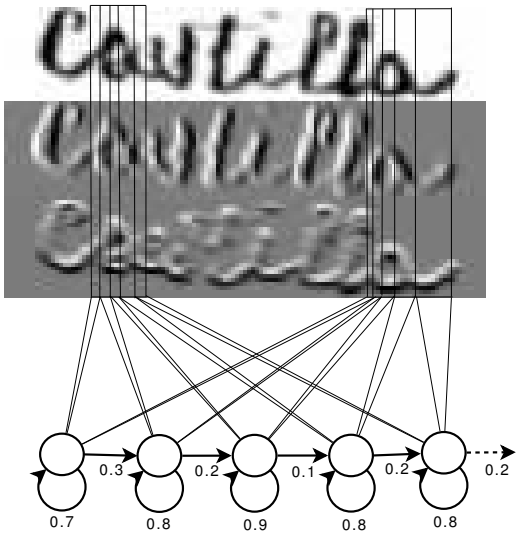

Figure 1: Example of 5-states HMM modelling (sequences of feature vectors of) instances of the character "a" within the Spanish word "Castilla". The states are shared among all instances of characters of the same class.

The HTR system used here follows the classical architecture composed of three main modules: document image preprocessing, line image feature extraction and HMM training/decoding [1].

The following steps take place in the preprocessing module. First, the skew of each page is corrected; we understand "skew" as 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, whose output is then fed to the text line extraction process which divides it into separate text lines images. Finally, slant correction and size normalisation are applied on each separated line. A more detailed description of the preprocess can be found in [1, 5].

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. From each cell, three features are calculated: normalised gray level, horizontal gray level derivative and vertical gray level derivative. The way these three features are determined is described in [1]. 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 $3N$-dimensional feature vectors is obtained. In Figure 1 an example of the sequence of feature vectors for the word "Castilla" is shown graphically.

3. Dictation of handwritten documents

Consulted paleographs agreed that an interesting way of transcribing handwritten text documents would be by dictating the document's content. Basically, the problem consists in obtaining a sequence of words $\boldsymbol{w}$ which is, at the same time, a transcription of the handwritten text $\boldsymbol{x}$ (from the HTR problem) and a speech utterance $\boldsymbol{s} = (s_1, s_2, \ldots, s_{|\boldsymbol{s}|})$. Statistically, this problem can be formulated as

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{s}) = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{w}) P(\boldsymbol{w}|\boldsymbol{x}) \quad (2)$$

Making the safe assumption that $P(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{w})$ is independent of $\boldsymbol{x}$ Eq. 2 can be rewritten as

$$\hat{\boldsymbol{w}} \approx \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{s}|\boldsymbol{w}) P(\boldsymbol{w}|\boldsymbol{x}) \quad (3)$$

where $P(\boldsymbol{s}|\boldsymbol{w})$ is a conventional acoustic-phonetic HMM for speech recognition, and $P(\boldsymbol{w}|\boldsymbol{x})$ is a LM conditioned on $\boldsymbol{x}$. Note that if $\boldsymbol{x}$ is dropped, the LM can be approximated by a standard n-gram LM, $P_N(\boldsymbol{w})$. In that case, Eq. 3 can be decoded with a state-of-the-art ASR system. However, a more interesting approach would be to take advantage of the information given by $\boldsymbol{x}$.

Although in principle an integrated decoding could be possible [6], it would require a specific training and decoding. This is especially complicated since both input signals have different lengths and they are not synchronised. A possible alternative is based on a semi-coupled approximation, in which Eq. 1 can be transformed, after a HTR decoding, into a statistical LM that can be used with current speech recognisers.

A procedure to compute a n-gram LM which is conditioned on $\boldsymbol{x}$ from the posterior probabilities of Eq. 1 is the following. First, a subset of the search space can be obtained in the HTR search process as a word lattice [7]. A word lattice L is a directed, acyclic, weighted graph with an initial node q_I and a final node q_F. The nodes correspond to discrete points in time (or horizontal space, in the case of HTR) conditioned on a specific state of the LM. A link l is defined as any edge between two nodes; each link has associated a starting node $s(l)$, an end node $e(l)$, a hypothesis word $w(l)$, and a likelihood $f(l)$; each link can be considered as a hypothesis $w(l)$ between the nodes $s(l)$ and $e(l)$ with likelihood $f(l)$.

To obtain a n-gram model from a word lattice, posterior probabilities for each node and link must be computed. These probabilities are based on the forward and backward probabilities of the nodes. The forward probability $\alpha(q)$ can be defined as the sum of the probability of all the prefix paths in the lattice reaching a node q from q_I. Correspondingly, the backward probability $\beta(q)$ can be defined as probability of all the suffix paths in the lattice reaching q_F from q. These probabilities can be efficiently computed with the well-known *forward-backward* algorithm [8].

The posterior probability for a specific link l can be computed by summing up the posterior probabilities of all hypotheses of the word lattice containing it:

$$P(l \mid \boldsymbol{x}) = \frac{\alpha\left(s(l)\right) \cdot f(l) \cdot \beta\left(e(l)\right)}{\alpha(q_F)} \quad (4)$$

Similarly, the posterior probability for a specific node q is

$$P(q \mid \boldsymbol{x}) = \frac{\alpha\left(q\right) \cdot \beta\left(q\right)}{\alpha(q_F)} \quad (5)$$

Then, the expected count for a word sequence $w_{i-n+1}^{i} = (w_{i-n+1}, \ldots, w_i)$ can be estimated efficiently as in [9]:

$$C^*(w_{i-n+1}^{i} \mid \boldsymbol{x}) = \sum_{l_1^n \in N(w_{i-n+1}^{i})} \frac{\prod_k P(l_k|\boldsymbol{x})}{\prod_k P(s(l_k)|\boldsymbol{x})} \quad (6)$$

where $N(w_{i-n+1}^{i})$ are all the sequences of concatenated links in L generating w_{i-n+1}^{i}.

Now, word posterior probabilities can be calculated after a proper normalisation:

$$P_L(w_i|w_{i-n+1}^{i-1}, \boldsymbol{x}) = \frac{C^*(w_{i-n+1}^{i} \mid \boldsymbol{x})}{C^*(w_{i-n+1}^{i-1} \mid \boldsymbol{x})} \quad (7)$$

This simple estimation presents two problems: no back-off is included, and only words present in the lattice are included into the model (which implies a high number of out-of-vocabulary words, since lattices only contain the words of the most likely hypotheses). The estimation of the back-off probabilities for the n-gram is obtained by applying a suitable discount method before normalisation. The out-of-vocabulary (OOV) problem is solved by equally distributing among the OOV words the discounted probability mass of the 1-gram.

The resulting LM can be defined as

$$P_L(\boldsymbol{w}|\boldsymbol{x}) = \prod_i P_L(w_i|w_{i-n+1}^{i-1}, \boldsymbol{x}) \tag{8}$$

Finally, in order to avoid poor estimations of $P_L(\boldsymbol{w}|\boldsymbol{x})$, a linear interpolation with the original LM probability $P_N(\boldsymbol{w})$ can be used as well

$$P_\gamma(\boldsymbol{w}|\boldsymbol{x}) = \gamma P_L(\boldsymbol{w}|\boldsymbol{x}) + (1-\gamma)P_N(\boldsymbol{w}) \tag{9}$$

4. Experimental framework

4.1. Corpora

The experiments have been carried out using a Spanish corpus compiled from the legacy handwriting document from the nineteenth century identified as "Cristo-Salvador" (CS), which was kindly provided by the *Biblioteca Valenciana Digital (BIVALDI)*[1]. This corpus is composed of 53 text page images, written by only one writer. 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.

The 33 initial pages, that correspond to the training set in the partition called "book" in [5], were used to train the models. This set is composed by 675 lines with $6,432$ running words and $2,277$ different words. On the other hand, to test the dictation method proposed in this paper we have used the page 41, which is the page with most similar error distribution with respect to the global error. It is important to remark that this corpus has a quite small training ratio (around 2.8 training running words per lexicon-entry). This is expected to result in under-trained LMs, which will clearly increase the difficulty of the recognition task for the system.

In order to assess the speech dictation systems five different users dictated the selected page line by line. It resulted in a test data-set composed by 120 dictated lines.

4.2. Models

As was mentioned above, the recognition process is based on HMMs. For the HTR system, the characters are modelled by continuous density left-to-right HMMs with 12 states and 32 Gaussian mixture components per state. The optimal number of HMM states and Gaussian densities per state were tuned empirically. Speech models are HMM with three states, with left-to-right with loops topology and 64 Gaussians per state (a total number of 4.6K Gaussians). Each speech model represents a phonetically context-independent unit (monophone). These models were estimated using the data of the Albayzin Spanish speech corpus [10], of about 4 hours of speech signal. In the two systems, each lexical word is modelled by a stochastic finite-state automaton (SFSA), which represents all possible concatenations of individual characters or phonemes to compose the word. On the other hand, text lines are modelled using 2-grams with Kneser-Ney back-off smoothing [11] directly estimated from the training transcriptions of the text line images.

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

4.3. Evaluation metrics

Our system was assessed by means of *word error rate* (WER), which obtains the ratio between the number of editions of the Levenshtein distance and the number of words in the reference. Similarly, *oracle* WER is the best WER that can be obtained from the word lattice resulting from the decoding process.

In addition, in order to evaluate the quality of the chosen LM, we have employed perplexity [12]. Perplexity, measured for a text with respect to a LM, is a function of the likelihood of that text being produced by repeated application of the model.

Finally, significance of our results has been assessed by the *paired bootstrap resampling* method, described in [13]. This technique compares two systems and finds out whether one of them significantly outperforms the other one.

5. Results

This section is devoted to analyse the experimental results of the methods proposed in Section 3. First of all, it should be noted that, despite being CS a small corpus to what the speech community is used to, CS is a realistic example of what can be found in transcription of historical documents.

There are some characteristics of this kind of tasks that must be explained. On the one hand, the topic addressed in CS is a very specific one. Since the training corpus is rather small ($6.4k$ running words), LMs are poorly estimated. This is reflected in the perplexity for the test page (552 for a bigram). Higher order n-gram models cannot improve perplexity since segments longer than 2 words rarely occur more than once. Furthermore, as far as we know, there are no other electronic texts dealing with the same topic, and consequently no robust LMs can be estimated. As a result, both HTR and ASR baseline systems must rely more on the good estimation of the HMM models.

On the other hand, each book presents a particular handwriting style which not only depends on the author, but on the period of the history the book was written. This makes very complicated to estimate generic book independent HMM models. In fact, the usual approach is to take part of the book for training and the rest for test. However, ASR HMMs are usually speaker independent.

Two baseline systems have been considered. The first is to transcribe the page using a HTR system. To do this the page must be digitised, the noise must be reduced and the lines segmented. This process is partially manual so it must be considered when evaluating the convenience of using this approach. A 2-gram LM was used in the HTR system. In the second baseline the test is read aloud and the transcriptions come from a dictation ASR system. This system uses the same LM as the HTR system (that is why the perplexities coincide). The results are summarised in Table 1 and show that the HTR system outperforms the ASR system. The explanation for this comes naturally from the previous comments. The LM is poorly estimated and the search process depends greatly on the HMM estimates in both cases. Nevertheless, in the HTR case the HMMs have been specifically trained for the particularities of the test (book and writer), whereas the ASR HMMs were trained from a completely different corpus (distinct speakers).

An intermediate approach is to use information from both

Table 1: Summary of perplexity and WER for the different approaches to transcription of handwritten historical documents.

model	LM	γ	perplexity	WER
HTR	2-gram	—	552	29.2 ± 8.2
ASR	2-gram	0.0	552	43.4 ± 4.0
SHR	Eq. 8 (3gr)	1.0	391	45.8 ± 4.0
SHRi	Eq. 9 (3gr)	0.2	**54**	$\mathbf{18.6 \pm 2.8}$

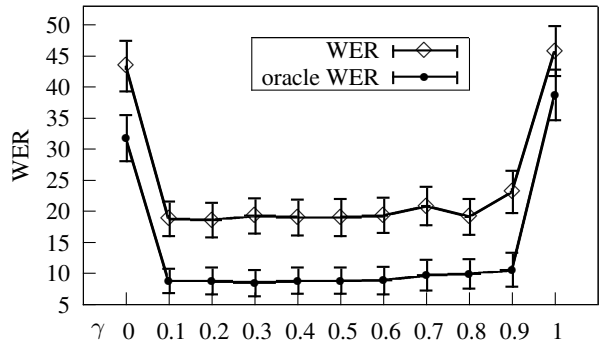

Figure 2: WER and oracle WER for dictation systems.The first value is an ASR system ($\gamma = 0$), the last a SHR system ($\gamma = 1$). The values within are SHRi systems with γ interpolation factor.

the handwritten text and the speech signal by means of the HTR posterior n-grams of Eq. 8. This system, referred to as Speech and Handwriting Recognition (SHR), follows a dictation scenario, as in ASR, but fusing the information from a previous HTR recognition. The parameters needed for this model were estimated using a *leaving-one-out* scheme over the lines of the page. Although it has quite a better perplexity than the original 2-gram model, this system achieves worse WER results. Nevertheless, confidence intervals at 95% overlap with the ASR system. This high WER is mainly due to the poor smoothing for OOV words when computing the HTR posterior n-grams. HTR lattices contain only a small part of the vocabulary so the rest of the vocabulary was introduced with equal probability (see end of Section 3). Thus, the probabilities for these words are low and the recognition performance decreases for them.

To prevent from poor estimations of OOV in the HTR posterior 3-grams, this model has been linearly interpolated with the baseline 2-gram as in Eq. 9, and it is referred to as SHRi. Figure 2 shows the results when changing the value of the interpolation factor. The ASR baseline is represented by $\gamma = 0$ while the HTR posterior 3-gram system is $\gamma = 1$. The graph shows the WER with confidence intervals at 95% along with the oracle WER. The scale factors where estimated in a *leaving-one-out* scheme over individual utterances. All the interpolated models improve the baseline with 100% *probability of improvement* (POI). It must be noted that almost all set-ups perform in the same range, although when γ approaches 1, the curve slowly raises. However, confidence intervals still overlap among interpolated models. The same behaviour can be observed on the *oracle* WER. Best oracle WER achieves an 8.5%, which suggests that there is still room for improvement.

6. Conclusions

In this paper we have tackled the transcription of handwritten historical documents from a multimodal perspective. As comparison, two uni-modal baseline systems have been used: a HTR system to transcribe handwritten text images and an ASR system to transcribe dictations of the documents. Our approach successfully integrates the two sources of information to achieve remarkable improvements over both baselines. However, manual post-editing is still necessary to obtain high quality transcriptions. Hopefully, the results indicate that there is room for reduction of the WER. First, it would be interesting integrate the decoding of text images and speech. Finally, a simpler option would be to compute ASR posterior n-grams to perform HTR recognition, and embed it into an iterative procedure.

7. Acknowledgements

Work supported by the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), iTrans2 (TIN2009-14511) and MITTRAL (TIN2009-14633-C03-01) projects. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant GV/2010/067, and by the UPV under grant UPV/2009/2851.

8. References

[1] A.H.Toselli, A. Juan, J. González, I. Salvador, E. Vidal, F. Casacuberta, D. Keysers, and H. Ney, "Integrated Handwriting Recognition and Interpretation using Finite-State Models," *IJPRAI*, vol. 18, no. 4, pp. 519–539, 2004.

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

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

[4] P. Liu and F. Soong, "Graph-based partial hypothesis fusion for pen-aided speech input," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 478 –485, march 2009.

[5] V. Romero, A. H. Toselli, L. Rodríguez, and E. Vidal, "Computer Assisted Transcription for Ancient Text Images," in *ICIAR 2007*, ser. LNCS. Montreal (Canada): Springer-Verlag, August 2007, vol. 4633, pp. 1182–1193.

[6] S. Bengio, "Multimodal speech processing using asynchronous hidden markov models," *Information Fusion*, vol. 5, no. 2, pp. 81 – 89, 2004.

[7] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, Jan. 1997.

[8] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 3, Mar 2001.

[9] W. Campbell and F. Richardson, "Discriminative keyword selection using support vector machines," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 209–216.

[10] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mariño, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proceedings of EuroSpeech'93*, Berlin, Germany, sep 1993, pp. 175–178.

[11] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of ICASSP 95*, vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, 1995, pp. 181–184.

[12] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" in *Proceedings of the IEEE*, vol. 88, 2000, pp. 1270–1278.

[13] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *Proceedings of ICASSP2004*, vol. 1, Montréal, Canada, May 2004, pp. 409–412.