Hybrid grammar language model for handwritten historical documents recognition

Núria Cirera[1], Alícia Fornés[1] Volkmar Frinken[1], and Josep Lladós[1]

Computer Vision Center, Universitat Autònoma de Barcelona, SPAIN,
{ncirera,afornes,vfrinken,josep}@cvc.uab.cat,
WWW home page: http://dag.cvc.uab.cat/

Abstract. In this paper we present a hybrid language model for the recognition of handwritten historical documents with a structured syntactical layout. Using a hidden Markov model-based recognition framework, a word-based grammar with a closed dictionary is enhanced by a character sequence recognition method. This allows to recognize out-of-dictionary words in controlled parts of the recognition, while keeping a closed vocabulary restriction for other parts. While the current status is work in progress, we can report an improvement in terms of character error rate.

Keywords: Handwriting Recognition, Language Models, Historical Documents, Document Analysis, Syntactic Pattern Recognition

1 Introduction

Handwriting recognition aims to retrieve and interpret the sequence of characters, words, and other symbols drawn in handwritten documents. This is of great convenience in order to preserve, store, make accessible, and relate textual information. Despite many years of research, it can still not be considered as a solved problem [1]. Here, we focus on the transcription process of syntactically structured historical documents in the absence of segmented data. Manuscripts share many and varied problems, such as line distortions, warped pages and physical degradation, ruling lines interference, bleeding, etc. or the variety of their contents such as multiple writers, languages, writing styles, etc.

A common technique for recognizing unsegmented handwritten text is based on hidden Markov models. In this approach, language models are integrated into the recognition process in order to return a character sequence that corresponds to a given language. This is done by restricting recognition hypotheses and re-ranking them accordingly to the recognized context. The different methods to include syntactic information vary from n-grams word models, n-grams class models, context free grammars to stochastic context free grammars. N-grams estimate the probability of word sequences, while grammars define the syntactically valid sequences of word categories and structure.

For example, word bi-grams are used in [2] to recognize handwritten medieval documents, and word bi-grams and tri-grams are used in [3] for modern handwriting recognition. An example of a grammar-based language modeling in [4] is applied for handwritten number words recognition.

The usage of a dictionary containing any existing word is computationally infeasible, hence the trend is to add *a few* extra words to a limited set. Yet, there are some drawbacks. One of them is that the exact number of words to be used can not easily be determined and often rely on an expert's criterion. Most importantly, however, any language model's performance that uses a closed dictionary is upper bounded by the impossibility of recognizing out-of-dictionary words.

As we aim to recognize text that follows a certain syntactical structure, we propose in this paper a hybrid grammar that includes character n-gram information in order to allow an open recognition of character sequences while at the same time honoring the syntactic structure of the text. This approach is useful in scenarios where new words are likely to appear in the text, for instance where few training data is available or where the training and test sets' vocabularies are not homogeneous. In the case of historical documents both problems emerge.

The presented approach is applied to recognize handwritten marriage licenses [6]. These records are quite syntactically structured. The experiments show that using our language model it is possible to recognize words that are not contained in the closed dictionary. We call it dictionary opening. We apply this dictionary opening in certain word categories where new words can occur.

The layout of the remainder of this paper is organized as follows. In Section 2, we explain the used methodology and techniques. Then, in Section 3 we expose the evaluation database, the experimental set-up and the results obtained. To conclude, in Section 4 we summarize the work and shed light on what a plausible and profitable future work could be as an extension to the present one.

2 Methodology

When the characters of a word are connected, *Sayre's paradox* [8] clearly states that segmentation and recognition require each other as a pre-processing step. This can be overcome by representing a text line as a sequence, which implicitly over-segments the image. For recognition, we use hidden Markov models (HMM).

2.1 Preprocessing and Feature Extraction

The preprocessing of a text image aims to reduce the writing variability produced by the image acquisition and the writer-specific characteristics.

We first binarize the images using Otsu's method. Then we apply skew correction to normalize the horizontal slope by means of a linear regression of the lower contour. After this, we correct the slant to obtain vertical text strokes using a horizontal shear mapping on the image. Finally we identify and normalize the three vertical zones, ascenders, descenders, and x-height.

After pre-processing, we extract three global and six local features from a one pixel width sliding window. The three global features we use are the 0^{th}, 1^{st} and 2^{nd} moments of the foreground pixels. The six local features are the positions and derivative of the top and bottom contour, the number of transitions from foreground to background, and the number of foreground pixels between the contour. For further details we refer the reader to [5].

2.2 Hidden Markov Models for handwriting recognition

We use a semi-continuous, character-based HMMs as the underlying recognition approach. Each character HMM is of linear topology and trained on images containing continuous text. In the recognition step, the character HMMs are concatenated into word-HMM which are, in turn, composed into a large recognition network. The design of this recognition network is guided by our proposed hybrid language model.

2.3 Hybrid grammar language model

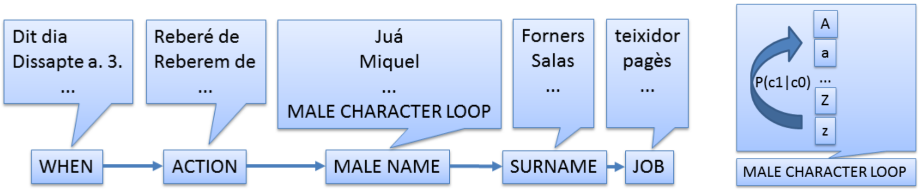
Language models give a probability distribution over text strings and are used to improve the results by favoring more likely recognitions according to the implicit lexical and syntactic structure.

Lexicon-driven recognizers usually specify the language model in terms of a provided dictionary. As long as a word is listed in the dictionary, there are chances to retrieve the right word. With word n-gram probabilities we can increase the recognition probabilities of words that usually are adjacent. However, if a word is not contained in the dictionary, it can not be recognized. In contrast, in unconstrained character recognition, any arbitrary character sequence might be recognized. This approach, however, is not stable and has a high risk of producing character strings which do not form existing words of a language, even by estimating character n-gram probabilities. Another limitation of character-level recognition is that we can not define a grammar or sentence structure at all.

In short, character n-grams can overcome the limitations posed by out-of-dictionary words and word n-grams can overcome single character mis-recognitions, yet both lack global and syntactic information. Such information can be included using a grammar. But, as it is defined at a word level, it also suffers from the drawback of being unable to recognize words that do not appear in the dictionary.

It is evident that both n-gram and grammar approaches have a great potential but also intrinsic drawbacks. To overcome the weak points of these language models, we propose as a novelty a hybrid language model consisting of character n-grams and a grammar. Hence, the language model is described by a word grammar that allows at its lexical level single character recognition.

The idea is that, given a grammar that contains categories of tokens recognized by the parsing process, we add to key categories the option of shifting from a closed word dictionary to words out of the dictionary but recognized by a character bi-gram model (see Figure 1). In this way we open a closed dictionary to new recognition options.



(a) (b)

Fig. 1: Hybrid grammar: a) grammar derivation sequence example, with a hybrid option for the male name category, b) open recognition of character sequences, based on character bi-gram probabilities estimated on the male name category.

2.4 Character bi-grams

When the text is syntactically structured, there are syntactic fields in the grammar and its content shall vary among a certain lexicon. For instance, if a male name is expected, the lexicon of this field is a collection of male names that appeared during the training process, so they are in a closed dictionary. Our purpose here is to also recognize names that are not in this dictionary, yet should resemble possible names from the same language. In order to do it, we transform the names into character sequences and estimate the character bi-gram probabilities on the training set.

At this point, we are treating the names as character sequences, so the learned character bi-grams also take into account the probability of a name to start with a certain character or either to finish with it. This would restrict the open recognition to character sequences that started only with character that were first character of training names. The same would happen for ending characters. In order to avoid this and to create an open recognition that can cope with any new name, we modify the learned character bi-grams to include the option to start and finish with any character using Kneser-Ney smoothing [10].

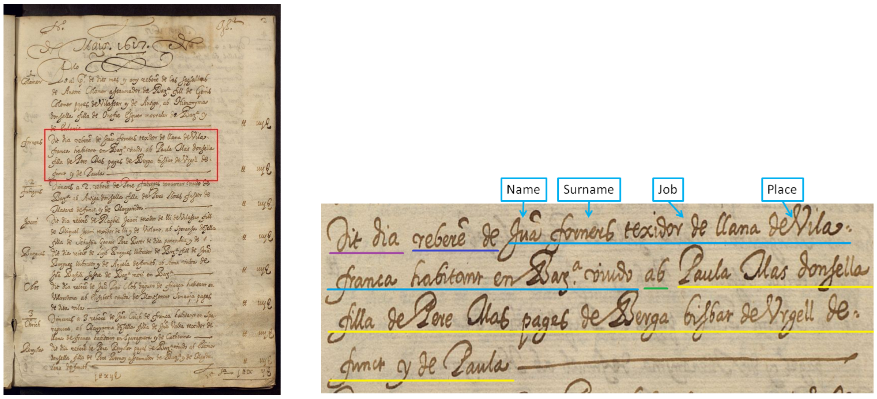
These character bi-grams are not external information that we add to the language model. All the new characters come from the character list that is used to learn the character HMMs, and the character bi-gram probabilities are also based on the training frequencies.

3 Experimental validation

3.1 Dataset description

The database in which we perform the evaluation is part of the *5CofM*'s *Llibres d'Esposalles*. The subset we use is Volume 69, and it consists of 1 714 marriage registers distributed in 173 pages and annotated from May 1617 to April 1619. The language used in these is Catalan and the conservation status is satisfactory. We use 1 480 registers as training set and 256 as test set, 1-fold. The test set consists of a total of 8 925 words and 38 635 characters.

The layout of these documents consists of separate registers, each in one paragraph. The structure of each of these registers is comprised by certain characteristics, such as the date of the wedding, the groom's name, the job and origin, the groom's parents' names, the bride's name, her father's name, and her home-town. We can see an example in Figure 2. Note, that not all data have to be present and that noise such as crossed out words may be present.



(a) Page with 8 registers.

(b) Register contained in 2a.

Fig. 2: Layout of the dataset *Llibres d'Esposalles*.

3.2 Experimental setup

We train a total of 83 single HMMs corresponding to the characters and symbols listed that appear in the training set. Each of these characters and symbols has six hidden states that model the emission probability using a mixture of 64 Gaussians. The number of Gaussian and states are taken from [7].

For the Baum-Welch training and Viterbi decoding, the HTK implementation is used [9]. The word insertion penalty has been set to 0.5 and the grammar scale factor to 10.5, based on empirical results on a small training subset.

3.3 Results

We define a grammar that generalizes the syntactical structures seen in the training set, in particular the structure in Figure 2b. This grammar uses for specific categories (namely *male name*, *female name*, *surname*, *job* and *town*) a closed dictionaries extracted manually by laymen and checked by linguist experts. This is standard procedure for real-world tasks, yet it imposes severe drawbacks, since this transcription is not character-accurate. A hyphenated word in the image, e.g. may be transcribed as a single word, or abbreviations might not occur as such. From the *male name* and *female name* fields we compute an estimation of the character bi-grams, and we include a character loop option in these two dictionaries, as explained in Subsection 2.3. The remaining of this grammar is built manually, based on the training set.

In Table 1 we can see the character error rate (CER) computed on the hybrid grammar's result. To calculate CER we transform the recognized words in sequences of characters, following the rule that after a word we include a empty space character, and we do leave the character recognitions as they are.

Language Model	CER
Character uni-grams	645.21
Character bi-grams	32.74
Word uni-grams	22.44
Grammar	20.89
Hybrid Grammar	**20.42**

Table 1: Language models' performance results on the test set, in terms of CER.

As a comparison, we also show the performances of other language models. We implemented a grammar without any hybrid part. A different case is the word unigram language model. We also have tested character n-grams for $n = 1$ and $n = 2$. In these cases, in order to learn character unigrams and bi-grams, we use the character-accurate transcriptions of the training set.The quantitative results show that the proposed hybrid grammar improves the character recognition rate over all reference systems.

The potential of the hybrid grammar can be appreciated in Table 2, where portions of register's recognitions are shown. In the first example, the male name *Jua$* is not in the closed dictionary, so the regular grammar recognizes a similar name instead while the hybrid grammar succeeds in the recognition. In the second example, the composed female name *Anna Victoria* is correctly recognized by the hybrid grammar as a sequence of characters and spaces. In this case the regular grammar does not recognize it because, although the female names *Anna* and *Victoria* are present in the training set, their composition is not. The second example also reports a non desired extension of the character recognition's length, although the recognition is perfect but in one single character.

The hybrid grammar presented here can be used to limit the recognition to an assumed syntax while permitting at the same time the flexibility to recognize out-of-dictionary words at pre-defined positions. Our grammar is currently work in progress and does not include any word bi-grams. Word transition probabilities

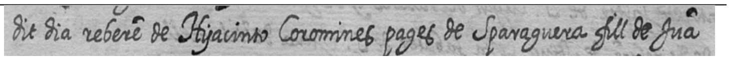
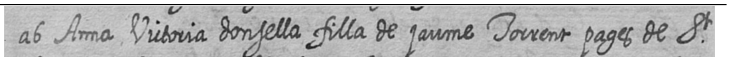
Image	 dit dia rebere$ de Hyacinto Coromines pages de Sparaguera fill de Jua
Transcription	dit dia rebere$ de Hyacinto Coromines pages de Sparaguera fill de Jua$
Grammar	dit dia rebere$ de Hyacinto Coromines pages de Sparaguera fill de Jua
Hybrid Grammar	dit dia rebere$ de Hyacinto Coromines pages de Sparaguera fill de _J _u _a$ _@
Image	 ab Anna Victoria donsella filla de jaume Torrent pages de St.
Transcription	ab Anna Victoria donsella filla de jaume Torrent pages de Sˆ(t).
Grammar	ab Anna Maria donsella filla de jaume Torrent pages de Sˆ(t).
Hybrid Grammar	ab _A _n _n _a_@ _V _i _c _t _o _r _i _a _@ donsella filla de _j _a _u _m _e _@ _T _o _r _r _c _n _t _@ _p _a _g _e _s _@ _d _e _@ _S _tt _.

Table 2: Qualitative results of the hybrid grammar compared with a non-hybrid grammar. Note that a character recognition is preceded by the symbol _ , a superscript character is surrounded by *ˆ( )* at word level, symbol *_tt* stands for the superscript character *t* and the empty space is represented by _@ symbol.

are additional information that can be estimated on the training set or even external data. Hence, we deliberately did not include a word bi-grams in the list of reference language models, since it would not be a fair comparison.

4 Conclusion

This paper has presented a hybrid grammar language model for handwritten historical documents that have a syntactical layout repeated over the document. This language model has a word-grammar layout that reflects the syntactical structure and uses a closed dictionary, enhanced with a character sequence recognition option embedded in the grammar, which allows to recognize words outside the closed dictionary. We compared this language model's performance with character unigrams and bi-grams, word unigrams, and a regular grammar.

The qualitative results prove successful performance of the proposed language model, recognizing words out of the dictionary. Furthermore, we notice that the open recognition can be more likely than the closed one even for other categories, resulting in an unexpected extension of the character sequence recognition. This suggests to make the open recognition more restrictive or to extend it to other categories.

In its current form, word transition probabilities are not yet considered, neither in the hybrid grammar approach nor as a reference system. However, the performance comparison with stochastic and semantic language models using the same information is very promising. In the future, word transition probabilities will also be included.

Furthermore, we can also investigate the performance when using higher order n-grams, *e.g.* 3-grams, 5-grams or 7-grams. Along a separate line of research are investigations into opening not only the names but also other entities, such as job titles or city names. Finally, cross-references within the database or between other sources could lead to an enhanced estimation of the sought-after probabilities of names, job title, or cities.

Acknowledgement

This research was partially supported by the European Research Councils Advanced Grant ERC-2010-AdG 20100407 and the Spanish R&D projects TIN2009-14633-C03-03 and TIN2012-37475-C02-02.

References

1. Espana-Boquera, S. , Castro-Bleda, M.J. , Gorbe-Moya, J. , Zamora-Martinez, F.: Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models. IEEE Transactions onPattern Analysis and Machine Intelligence, Vol. 33(4), 767 -779, (2011)
2. Wuthrich, M. , Liwicki, M. , Fischer, A. , Indermuhle, E. , Bunke, H. , Viehhauser, G. , Stolz, M.: Language Model Integration for the Recognition of Handwritten Medieval Documents. 10th International Conference on Document Analysis and Recognition, (2009)
3. Zimmermann, M. , Bunke, H.: N-gram language models for offline handwritten text recognition. 9th International Workshop on Frontiers in Handwriting Recognition, (2004)
4. Toselli, A. H., Juan, A., González, J., Salvador, I., Vidal, E., Casacuberta, F., Keysers, D., Ney, H.: Integrated Handwriting Recognition And Interpretation Using Finite-State Models. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18 (8), 519–539 (2004)
5. Marti, U.-V., Bunke, H.: Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems. Hidden Markov models, 65–90 (2001)
6. Romero,V. , Fornés, A. , Serrano, N. , Sánchez, J.A. , Toselli, A.H. , Frinken, V. , Vidal, E. , Lladós, J.: The ESPOSALLES Database: An Ancient Marriage License Corpus for Off-line Handwriting Recognition. Pattern Recognition. Vol. 46(6), 1658-1669 (2013)
7. Romero, V., Sánchez, J.A., Serrano, N., Vidal, E.: Handwritten Text Recognition for Marriage Register Books. Proceedings of the International Conference on Document Analysis and Recognition, 533-537 (2011)
8. Sayre., K.M.: Machine Recognition of Handwritten Words: A Project Report. Pattern Recognition, Vol. 3 (3), 213–228 (1973)
9. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Technical report, Cambridge University Engeneering Department, (Dec. 2006)
10. Goodman., J. T.: A Bit of Progress in Language Modeling - Extended Version. Technical Report MSR-TR-2001-72, Microsoft Research, One Microsoft Way Redmond, WA 98052, 8 (2001)