On the Use of Context-Dependent Modeling Units for HMM-Based Offline Handwriting Recognition

Gernot A. Fink Thomas Plötz
University of Dortmund, Robotics Research Institute,
44221 Dortmund, Germany
{Gernot.Fink,Thomas.Ploetz}@udo.edu

Abstract

The use of context dependent modeling units in handwriting recognition has been considered by many authors as promising substantial performance improvements in systems based on Hidden-Markov models. Interestingly, in the literature only a few approaches limited to online recognition are documented to make use of this technology. Therefore, we investigated whether context dependent modeling also offers advantages for offline recognition systems. The moderate performance improvements we achieved on a challenging unconstrained handwriting recognition task suggest that context dependent modeling can not easily be exploited for offline recognition. In this paper we will present the principles behind context dependent modeling and discuss the reasons for its limited applicability in recognizing offline handwriting data.

1 Introduction

The use of Hidden Markov Models (HMMs) for recognizing machine printed or handwritten text is motivated by the success of this method in the field of automatic speech recognition (cf. [10]). In contrast to classical approaches to character recognition that rely on a combination of explicit segmentation and classification steps, HMMs allow for a classification of characters or words with *implicit* segmentation. Therefore, the approach is especially attractive for handwriting recognition where reliable character segmentation is extremely hard to achieve.

In order to make a technique for the statistical modeling of sequence data applicable to images of printed or handwritten text, those must be transformed into a suitable sequence representation. This is achieved by sliding a narrow analysis window along a text line in the direction of writing. Then features are extracted at each position from the small patch of the text image covered by the analysis window. The sequence of feature vectors generated by this *sliding window* approach (cf. [1]) is then – from a principal point of view – comparable to the feature vector sequences extracted from speech data, and can, therefore, be modeled with the same statistical techniques, namely HMMs.

From the similarity of the data representation it follows immediately that similar structures for HMMs can be used. In speech recognition complex models describing arbitrary sequences of words from a given lexicon are constructed from elementary units for speech sounds (phones). Those units, which are usually realized as HMMs with either linear or Bakis topology, can be concatenated to form models for whole words. By adding further edges to the model arbitrary word sequences can be represented. Searching the path through the model, along which the data is generated with maximum probability, gives the segmentation and recognition result for a spoken utterance.

In order to apply this model structure to text recognition only the elementary units need to be exchanged. Instead of phones now models for characters, numerals, and punctuation symbols have to be defined. An important difference in the models applied for speech or text recognition is, however, that in speech recognition reasonable modeling quality is only achieved by using context-dependent phone units – a modeling strategy found in almost any speech recognition system today. In text recognition, however, research efforts involving context-dependent modeling can be found for online tasks only. Therefore, in this paper we try to shed some light on the inherent difficulties faced when using context-dependent models for offline handwriting recognition.

The paper is organized as follows. First, we will describe the idea of context dependent modeling developed originally for speech recognition. Afterwards, we will review the approaches for handwriting recognition using context dependent models. In section 4 we will then discuss the principal challenges that context dependent modeling is faced with in offline recognition tasks. Results of experimental evaluations will be presented in section 5 and conclusions will be drawn in section 6.

2 Context-Dependent Modeling

In speech recognition the goal is to transcribe spoken utterances into sequences of words. However, for reasonable lexicon sizes it is not feasible in practice to model all known words with separate HMMs, as in general not enough training samples will be available. Therefore, word models are constructed from smaller *sub-word units*, which will occur more frequently in the data as they are shared among words. The most natural choice for such building blocks in speech modeling is phones, i.e. speech sounds.

Though building word models from a small set of elementary phone units is a very versatile approach, it can not cope with the effects of co-articulation between neighboring phones – i.e. context-dependent variations in pronunciation – that severely effect their realization. In order to improve the modeling quality, already in the 1980s researchers proposed to distinguish phone models according to their context. The basic idea was refined and lead to the definition of so-called *triphone* models representing phone units in the context of the neighboring left and right phones.

However, pure triphone units suffer from the fact that the associated speech events will in general be very special and, therefore, tend to occur too rarely in the training data available. Consequently, in practice only suitable variants of those models can be used which are general enough to be estimated robustly. As the specialty of triphones results from their context restrictions more general units can be found by reducing those constraints – a process which leads to so-called *generalized triphones* [5].

The most straight-forward way of generalizing context restrictions is to merge similar contexts on the symbolic level. The method has the disadvantage, however, that the appropriate rules need to be defined by experts. Therefore, a very successful alternative method merges triphone units in a data-driven manner on the level of individual HMM states by applying an agglomerative clustering procedure. First, an intermediate model based on pure triphone units only is trained. The parameters of the model's states will in general be estimated on very few samples only and, therefore, will be almost unusable in practice. Secondly, groups of similar parameter sets are determined by performing agglomerative clustering in the parameter space of the intermediate model. The final model is then defined on the basis of the computed generalized state parameter sets which are shared between multiple triphone models.

3 Related Work

According to the literature the majority of state-of-the-art handwriting recognition systems applying HMMs is based on models estimated for context independent characters [3]. In order to model complete words they are, usually, concatenated and lexicon based decoding is performed.

Some authors pointed out the enormous potential context-dependent modeling units might offer for improving existing handwriting recognizers [3, 9]. The incorporation of contextual information at the modeling level has been identified as "future issue" and "very challenging". So far, however, only very few related attempts have been documented. To the authors' knowledge, in fact no *offline* recognizer based on context-dependent modeling units has ever been described. For *online* handwriting recognition two systems were developed.

In [4] for the first time the basic idea of context-dependent modeling originating from automatic speech recognition in terms of triphones was generalized to the domain of handwriting recognition. Kosmala et al. developed trigraphs describing characters – i.e. monographs – in the context of their left and right neighbors as base units for the recognizer. Reducing the complexity of the resulting model-space by either using only the most frequent trigraphs or by state-clustering, respectively, substantial improvements in classification accuracy were reported. In [8] the authors enhanced their recognition system towards a very large lexicon size (200k). Due to the enormous computational complexity in decoding the use of context-dependent models has, however, been given up.

In the second system context-dependent modeling has been introduced for an online Japanese handwriting recognizer [11]. Cursive Kanji or Hiragana characters were modeled using sub-stroke models additionally covering adjacency relations. Comparable to [4] the basic idea for context-dependent modeling is also related to the general concept of triphones. Every sub-stroke is considered depending on its surrounding sub-strokes. For computational feasability a successive state splitting algorithm is applied allowing for efficient organisation of the search space in a so-called Hidden Markov net. For the resulting recognition system promising classification results for an experimental evaluation have been reported.

4 Offline Handwriting Recognition with Context-Dependent Units

On a purely technical level the application of context-dependent units for offline handwriting recognition is straight forward. There are, however, several important aspects distinguishing the offline recognition problem and the data to be described by the statistical models from both online handwriting recognition and speech recognition.

Configuration of modeling units: In speech recognition the basic phone units have a very simple configuration. The topology is either linear or Bakis – i.e. skipping of one state is allowed – and the models use mostly 3 independent states.

The 3-state configuration follows an articulation model that assumes a transition phase into the current phone, a stationary phase, and a transition to the next phone. Therefore, the degrees of freedom within the basic units are rather limited (approx. 50 phone models amount in 150 total HMM states) and can only be increased by using more complex output probability density functions. In this situation adding a substantial number of states for more specialized models can greatly improve recognition accuracy while still maintaining trainability of the generalized context dependent models. In our work on speech recognition (cf. [2, 7]) we typically observe an increase of a factor of 30–40 in the state space when going from context independent to context dependent models and a reduction in error rate of up to 50%. Similar improvements are reported for online handwriting recognition using context dependent models [4].

In offline handwriting recognition the situation is different, as a reasonable modeling quality for context independent character units is only achieved by models with substantially more states. As the analysis windows of the sliding-window feature extraction need to be quite narrow, characters are decomposed into long sequences of feature vectors that exhibit a substantial amount of variation. In our offline recognition systems we typically use an average number of 30 states for the 75 basic units. Therefore, it is already clear from considering availability of training data only, that a 30-fold increase of a 2k state space would result in a model for which parameters could not be estimated robustly. However, increasing the state space only moderately has the disadvantage that not enough specialized modeling capacity is added.

Principles of context influence: The second important difference between offline handwriting recognition on the one and speech and online handwriting recognition on the other side is that in the latter two approaches contextual influence is caused by well understood underlying principles. In speech recognition phone articulation is primarily dependent on the configuration of the vocal tract, i.e. relative positions of the active and passive articulators such as lips, tongue and teeth. Ideally, a speech sound is produced with one specific of those configurations and the next phone with a different one. Therefore, articulators have to move from one configuration to the next producing intermediate configurations and, consequently, contextually influenced portions of speech sounds. This sequential context influence can finally be observed in the feature representations of speech that all try to compute a parametric representation of the gross vocal tract configuration.

In online handwriting recognition the situation is similar. Here the contextual influence arises from the need to move the pen between realizations of adjacent characters or strokes. Also the pen movements can directly be observed in the feature representations used, where shape parameters of the pen trajectories are computed.

Though, in offline handwriting recognition there surely is some contextual influence, it is completely unclear, what *general* principles – i.e. mechanisms independent of any particular writer – this context influence follows. This becomes even more evident when considering the features calculated in the sliding-window framework. For every analysis window either some sort of local statistics of the grey-value distribution or combinations of expert-crafted geometric properties of the small text-image slices are computed (cf. e.g. [1]).

Taking both the different configuration of models and the lack of a general principle of contextual influence into account, it becomes clear that improvements with context dependent models will be significantly harder to achieve for offline handwriting recognition tasks.

5 Experimental Results

In order to investigate the potential of using context dependent sub-word units for unconstrained offline handwriting recognition we conducted a series of experiments on a challenging task defined on the IAM database [6]. The results obtained will be compared to those of a baseline system using context independent models only.

Data: The IAM database of handwritten texts consists of several hundred documents scanned at 300 dpi which were generated by having subjects write short paragraphs of text from several different text categories. No instructions concerning the writing style were given so that the data can be considered to represent truly unconstrained handwriting.

As in our previous works (cf. e.g. [12]) we used all documents from text categories A to D (485 documents, 4222 text lines) for training and the documents from categories E and F (129 documents, 1076 text lines) for testing.

This choice of training and testing material implies, however, a serious drawback for lexicon based recognition, as almost 50% of the word forms appearing in the test material were never seen in the training texts. Therefore, we decided to extend the lexicon of the training data by additionally including all those word forms, which are necessary to describe the text prompts from which the test set was generated and which could be constructed by characters found in the training data. The resulting recognition lexicon consists of 7485 entries including punctuation and word fragments resulting from hyphenation.

Baseline System: The baseline recognition system used in this study results from further development of the systems used in our previous research (cf. [12]). First, text lines are extracted from the input documents. The lines are normalized with respect to baseline orientation, slant, and estimated average character width. A pre-segmentation

into words is, however, *not* attempted. From the normalized text lines feature vector sequences are extracted using the sliding-window approach. For every analysis window a set of geometric and their discrete derivatives is computed.

The HMMs for the 7485 words in the lexicon are constructed from 75 basic units representing context independent characters, numerals, punctuation symbols, and white space. The model for the word "Brazil" is, e.g., formed by concatenating the context independent character models:

$$\lambda(\texttt{Brazil}) = \lambda(\texttt{B}) \circ \lambda(\texttt{r}) \circ \lambda(\texttt{a}) \circ \lambda(\texttt{z}) \circ \lambda(\texttt{i}) \circ \lambda(\texttt{l})$$

The models have *Bakis*-type topology. The number of states depends on the length of the associated segments in the initialization data resulting in 2248 states. The context independent HMMs as well as all context dependent models described below are semi-continuous sharing a set of 1.5k Gaussian densities with diagonal covariance matrices.

From the transcription of the training texts a bi-gram language model was estimated using absolute discounting and back-off smoothing. The model only achieves a perplexity of 757 on the test set due to the limited amount of training data and the inherent differences of the text categories.

Recognition Results: In our first approach we used context dependent sub-word units in a straight-forward way in complete analogy to our model building process for speech recognition systems (cf. e.g. [2]). All occurrences of characters with different left and right context (denoted as 'cdc3') found in the training data were established as context dependent models. The word "Brazil" is now built as

$$\begin{aligned}\lambda(\texttt{Brazil}) \quad = \quad & \lambda(\texttt{\#/B/r}) \circ \lambda(\texttt{B/r/a}) \circ \lambda(\texttt{r/a/z}) \circ \\ & \lambda(\texttt{a/z/i}) \circ \lambda(\texttt{z/i/l}) \circ \lambda(\texttt{i/l/\#})\end{aligned}$$

where $\lambda(\texttt{r/a/z})$ denotes a model for the base character `a` with left context `r` and right context `z` and # represents the word boundary. For this overall model an intermediate parameter set is estimated. Afterwards, all states corresponding to the same position in models with the same base character are subject to an agglomerative clustering. States with similar parameters are merged until a minimum of $c = 75$ training samples is available for every newly created state cluster. Finally, Baum-Welch re-estimation is applied to the state-clustered models (in total 10 re-estimation steps).

Fig. 1 shows that this modeling approach achieves clearly a better representation of the training data, as the average (negative) log-likelihood is considerably reduced with respect to the context independent baseline. The generalization capabilities of the model are, however, quite poor as can be seen from the 80% increase in word error rate (WER) on the test set (cf. table 1).

In order to improve the generalization capabilities of the model the specificity of the sub-word units needs to be reduced. This can be achieved by either increasing the threshold c used for state clustering or by reducing the number of potential contexts already on the symbolic level.

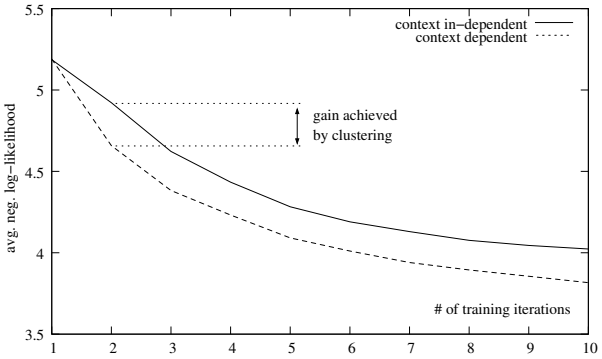

Figure 1. Modeling quality on training data

Model	c	# states	WER	Δ WER
baseline	-	2248	24.0	-
cdc3, clustering	75	14387	42.8	+78.3
	250	6143	36.3	+51.3
	500	4454	36.2	+50.8
	1000	3195	34.0	+41.6
	2000	2640	34.4	+43.3
cdc3, 5 categories,	75	6333	25.6	+6.7
clustering	250	4538	23.3	-3.0
	500	3778	**22.9**	-4.6
	1000	2996	**22.5**	*-6.3*
	2000	2574	**22.9**	-4.6
cdc3, 5 cat., 10× ci	250	4309	**22.7**	-5.4
	1000	3610	**22.7**	-5.4
	1000	2938	**22.9**	-4.6

Table 1. Recognition results on IAMdb test

When retaining all possible contexts a change of the clustering threshold results in a substantial performance increase though never matching the baseline system. The results are summarized in the upper part of table 1, where the model configuration, the clustering threshold, the effective number of HMM states, the WER, and the relative improvement w.r.t. the baseline (both in percent) are given.

One might expect that the search in parameter space by the clustering could also specialize to something different than context. Therefore, we tried to guide the clustering by first reducing the degrees of freedom by grouping contexts on the symbolic level. As it is far from obvious how classes of similar context should be defined for offline data we came up with 5 rather simple context categories that – to some extent – show similar geometric properties: characters that occupy the core area only (`core`), characters with ascenders (`asc`), descenders (`desc`), or both (`adc`), numerals (`num`), and capitals (`cap`). Punctuation symbols and white

space were treated separately. “Brazil” is now defined as:

$$\begin{aligned}\lambda(\texttt{Brazil}) \;=\; & \lambda(\texttt{\#/B/core}) \circ \lambda(\texttt{cap/r/core}) \circ \\ & \lambda(\texttt{core/a/core}) \circ \\ & \lambda(\texttt{core/z/core}) \circ \\ & \lambda(\texttt{core/i/asc}) \circ \lambda(\texttt{core/l/\#})\end{aligned}$$

As can be seen from the results given in the middle part of table 1 the performance could be increased significantly (typeset in boldface) beyond the baseline.

Another drawback of the state clustering procedure lies in the estimation of the intermediate model. Therefore, instead of finding state clusters after the first iteration of parameter re-estimation we investigated, whether a well trained context independent model (here after 10 re-estimation steps) might improve the performance when used as initial parameter set. All models estimated with this two-step procedure achieve a significant improvement beyond the baseline and consistently perform on a similar level as the best configuration of the directly trained models (cf. lower part of table 1).

6 Conclusion

As the success of context dependent models in the domain of speech recognition was so triumphant there were no obvious reasons to doubt that this modeling technique would deliver vast performance improvements also in the related area of handwriting recognition. However, the documented approaches in the literature, which deal with online data only, and our own experiments with offline recognition strongly suggest that context dependent models can’t be as successfully applied in offline approaches.

In this paper we presented an extensive experimental analysis of different methods for using context-dependent modeling units for HMM-based handwritten text recognition. Though significant improvements in performance could be achieved the reduction in word error rate of approx. 6% relative is far from matching the performance gains reported on speech or online handwriting recognition tasks.

We assume that the primary reason for this is that the underlying assumptions of a well defined mechanism of contextual influence are violated when computing feature representation of offline handwriting data via the sliding window approach. Though the training data available in the IAM database is quite substantial, a secondary reason might be still insufficient data. As can be seen from the results the best performing models only used a few hundred additional state clusters for a more accurate representation of the data. Finally, as the nature of the contextual variation is completely unclear for offline data the specialization of the additional models might also capture variation caused by some totally different effect. However, adopting some other definition of “context” (e.g. writing style) which deviates from the strict linear context of the symbol sequence itself (i.e. word context) leads to the creation of model variants and substantially complicates model decoding.

Consequently, more accurate modeling for offline handwriting recognition still remains a challenging and largely open problem which we will address in future research.

References

[1] H. Bunke. Recognition of cursive roman handwriting – past, present and future. In *Proc. Int. Conf. on Document Analysis and Recognition*, pages 448–459, Edinburgh, 2003.

[2] G. A. Fink and T. Plötz. Integrating speaker identification and learning with adaptive speech recognition. In *2004: A Speaker Odyssey – The Speaker and Language Recognition Workshop*, pages 185–192, Toledo, 2004.

[3] A. L. Koerich, R. Sabourin, and C. Y. Suen. Large vocabulary off-line handwriting recognition: A survey. *Pattern Analysis and Applications*, 6(2):97–121, 2003.

[4] A. Kosmala, J. Rottland, and G. Rigoll. Improved online handwriting recognition using context dependent hidden markov models. In *Proc. Int. Conf. on Document Analysis and Recognition*, volume 2, pages 641–644, Ulm, Germany, 1997.

[5] K.-F. Lee. Context dependent phonetic hidden markov models for continuous speech recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38(4):599–609, 1990.

[6] U.-V. Marti and H. Bunke. The IAM-database: An English sentence database for offline handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[7] T. Plötz and G. A. Fink. Robust time-synchronous environmental adaptation for continuous speech recognition systems. In *International Conference on Spoken Language Processing*, volume 2, pages 1409–1412, Denver, 2002.

[8] G. Rigoll, A. Kosmala, and D. Willett. An investigation of context-dependent and hybrid modeling techniques for very large vocabulary on-line cursive handwriting recognition. In *Proc. 6th Int. Workshop on Frontiers in Handwriting Recognition*, Taejon, Korea, Aug. 1998.

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

[10] T. Starner, J. Makhoul, R. Schwartz, and G. Chou. On-line cursive handwriting recognition using speech recognition methods. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 5, pages 125–128, Adelaide, 1994.

[11] J. Tokuno, N. Inami, S. Matsuda, M. Nakai, H. Shimodaira, and S. Sagayama. Context-dependent substroke model for HMM-based on-line handwriting recognition. In *Proc. Int. Workshop on Frontiers in Handwriting Recognition*, pages 78–83, 2002.

[12] M. Wienecke, G. A. Fink, and G. Sagerer. Toward automatic video-based whiteboard reading. *Int. Journal on Document Analysis and Recognition*, 7(2–3):188–200, 2005.