A Hybrid CRF/HMM Approach for Handwriting Recognition

Gautier Bideault, Luc Mioulet, Clément Chatelain, Thierry Paquet

A hybrid CRF/HMM approach for handwriting recognition

Gautier Bideault, Luc Mioulet, Clément Chatelain, and Thierry Paquet

Laboratoire LITIS - EA 4108, Université de Rouen, FRANCE 76800
Email: firstname.lastname@univ-rouen.fr

Abstract. In this article, we propose an original hybrid CRF-HMM system for handwriting recognition. The main idea is to benefit from both the CRF discriminative ability and the HMM modeling ability. The CRF stage is devoted to the discrimination of low level frame representations, while the HMM performs a lexicon-driven word recognition. Low level frame representations are defined using n-gram codebooks and HOG descriptors. The system is trained and tested on the public handwritten word database RIMES.

1 Introduction

Handwriting Recognition (HWR) is a difficult problem due to the high variability of the data. Currently, the most widely used probabilistic models for handwriting modeling are Hidden Markov Model (HMM) [12]. Multiple training frameworks have been proposed to train these generative models. The original generative framework relies on a Maximum Likelihood (ML) criterion [10], but it has been shown that a discriminative framework based on a Maximum Mutual Information (MMI) criterion [2] could lead to some improvement. Regardless of the criterion, HMM rely on strong observation independence assumptions and they perform poorly on high dimensional observations.

Conditional Random Fields (CRF) [16] became more and more popular models during the last decade for sequence modeling because they are discriminative models and they do not rely on the same restrictive assumptions. The original CRF framework [16] was proposed to process symbolic data in the field of automatic language processing [6]. A major drawback concerning CRF is there inability to process numerical data, they only process discrete values. When facing numerical data, they are generally introduced at a second stage of the model in order to model the dependency between classes, while raw numerical data are analyzed through a classification stage such as Artificial Neural Networks (ANN) for example in the field of Automatic Speech Recognition (ASR) [8, 18].

Despite their ability to deal with symbolic data, CRF models are limited to label the observation sequence, *i.e.* to provide a label to each frame of the sequence. As a consequence, the CRF is not able to integrate high level knowledge through the integration of lexicons and/or language models, as it is whith HMMs. A second limitation of CRF, as opposed to HMMs, is the requirement of having

groundtruthed data at frame level in order to train the models, thus preventing using embedded training afforded within the HMM framework.

In this paper, we propose a hybrid model that takes advantage of both generative and discriminative models in order to tackle Off-Line omni-writer handwriting recognition. The paper is organized as follows: first a review of the related works is given in section 2, then we present the hybrid model devoted to handwriting modeling in section 3. Experimental setup and results reported using the RIMES database [5] are presented in section 4.

2 Related Work

In the early nineties, hybrid architectures have been proposed to combine the advantages of both discriminative and generative models. They were initially designed for ASR by combining ANN (mostly Multi Layer Perceptron) with HMM [15]. Such hybrid models have also been proposed for HWR [1].

In general, these models use the ANN discriminative stage to analyse and classify local observations at frame level, whereas the HMM generative stage is devoted to the integration of higher level information such as lexicon, language models, ... More precisely, the Gaussian Mixture Models (GMM) of the HMM stage are substituted for local posteriors computed by the ANN stage.

Recently, the Bilateral Long Short Term Memory (BLSTM) neural networks combined with a Connectionist Temporal Classification (CTC) stage [4] has proven to be a powerful alternative hybrid structure for sequence classification. Such a structure combines an efficient low level frame modeling stage with the ability to model long time dependencies, with a discriminative classification stage made of a simple logistic classifier. This structure has proven to perform extremely well for ASR and HWR [3].

CRFk [7] were originally formulated for language processing tasks, due to their interesting theoretical properties they have also been applied in fields in which the ability to process numerical data is important. Hence, in order to process this data the CRF model has been adapted to be applied to applications fields such as Gesture Recognition (GR) [9] or ASR [11].

In the field of HWR, some attempts have been reported on using CRF models. In [14], the authors introduce a CRF model to perform character sequences recognition. However, this method is applied on an already segmented character sequence consequently the segmentation is not modeled by the CRF stage. In order to perform both segmentation and recognition of characters [17] introduced a non linear HCRF model that consists in a Deep Neural Network (DNN) and a CRF. The deep structure improves the discrimination at the low level while the HCRF allows high level modeling.

Most of the previous works of the literature that have developed hybrid models introduce a discriminative stage that deals with the low level input observable raw data. Neural Networks such as MLP, BLSTM or DNN are suitable models that provide higher level informative features (e.g class conditional probabilities) to the second stage of the hybrid architecture. This second stage is most of the

time devoted to the contextual analysis of the hypothesis given by the first stage. It is generally based on a generative model that can introduce constraints such as lexicons and/or language models. In most cases, HMMs are implemented, but dynamic programming stages, such as CTC, have proved to be a possible alternative architecture.

HCRF have the specificity to be discriminative at both low and high level stages. But they are limited to the task of sequence labelling they have been trained for. Moreover, they cannot embed higher level information such as lexicon or language model at decoding time.

The following section presents the proposed hybrid model.

3 A CRF-HMM hybrid Approach

3.1 Overview of the proposed approach

The proposed CRF / HMM architecture has been chosen in order to take advantage of both generative and discriminative frameworks. As described on Figure 1, the CRF stage performs the discrimination of the low level frame representation. It extracts the local posterior probabilities of every character at every time using a forward-backward inference :

$$p(s_t = q_k|O^{(n)}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{1}$$

The forward $\alpha_t(j)$ and backward variable $\beta_t(j)$ are defined as :

$$\alpha_t(j) = P(O_1O_2...O_t, q_t = S_i|\lambda) \tag{2}$$

i.e the probability of the partial observation sequence, $O_1O_2...O_t$ and state S_i at time t, given the model λ.

The backward variable $\beta_t(j)$ is defined as :

$$\beta_t(j) = P(O_{t+1}O_{t+2}...O_T|q_t = S_i, \lambda) \tag{3}$$

i.e the probability of the partial observation sequence $O_{t+1}O_{t+2}...O_T$, given the state S_i and the model λ.

In order to use the discriminative and highly contextual information of the CRF, the GMM of the HMM stage are substituted for these local posteriors, as it is traditionally the case for hybrid Neuro-HMM structures. Doing this, we can use the HMM generative stage to analyze the information in context with the possibility to introduce lexical and language constraints.

As CRFs are not able to cope well with numerical data, we propose an unsupervised classification stage based on k-means devoted to the discretization of the numerical Histograms of Oriented Gradient (HOG) feature vector (for further details see section 3.3 and 4.1).

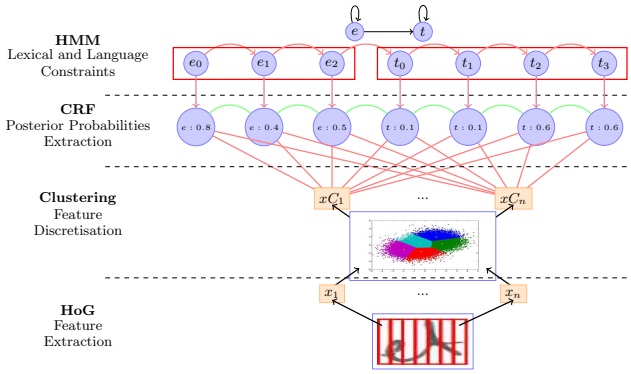

Fig. 1. Hybrid structure CRF/HMM : Detail of every step of the whole hybrid structure from feature extraction to word recognition for the word **et**.

3.2 CRF-HMM training

In order to train our hybrid CRF/HMM structure, we have to train both CRF stage and the transition probabilities of the HMM stage. An important issue when training a discriminative model, such as CRF, is that it requires a labelled training set at the frame level, whereas the groundtruth of handwriting databases is generally given at the word level. In order to get this frame level segmentation, we need to use first a standard HMM model trained on the same learning dataset, and used in a forced Viterbi alignment mode of the frame data on the word character sequence groundtruth. Following this frame labelling stage, the CRF is trained using Stochastic Gradient Descent (SGD). The convergence and the overfit of the training is controlled on a validation dataset during training. The HMM parameters (the conditional transition probabilities) are also computed on the labelled dataset.

3.3 *N*-gram data representation

CRF have been originally proposed to deal with high dimensional discrete symbolic features (words) for automatic language processing tasks. Therefore, HCRF have been introduced to deal with real valued raw data, in a way similar to neural networks or deep neural networks can do. Deep architectures have the ability to learn high level features from the raw numerical data an unsupervised training , whereas HCRFs introduce a fixed number of hidden states that act as sequentially structured features optimized during training.

The drawback of these architectures is their very long training time and their sensitivity to the initial conditions, which make them difficult to optimize with

standard computational resources. The use of GPU is recommended to learn the model under a reasonable time.

Taking advantage of the ability of CRFs to deal with very large discrete features (several thousands in the case of language processing), which can even be extended to n-gram features as a result we use n-gram feature codebooks. In a way similar to the pre-training stage of a DNN, feature codebooks are trained in an unsupervised manner, so as to minimize the mean square error of the training set, using k-means, or LindeBuzoGray clustering for example.

This stage provides a high dimensional symbolic feature codebook representation of the data (see Fig. 2). In the experiments described below, we explore the use of uni-gram, bi-gram and tri-gram feature codebooks.

4 Experiments

4.1 Discretization of frame level numerical features

An initial 70 continuous feature set has been designed, based on Histograms of Oriented Gradient (HOG) [13] extracted from each frame using a 8-pixels width sliding window. It is composed of 64 HOG features (8 directions from the frame divided into 2 columns $\times$ 4 rows), and 6 high level information features: the position of the vertical and horizontal centroids, the position of the highest and lowest black pixels in the frame, the distance between them, and the number of black pixels in the frame. This continuous representation is fed to the

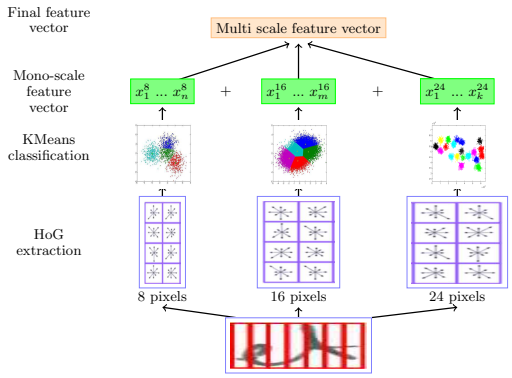

Fig. 2. Feature Extraction : Detail of every step during the feature extraction from the initial image to the final feature vector with multi-scale information of the word et.

unsupervised clustering stage allowing the definition of a discrete codebook. In our experiments, we explore the use of uni-gram, bi-gram and tri-gram codebooks extracted respectively from 1, 2 and 3 consecutive frames. Using KMeans clustering, three different codebooks are generated, providing three discrete representation levels of the input numerical data (see Fig. 2). After a validation

step, 1000, 2000 and 5000 clusters has been determined to be the optimal size for 1, 2 and 3 consecutive frames.

Finally, the CRF is fed with uni-gram, bi-gram and tri-gram codebooks in context:

- The unigram representation is composed of 9 cluster numbers (symbols): the current symbol and its 4 previous and next neighbours (I)
- The bigram representation is composed of 3 cluster numbers (symbols) computed from frames $[t, t+1]$ and frames $[t-1, t]$ (II)
- The trigram representation is composed of 1 cluster numbers (symbols) computed from frames $[t-1, t, t+1]$ (III)

We evaluated the following configurations: (I), (I+II) and (I+II+III) (see Table 1).

4.2 Results and Discussion

The CRF training converged in 80 iterations of 135s each (average value). We carried out the experiments on the public RIMES 2009 database of isolated words [5]. The participants were given about 43000 words snippets to train their system, and a validation database of more than 7000 words to test them. The unknown test dataset is composed of 7464 snippets. The system is evaluated on this test dataset with a lexicon of 1600 entries. The results of our experiment are summarized in Table 1. We provide the frame error rate (FER) in Top 1, and the word error rate (WER) in Top 1, Top 2, Top 3 and Top 5 of the whole system.

Table 1. Results on Rimes database

Features	FER Top 1	WER Top 1	WER Top 2	WER Top3	WER Top5
HMM (standard HOG)	88.6 %	36 %	32 %	30 %	25 %
CRF-HMM (I)	53.8 %	38 %	33 %	29 %	21 %
CRF-HMM (I+II)	52.5 %	34 %	29%	25 %	18 %
CRF-HMM (I+II+III)	51.2 %	31 %	29%	23 %	18 %
BLSTM-HMM [4]	33.93	12.19%	x	x	x

It can be seen that the multi-scale feature set improves the performance of our system at frame and word level. We observe an enhancement of 1.6% at frame level and 6% at word level between the set of features without multi-scaling information (I) and the set of features adding the bi-grams and tri-grams information. Figure 3 presents an example showing the ability of the model to perform a frame level recognition, and to retrieve the correct character alignment (shown in red) thanks to the HMM lexicon-driven decoding. Our best system achieves 69% word recognition (Top 1) which is under the best performance reported on this database. However, these are promising results if we look at

the potential improvements of the method. From our point of view, one of the main limitation of the system is that the CRF is trained on a frame-labelled dataset obtained from an initial Viterbi forced-alignment using an initial trained HMM. This means that the CRF is trained to recognize characters, but not to segment them. Some improvements are expected by introducing a lexicon-based training procedure of the proposed hybrid architecture. As a result, recognition and segmentation could be trained in conjunction. In addition, such scheme would allow to avoid training an initial HMM.

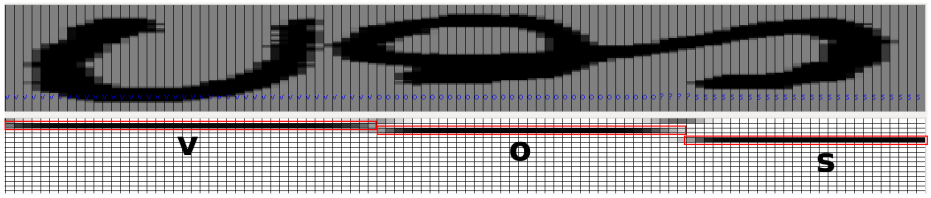

Fig. 3. Posteriors probabilities given by the CRF on the word "vos" and the alignment provide by the HMM.

In order to avoid similar wrong recognition events, we have to keep working on our features, try to find a better representation of our data. A major uncertainty we are faced with, is that we do not know if the segmentation performed by the HMM is suitable for the CRF. This is why we intend to design a system in which the CRF could impact the labelling processing of each frame during the learning stage.In order to achieve this we could introduce a joint training of the whole system CRF/HMM. After training the CRF a first time, the HMM produces a new alignment on the learning database using the CRF outputs. This new labelled database is used to retrain a new CRF. This two step learning method is repeated until the system stops improving the word recognition rate. By using this training method, the CRF outputs impact the global result of the system, and are not a simple byproduct of it, therefore improving the recognition of the CRF/HMM system.

Last but not least, in our CRF training the criterion is based on frame recognition rate, they are not trained to perform word recognition directly. To infer this information we have to add the word level information of the HMM stage in the training criterion of the standard CRF.

5 Conclusion and future work

In this paper, we have proposed a hybrid CRF/HMM model to perform off-line omni-writer handwriting recognition. We showed the architecture has promising performance even if the recognition rate is still below the best performance of the literature obtained on the same database.

Further improvements are expected by introducing embedded training of the hybrid model allowing joint training of the CRF and the HMM stage to perform both segmentation and character recognition, bypassing the need of an initial labelling.

Another expected improvement lies in the optimization of the HMM structure including character duration.

References

1. Yoshua Bengio, Y LeCun, and Y LeRec. Ann/hmm hybrid for on-line handwriting recognition. *Neural Computation*, 7(6):1289–1303, November 1995.
2. J Gauvain and Lee Chin-Hui. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing*, pages 291–298, April 1994.
3. A Graves, M Liwicki, S Fernandez, R Bertolami, H Bunke, and J Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *PAMI*, pages 855–868, May 2009.
4. Alex Graves, Santiago Fernández, Marcus Liwicki, Horst Bunke, and Jurgen Schmidhuber. Unconstrained online handwriting recognition with recurrent neural networks. *NIPS*, December 2007.
5. E Grosicki and H El Abed. Icdar 2009 handwriting recognition competition. *ICDAR*, 2009.
6. Asela Gunawardana, Milind Mahajan, Alex Acero, and John C. Platt. Hidden conditionnal random fields for phone classification. *InterSpeech*, 2005.
7. John Lafferty, Andrew McCallum, and Fernandon C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, June 2001.
8. Abdel-Rahman Mohamed, Dong Yu, and Li Deng. Investigation of full-sequence training of deep belief networks for speech recognition. *InterSpeech*, 2010.
9. Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latten-dynamic discriminative models for continuous gesture recognition. *CVPR*, 2007.
10. Ara V Nefian and Monson H Hayes III. Maximum likelihood training of the embedded hmm for face detection and recognition. *Image Processing*, 1:33–36, September 2000.
11. Ariadna Quattoni, Michael Collins, and trevor Darrel. Conditional random fields for object recognition. *NIPS*, December 2005.
12. Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), February 1989.
13. JA Rodriguez and F Perronin. Local gradient histogram features for word spotting in unconstrained handwritten documents. *ICFHR*, 2008.
14. Shravya Shetty and Harish Srinivasan. Handwritten word recognition using conditional random fields. *ICDAR*, pages 1098–1102, September 2007.
15. Todd A. Stephenson, H Bourlard, Samy Bengio, and Andrew C Morris. Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. *ICSLP*, 2:951–954, October 2000.
16. Charles Sutton and Andrew McCallum. Introduction to conditional random fields for relational learning. *Introduction to Statistical Relational Learning*, pages 94–126, 2006.
17. A Vinel, Trinh Minh Tri Do, and T Artieres. Joint optimization of hidden conditional random fields and non linear feature extraction. *ICDAR*, pages 513–517, September 2011.
18. G Zweig and P Nguyen. A segmental crf approach to large vocabulary continuous speech recognition. *Automatic Speech Recognition & Understanding*, pages 152–157, December 2009.