Benchmarking discriminative approaches for word spotting in handwritten documents

Gautier Bideault*, Luc Mioulet*, Clément Chatelain† and Thierry Paquet*
*Laboratoire LITIS - EA 4108, Universite de Rouen, FRANCE 76800
†Laboratoire LITIS - EA 4108, INSA Rouen, FRANCE 76800

***Abstract*—In this article, we propose to benchmark the most popular methods for word spotting in handwritten documents. The benchmark includes a pure HMM approach, as well as hybrid discriminative methods MLP-HMM, CRF-HMM, RNN-HMM and BLSTM-CTC-HMM. This study enables us to observe the increase ratio of performance provided by each discriminative stage compared with the pure generative HMM approach. Moreover, we put forward the different abilities of all these discriminative stages from the simplest MLP to the most complex and current state of the art BLSTM-CTC. We also propose a more specific and original study on BLSTM-CTC, showing that when used as a lexicon-free recognizer, it can reach very interesting word-spotting performance.**

***Keywords*—Benchmark, word spotting, Handwriting recognition, CRF, CRF/HMM, BLSTM/CTC, hybrid systems.**

I. Introduction

Word spotting consists in detecting a given keyword or a set of keywords in a whole document image, or a set of document images. Detecting keywords can be useful for higher level processing stages, such as document categorization [1], customer identification [2], Named Entity detection, or simply to find a relevant section in a huge quantity of information. This problem has been extensively addressed in document images these last years, using different approaches [3], [4], [5], [6], [2]. All these systems can be classified into two main approaches : matching approaches and recognition based approaches.

Matching approaches are suited for image based query. They consist in extracting morphological features such as ascenders, descenders, horizontal and vertical strokes, ovals, etc. of the query image and then compute a distance to the known images of the database. This approach is rather straightforward because it only requires to locate words in the dataset but no character segmentation nore classification stage is needed. It is fast and simple to set up. However, it cannot cope with multi-scriptor problem or highly variable data such as handwritten document images.

Recognition based approaches include a recognition step using a classifier, which enables them to be more robust to a multi-scriptor and noisy context. These approaches are refered as ”query as string” approaches in the literature because of the introduction of a recognition stage. However, they require training of a classifier on a labelled dataset.

In addition, these approaches can be divided into two categories : word based and line based approaches.

Word based approaches as proposed in [7] consists in two steps. First, a segmentation is applied on the whole document in order to locate every words in the document. Then words are latched against the query using either a pure matching approach or a recognition based matching approach. The problem in this case is that word localisation errors cannot be recovered by the recognition step.

Line model approaches [5], [6], [2], [4], [8] try to overcome this limitation by modelling the whole text line. The line model contains the model of the word to be spotted (the string query), surrounded by filler models. These filler models are ergodic models of characters that model any other possible sequence of character. Some works are based on pure HMM approach [5], [6], [2], and have proposed interesting results. However, hybrid structures have proved to be a powerful alternatives to address the word spotting task. There are usually made of a discriminative stage which deals with the image representation (features) and a generative stage which embeds high level information such as lexicons or language models [3], [4], [8]. The discriminative stage is generally made of a neural network classifier, while the generative stage is most of the time made of a HMM model.

In this paper, we compare various hybrid methods for the handwritten word spotting task. This benchmark uses the Rimes Database [9]. The benchmark includes: i) hybrid neural network/HMM, with either MLP or recurrent neural network, ii) the fully neural based state-of-the-art BLSTM/CTC, and iii) a specific discriminative hybrid CRF/HMM structure. The benchmark is also include a standard generative HMM method. These methods are fairly compared using the same input features and pre-processing steps. This paper is organized as follows: the section 2 presents a review of the related works, as well as a brief description of the selected methods. The benchmark protocol and the results using the Rimes Database [9] is presented in the section 3.

II. Related Works: Hybrid structures for word spotting

For a long time, pure HMM methods were state of the art for the handwritten word spotting task [10], [11], [12], [13], [14]. However these models suffer from the observation conditional independence assumption and their generative modeling ability.

In the early nineties, hybrid architectures were introduced to combine both discriminative and generative methods. They were at first dedicated to speech recognition using

ANN (mostly Multi Layer Perceptron)/HMM hybrid architectures [15]. These hybrid models have also been applied to handwriting recognition [16]. In most cases, these models are made of a neural network discriminative stage to compute and classify local observations at frame level, whereas the HMM generative stage is fully dedicated to combine the frame decision level using higher level information such as a lexicon or a language model. Many neural network architectures have been proposed in this spirit such as DNN/HMM [4], but more original CRF/HMM hybrid structures has also been proposed [17].

Recently, the BLSTM-CTC hybrid structure has been proposed, and has shown to perform extremely well for sequence classification [18]. Experiments using BLSTM-CTC framework for word spotting have also been reported [19], [8]. In [19], the authors proposed a BSTLM-CTC approach and they report very promising results on the IAM-database. In this system, the high level knowledge such as lexicon or language model are manage at the CTC level in order to constrain the alignment of the BLSTM outputs.

In this paper, we propose to conduct a fair benchmark on a word spotting task using Rimes database [9], using all these popular approaches [18], [16], [17], [4]. For this purpose we adopted a common modular hybrid model for word spotting. The high level structure of the model is designed using the HMM framework and is shared by every hybrid structure whereas the low level part of the model varies, depending of the classification framework.

In this paper, the following low level classification structures are explored : MLP, RNN, CRF and BLSTM. Experiments are carried out on the Rimes 2011 database [9]. For a fair comparaison of the performance, we use the same pre-processing and the feature representation stages, for any of the experiments. These are presented in the next section below.

III. Hybrid Models

In order to perform a fair benchmark, all the methods are evaluated using the same preprocessing and the same feature space. We emphasize that all hyperparameters of methods have been tuned over the validation set.

A. Details of the preprocessing

The image preprocessing are applied over the whole lines of text, and are made of three classical steps. First we applied a Sauvola binarization by thresholding. Then deslanting is performed. Finally, we applied a height normalization of 64 pixels, in order to center the baseline and normalize the heights of the ascenders and descenders.

B. Features detection and representations

Histograms of Oriented Gradient (HOG) [20] have been chosen for this experimentation. They are extracted from sliding windows of 8 pixels width and 64 pixels height. Each windows is divided into sub-windows of 4 pixels width and 16 pixels height. In each sub-window the histogram of intensity gradient in the 8 directions is computed. The 16 HOG are finally concatenated into a single 64 dimension feature vector.

C. Low level classifcation stages

As mentioned earlier, every hybrid structures share the same high level model designed for word spotting, whereas their low level character model is implemented in various fashions.

1) GMM-HMM character model: The basic GMM-HMM model is evaluated follows a standard structure. It is composed of 4 states left-right HMM with 20 Gaussian models per state. It was trained using the Baum-Welch algorithm on the whole training database.

2) MLP-HMM character: The MLP stage is composed of a single hidden layer of 80 neurons using an intern hyperbolic tangent function. It was trained using back propagation of the gradient using rnnlib[21]. A frame level ground-truth was previously generated in order trained this system using forced alignment. This stage is combined into a hybrid structure with a one state per character HMM model. The same frame level ground-truth and HMM stage are used for every other discriminative methods described below.

3) CRF-HMM character: In their initial formulation CRF [22], [23] are not able to cope with raw numerical data. We choose to take advantage of the ability of CRFs to deal with a huge amount of discrete features (several thousands in the case of language processing), and extended this framework to n-gram features : n-gram feature codebooks. Similarly to training Deep Neural Network (DNN), feature codebooks are trained in an unsupervized classifier, in order to minimizing the mean square error on the training dataset, using either k-means, or LindeBuzoGray clustering. Thanks to this stage, we can provide our CRF with a high dimensional symbolic feature codebook representation of the data (see Fig. III-C3). In this experiment, we propose the use of uni-gram, bi-gram and tri-gram feature codebooks extracted from 1,2, and 3 consecutive frames.

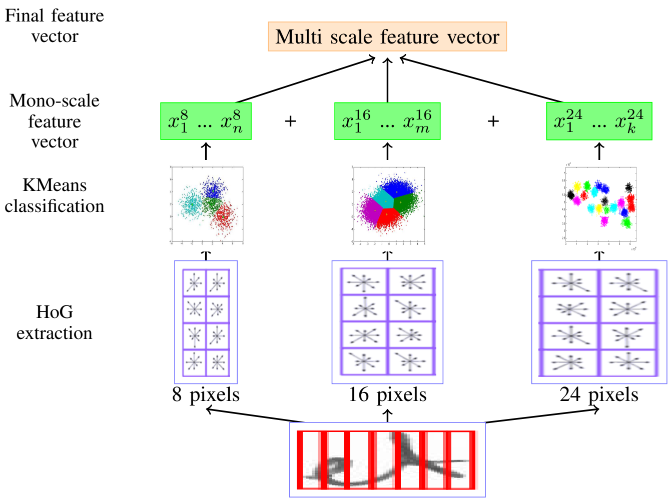

Fig. 1. Feature Extraction : Detail of every step of the creation of the high dimensional symbolic feature codebook from the raw image to the final feature vector of the word ”et”.

A validation step, led us to choose 1000, 2000 and 5000 clusters for coding uni-gram, bi-gram and tri-gram, respectively. This multi-level discrete representation is then fed to the CRF. Each frame of the CRF exploits the various feature codebooks as follows :

- a context of 9 uni-gram features is used (current unigram feature) with the 4 previous and the 4 next features
- a context of 2 bi-gram features is used (the current bigram feature spanning over $[t-1, t]$ and the next bigram spanning over $[t, t+1]$
- a single trigram feature spanning of $[t-1, t, t+1]$

The best combination as proposed in [17] is the combination (I)+(II)+(III), thus providing each frame with 12 (9+2+1) binary features.

4) RNN-HMM: The recurrent neural network is made of a single hidden layer of 80 recurrent neurons using an intern hyperbolic tangent function. It was trained using back propagation of the gradient through time using rnnlib[21].

5) BLSTM-CTC-HMM: Our BLSTM-CTC is composed of 2 hidden layers of 70 and 120 LSTM neurones. It was trained using back propagation of the gradient through time using rnnlib[21].

D. High level HMM word spotting model

This model is designed in order to model the possible occurrence of a specific keyword in a text line. Any spotting model integrates two models : the model of the keyword to be spotted and the model of any other possible word that can occure within a line. This second model, also called ”filler model” in the litterature, accounts for the occurrence of any other possible sequence of characters. It is generally made of an ergodic HMM model of characters. The overal structure of the spotting model is depicted in Figure 2, coupled with a BLSTM-CTC as a low level stage. Within the HMM framework the detection stage is usually performed using two Viterbi decoding stages : a first score is generated using the spotting model, then a second decoding is performed using only the filler model. The likelihood ratio of the two models serves as a detection score of the keyword to be spotted. Acceptance of the keyword is made if the score is higher than a predefined threshold. Within the discriminative framework introduced by any of the hybrid models. Only one Viterbi decoding step is necessary using the spotting model. Then the score of each spotted hypothesis is computed by the average character posterior probabilities of the hypothesis divided by the number of frames spanning the hypothesis. Indeed in this case, the decision is directly derived from the normalize outputs of the low level discriminative stage.

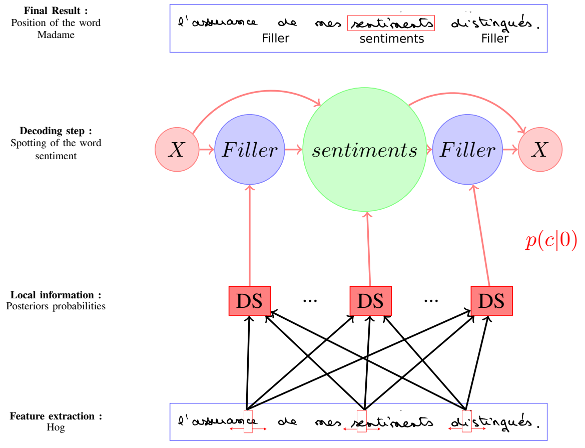

Fig. 2. Hybrid structure BLSTM/HMM : Details of every module of the hybrid structure from feature extraction to the extraction of the position of the word **sentiments** in the sentence for any Discriminative Stage (DS) : MLP, RNN, CRF, BLSTM. Posteriori probabilities of the choosen discriminative method are fed into the HMM line model.

IV. Benchmark

The evaluation dataset is the RIMES database used for the 2011 ICDAR handwriting recognition competitions [9]. The training database is composed of 1.500 documents, the validation and test sets are composed of 100 documents each. Recall (R) and Precision measures (P) are measured to evaluate the system, by counting the number of true positives (TP), false positives (FP), and false negatives (FN) using variable threshold values. From these values, a recall-precision curve is computed by accumulating these values over all word queries.

$$R = \frac{TP}{TP + FN} \quad P = \frac{TP}{TP + FP} \tag{1}$$

A. Word spotting results

The most common process to evaluate a word spotting system is to spot one word in a collection of documents. In these experiments, we decided to evaluate our framework by detecting a set of keywords (a keyword lexicon) at the same time, allowing to deal with confusion between words [5]. Spotting a set of keyword instead of a single word enable document categorization [1]. We evaluate the performance of the five systems on lexicon of size 25, 50 and 100 keywords respectively, following the same data-set and protocol as in *Kessentini et al.*[5].

Figure 3, 4 and 5 show the recall-precision curves obtained for the five system for each lexicon (25, 50 and 100). As expected the overall results of the five methods decrease with the size of the lexicon. We can see that the BLSTM-HMM architecture shows the best results with break-even points around 85% (the point where Recall=Precision). It is followed by the RNN-HMM system, the CRF-HMM, the MLP-HMM and finally the pure HMM system. We can see how these various discriminative approaches contribute to the spotting task and how they compare with each other. As already demonstrated in other studies, the BLSTM-HMM structure is by far the best method. Then, interestingly, we see the CRF framework slightly more accurate than a standard MLP (gap of 5%). We also see that a standard recurrent neural network structure perform 5% better than a CRF. These results also show the significant contribution of a recurrent structure in comparaison with non recurrent classification stages.

The CRF provides slightly more accurate results than a standard MLP method (gap of 5%). However, the RNN is 5% ahead of the CRF, it seems that adding recurrent neurons is sufficient to outperform the CRF.

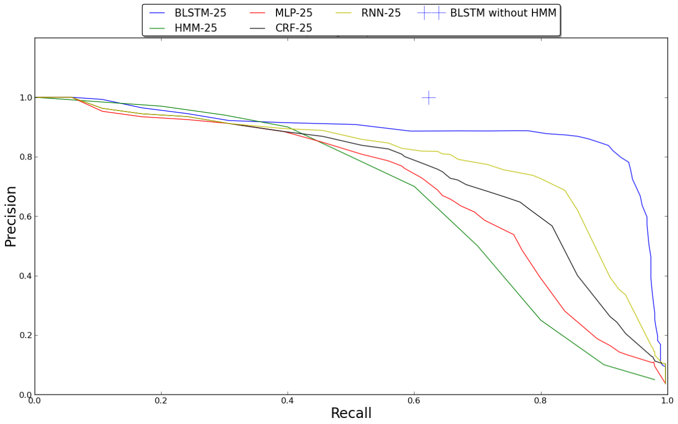

Fig. 3. Performance on a keyword lexicon of 25 keywords.

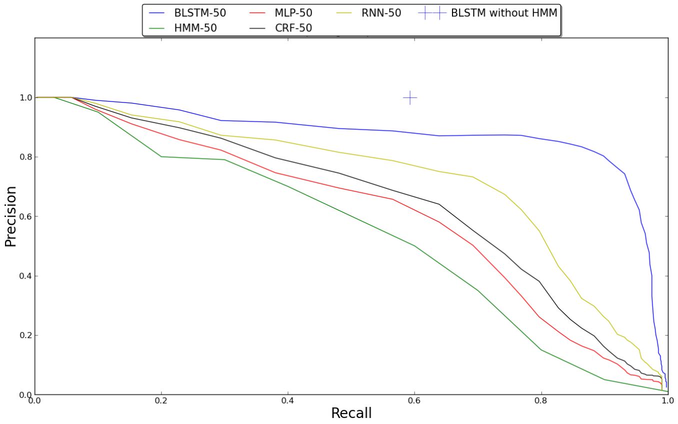

Fig. 4. Performance on a keyword lexicon of 50 keywords.

B. Frame Error Rate

During the training phase it is important to monitor the convergence of the system to avoid over-fitting. This is the reason why we tried to evaluate the relation between the frame error rate (at every epoch of the training algorithm) and the break-even point of every methods (final result). In this experiment, we compute the average the break-even points of the five methods for all the lexicons. Results are shown in the Figure 6.

We can see that the frame error rate is directly related to the global recognition performance of the systems. The lower the frame error rate, the better the break-even point. Based on this observation on this database, the frame error rate is a good target criterion to control the quality of hybrid architecture during the training phase.

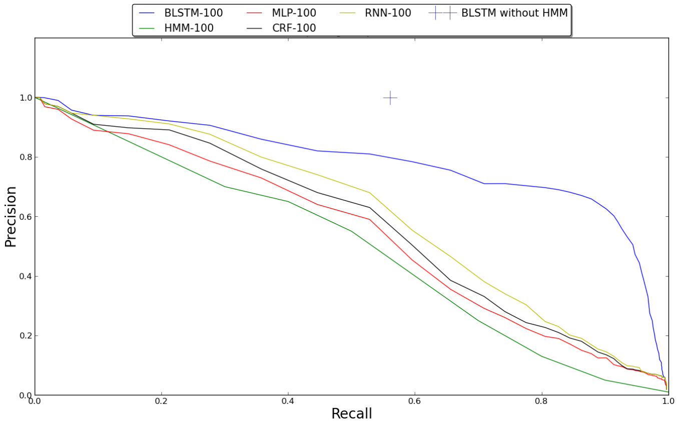

Fig. 5. Performance on a keyword lexicon of 100 keywords.

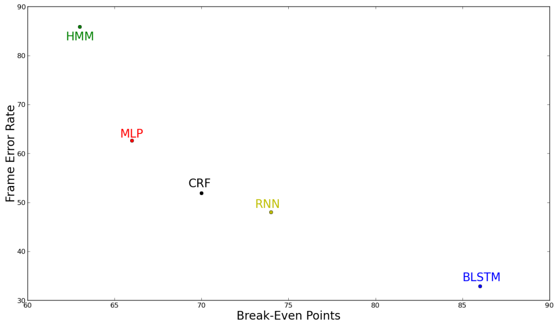

Fig. 6. Relation between Frame Error Rate and the Break-Even Points

C. Decoding without the HMM stage

Recall-Precision curves allow the user to choose the most appropriate threshold for his problem. Indeed, some applications need to maximize the precision whereas some others may privilege the Recall. That is why we perform a word spotting experiment without introducing the higher level HMM stage spotting model. This stage correspond to analyzing the raw transcriptions provided by the low level decision stage and then matching the searched keywords on this transcription. Results are shown in table I. Those particular point are also on previous figures 3, 4 and 5, represented by blue dots.

In most of the applications, you can improve your global results thanks to a language model, a lexicon or any kind of high level information. By performing this experiment, we once again prove the powerful capacity of the BLSTM-CTC to tackle the Sayre paradox as it is able to segment and recognize characters very accurately. We remind that these recall score are coupled with 100 % of precision which is a huge performance knowing the handwritten context. The BLSTM-CTC seems to be able perform recognition of out vocabulary elements such as named entities for example.

TABLE I. RESULTS OF WORD SPOTTING ON LEXICON OF 25, 50 AND 100 KEYWORDS WITHOUT THE HMM STAGE.

Method	Lexicon size	Recall	Precision
MLP-HMM	25	5.2%	100%
MLP-HMM	50	3.7%	100%
MLP-HMM	100	2.2%	100%
CRF-HMM	25	8.5%	100%
CRF-HMM	50	7.3%	100%
CRF-HMM	100	4.8%	100%
RNN-HMM	25	10.2%	100%
RNN-HMM	50	8.9%	100%
RNN-HMM	100	6.2%	100%
BLSTM-HMM	25	62.3 %	100 %
BLSTM-HMM	50	59.3 %	100 %
BLSTM-HMM	100	56.1 %	100 %

V. CONCLUSION

In this paper, we have benchmarked five different models for handwritten word spotting : pure HMM, MLP-HMM, CRF-HMM, RNN-HMM, BLSTM-CTC-HMM. We used three different lexicons of size 25, 50 and 100 keywords. We showed that recurrent neural methods cope better with handwritten documents than the no recurrent ones. It once again prove the need to take context through time into account. More precisely, we observe that the BLSTM-CTC is the current best method to solve this kind of problem. This structure showed break-even points at more than 80% even on lexicon of 100 keywords.

The CRF-HMM hybrid structure performs better than the pure HMM and the MLP-HMM but cannot compete with recurrent neural networks. Some additional results show that the BLSTM-CTC provides interesting performance even when no additional constraints are introduced, since it provides a recall higher than 55 %. These experiments show that the frontiers of processing handwritten documents are becoming more and more closer to those of processing printed or born digital documents which offers many perspectives for developing applications dealing with handwritten documents in the bear future.

REFERENCES

[1] T. Paquet, L. Heutte, G. Koch, and C. Chatelain, "A categorization system for handwritten documents," *International Journal on Document Analysis and Recognition*, vol. 15, no. 4, pp. 315–330, 2012.

[2] C. Chatelain, L. Heutte, and T. Paquet, "A two-stage outlier rejection strategy for numerical field extraction in handwritten documents," in *ICPR, Hong Kong, China*, vol. 3, 2006, pp. 224–227.

[3] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 2, pp. 211–224, 2012.

[4] S. Thomas, C. Chatelain, L. Heutte, T. Paquet, and Y. Kessentini, "A deep hmm model for multiple keywords spotting in handwritten documents," in *to appear in Pattern Analysis and Applications*, 2014.

[5] Y. Kessentini, C. Chatelain, and T. Paquet, "Word spotting and regular expression detection in handwritten documents," in *ICDAR*, 2013.

[6] M. E. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Segmentation and recognition of handwritten dates: an hmm-mlp hybrid approach." 2003, pp. 248–262.

[7] R. Manmatha, "Multimedia indexing and retrieval research at the center for intelligent information retrieval," in *Proceedings of the 1997 Symposium on Document Image Understanding Technology*, 1997, pp. 16–30.

[8] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "A tandem blstm-dbn architecture for keyword spotting with enhanced context modeling," in *Proc. of NOLISP*, 2009.

[9] E. Grosicki and H. El-Abed, "Icdar 2011-french handwriting recognition competition," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1459–1463.

[10] T. Artières and P. Gallinari, "Stroke level hmms for on-line handwriting recognition," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 227–232.

[11] U.-V. Marti and H. Bunke, "Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system," *International journal of Pattern Recognition and Artificial intelligence*, vol. 15, no. 01, pp. 65–90, 2001.

[12] J. Hu, M. K. Brown, and W. Turin, "Hmm based online handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 10, pp. 1039–1045, 1996.

[13] L. M. Lorigo and V. Govindaraju, "Offline arabic handwriting recognition: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 5, pp. 712–724, 2006.

[14] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 1, pp. 63–84, 2000.

[15] T. A. Stephenson, H. Bourlard, S. Bengio, and A. C. Morris, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," *ICSLP*, vol. 2, pp. 951–954, October 2000.

[16] Y. Bengio, Y. LeCun, and Y. LeRec, "Ann/hmm hybrid for on-line handwriting recognition," *Neural Computation*, vol. 7, no. 6, pp. 1289–1303, November 1995.

[17] G. Bideault, L. Mioulet, C. Chatelain, and T. Paquet, "A hybrid crf/hmm approach for handwriting recognition," in *Image Analysis and Recognition*. Springer, 2014, pp. 403–410.

[18] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2008, pp. 577–584.

[19] V. Frinken, A. Fischer, and H. Bunke, "A novel word spotting algorithm using bidirectional long short-term memory neural networks," in *Artificial Neural Networks in Pattern Recognition*, ser. Lecture Notes in Computer Science, F. Schwenker and N. El Gayar, Eds. Springer Berlin Heidelberg, 2010, vol. 5998, pp. 185–196.

[20] J. A. Rodrıguez and F. Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," in *Int. Conf. on Frontiers in Handwriting Recognition*, 2008.

[21] A. Graves, "Rnnlib: A recurrent neural network library for sequence learning problems," 2013.

[22] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[23] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, pp. 93–128, 2006.