Co-Training for Handwritten Word Recognition

Volkmar Frinken, Andreas Fischer, and Horst Bunke
Institute of Computer Science and Applied Mathematics
University of Bern
Neubrückstrasse 10, 3012 Bern, Switzerland
{*frinken,afischer,bunke*}*@iam.unibe.ch*

Alicia Fórnes
Centro de Visio per Computador
Universidad Autonoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{*afornes*}*@cvc.uab.es*

***Abstract*—To cope with the tremendous variations of writing styles encountered between different individuals, unconstrained automatic handwriting recognition systems need to be trained on large sets of labeled data. Traditionally, the training data has to be labeled manually, which is a laborious and costly process. Semi-supervised learning techniques offer methods to utilize unlabeled data, which can be obtained cheaply in large amounts in order, to reduce the need for labeled data. In this paper, we propose the use of Co-Training for improving the recognition accuracy of two weakly trained handwriting recognition systems. The first one is based on Recurrent Neural Networks while the second one is based on Hidden Markov Models. On the IAM off-line handwriting database we demonstrate a significant increase of the recognition accuracy can be achieved with Co-Training for single word recognition.**

***Keywords*-Semi-supervised Learning, Co-Training, Handwriting Recognition, Single Word Recognition, HMMs, BLSTM NN**

I. Introduction

Off-line handwriting recognition is the task of recognizing a handwritten text from a sheet of paper that was scanned, photographed or digitized otherwise. Opposed to on-line handwriting recognition, where temporal and spatial information about each stroke is available, the off-line recognition task is performed using only the image of the written text. Many important applications are based on off-line handwriting recognition, e.g. postal address identification [1], Bank check processing [2], prescreening of handwritten notes [3], and the creation of digital libraries of historical documents [4]. After several decades of ongoing research, however, off-line handwritten text recognition is still considered a difficult problem that is only partially solved [5].

To create an automatic handwriting recognition system, a set of images of handwritten text along with their correct transcription is needed for training. As it turns out, one of the key problems encountered when building a writer independent recognition system[1] is the great variety in writing styles between different writers. Hence, the amount of training data needed is extremely large. Unfortunately, the transcription of the handwritten text has to be done manually which makes the acquisition of training data costly and time consuming. On the other hand, collecting handwritten samples itself can be done very efficiently. Thus, although the computational complexity of Co-Training can be quite large, it runs completely off-line without any human interaction.

Consequently, unlabeled data can easily be made available in large amounts. Hence the question arises whether such unlabeled data can be helpful for handwriting recognition systems. It has been shown that in various classification scenarios unlabeled examples can indeed significantly improve the recognition accuracy using semi-supervised learning [6]. Most of the existing works deal with the standard classification scenario where a single point in a feature space has to be mapped into the label space [7]. The cursive, sequential nature of handwritten text put common approaches closer to speech recognition than OCR. Semi-supervised learning methods for handwritten text are restricted to frameworks that are general enough to make use of sequential data. In this paper, we propose to use Co-Training [8] to improve the recognition accuracy of weakly trained recognizers using unlabeled data. Co-Training is a semi-supervised learning paradigm under which two recognizers improve each other. This is done by retraining one recognizer with elements confidently recognized by the other recognizer and vice versa. Different approaches that make use of unlabeled data have been proposed before. In [9], [10] the authors adapt a recognition system to a single person by using unlabeled data. This system is highly specialized after the adaptation and not suitable for general handwriting recognition, though. Improving a single recognition system using Self-Training was proposed in [11], [12]. To the knowledge of the authors, this paper is the first report on using Co-Training for unconstrained recognition of handwritten words.

Two sets of experiments on single word recognition with a different size of labeled data are performed using a Hidden Markov Models and a neural network based recognition system. Several rules that determine which elements are used for retraining are investigated. Choosing too few confidently recognized words for retraining does not influence the original training set substantially. Selecting more data can only be done with the risk of adding noise to the training data. For a

[1] A writer independent system is one that recognizes text from writers that have not contributed to the training set.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.71

retraining rule that balances data quality and data quantity, a significant increase in both systems' recognition accuracies can be observed.

The remainder of this paper is structured as follows. The principles of the Co-Training are presented Section II. Details about the word recognition systems as well as the applied preprocessing steps are explained in Section III. Section IV-A covers the techniques for estimating the recognition confidences. An experimental evaluation of the proposed approach is given in Section V and conclusions are drawn in Section VI.

II. Co-Training

Originally proposed in [8], Co-Training states that two recognizers can train each other successfully, given the prerequisite that both recognizers have a conditionally independent view of the data. In other words, for a given class, the features used for one classifier must nor correlate with the features used for the other classifier. If this condition holds, an infinite amount of unlabeled data can be used to gradually reduce the classification error down to the Bayes risk. If a data point is classified by only one classifier with a high confidence, it is used to train the other classifier. Consequently, each classifier is presented with data that is both, very likely to be correct and highly relevant.

Unfortunately, that is the ideal case and a feature split which is conditionally independent is not very likely to exist. Instead, the Co-Training conditions are further relaxed by using the same set of features but two different recognizers with a different inherent bias, which has also been proven to work in [13]. In this paper, as shown in Fig. 1, two different recognizers, one based on Hidden Markov Models, one based on BLSTM neural networks, recognize the set of unlabeled data independently. After using post-processing methods to estimate a reliable confidence measure, a filter function is applied to select confidently recognized elements. Then, all such elements of the recognized set of one recognizer are added to the other recognizer's training set. Two different training sets are used, one for each recognizer. This is in contrast to proposed methods in the literature. In this setup more emphasis is put on the exchange of classification decisions than on the reinforcement achieved by self-training.

III. Recognition Systems

A. Preprocessing

The database used in this paper consists of 1,539 pages of handwritten English text, written by 657 writers[2] [14]. All pages of the database are already segmented into individual text lines. The segmented text lines are normalized prior to recognition in order to cope with different writing styles. First, the skew angle is determined by a regression analysis based on the bottom-most black pixel of each pixel column. Then, the skew of the text line is removed by rotation. Afterwards the slant is corrected in order to normalize the directions of long vertical strokes found in characters like 't' or 'l'. After estimating the slant angle based on a histogram analysis, a shear transformation is applied to the image. Next, a vertical scaling is applied to obtain three writing zones of the same height, i.e., lower, middle, and upper zone, separated by the lower and upper baseline. To determine the lower baseline, the regression result from skew correction is used, and the upper baseline is found by vertical histogram analysis. Finally the width of the text is normalized. For this purpose, the average distance of black-white transitions along a horizontal straight line through the middle zone is determined and adjusted by horizontal scaling. For more details on the text line normalization operations, we refer to [15].

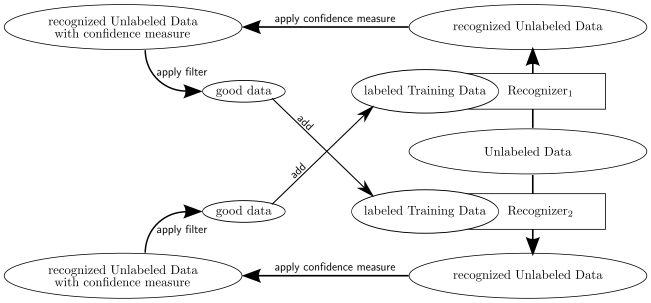


Figure 1. The approach to Co-Training investigated in this paper.

B. BLSTM NN based Handwriting Recognition

The recognizer used in this paper is a recently developed recurrent neural network, termed *bidirectional long short-term memory* (BLSTM) neural network [16]. A hidden layer is made up of so called *long short-term memory* blocks instead of simple nodes to circumvent the exponential increase or decay of information that is encountered in common recurrent neural networks.

The network is *bidirectional*, i.e. a sequence is fed into the network in both the forward and the backward mode using two separate input and hidden layers, joined in one output layer. The output layer contains one node for each possible character in the sequence plus a special ε node, to indicate "no character". At each position, the output activations of the nodes are normalized so that they sum up to 1, and can hence be treated as posterior probabilities of the characters' occurrences at each position.

The output of a sequence is therefore a matrix of probabilities and a path through the matrix represents a recognition. The probability of a recognition is given as the product of each element along the path and the optimal path can be found efficiently using dynamic programming. For more details about BLSTM networks, we refer to [16], [17].

[2] http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

C. HMM based Handwriting Recognition

Hidden Markov Models (HMMs) [18] are a time-discrete stochastic process, defined by a set of internal states, their transition probabilities as well as for each state a probability distribution function over a set of emitting symbols. The process works in two steps. At each time step, the system changes the internal state by respecting the Markov property. In each state, a symbol is emitted according to the current states output distribution function.

In the proposed approach, each character is modeled by a set of states, arranged in a linear topology. This way, a word is given by the concatenation of the character's states. Given a set of labeled training sequences, the transition and output probabilities can be learned efficiently using the Baum-Welch algorithm. To recognize an unknown word, the n most likely character sequences can be inferred using a Token Passing algorithm. An additional dictionary is used to only consider character sequences that form possible words.

IV. Retraining the System

A. Recognition Confidence

For Co-Training, the most confidently recognized elements are used for retraining. Therefore, a reliable confidence measure of the recognition is needed. Preliminary experiments as well as existing literature show that the pure recognition likelihood of the HMM recognizer can not be used as a confidence measure. In addition, it turned out that even the returned posterior probability of the neural networks is not reliable enough. Consequently we exploit the variability of different networks induced by their random initialization. For a more accurate recognition confidence, the following steps are applied. First, a preliminary estimate is computed differently for both recognizers. Then, the validation is used in the same way for both systems to increase the reliability of the recognition confidence.

The preliminary confidence measure for the HMM-based system is the likelihood ratio of the top two recognition results, mapped into the range between 0 and 1 according to the monotone function $1 - (1 + \frac{x}{f})^{-1}$, where f is a normalization constant. The second step requires a discrete input value, therefore the range between 0 and 1 is split into b bins of equal size.

In case of the neural networks, the natural variability due to the random initialization is exploited. In our experiments, 10 different neural networks are trained and the fraction of networks agreeing with the output of the best system, as tested on a validation set, is used.

In a next step, the preliminary recognition confidence $conf_{prel}$ of the word w is enhanced. Let $c \in \{0, 1\}$ indicate the correctness of a recognition. Then, $p_1(c|conf_{prel})$ indicates the probability of a recognition having the preliminary recognition measure $conf_{prel}$ being correct. In a further estimation, $p_2(c|conf_{prel}, w)$ also takes the recognized word w into account. For a robust computation, however, a minimum number of occurrences θ_v are needed. If, for example, the word 'abba' is recognized 12 times with a confidence measure of 0.5 on the validation set and 7 out of these recognition are correct, then $p_2(1|0.5, `abba') = \frac{7}{12}$. If, on the entire validation set, 100 words are recognized with a confidence measure of 0.5 and 65 of these are correct, then $p_1(1|0.5) = \frac{65}{100}$.

If the word 'abba' is now recognized on the test set with a preliminary recognition confidence of 0.5 and $\theta_v = 15$ is given, not enough elements exist on the validation set to estimate p_2 well enough and the final confidence is set to $p_1 = 0.65$. If not enough elements are given to estimate neither p_1 nor p_2, then the preliminary recognition confidence $conf_{prel}$ is used. More details on these steps can be found in [19].

B. Retraining Rules

Three different retraining rules based on the confidence are investigated. Each rule defines a confidence threshold and all words recognized with at least this threshold are added to the retraining set. The first threshold is called *High Threshold* and is set to the highest possible value 1. A second, more refined threshold is the *Medium Threshold*. It is set so that all elements added are more likely to be correct than wrong. This threshold is found by choosing the lowest value returning more correctly that incorrectly recognized samples in the validation set. The last threshold investigated is the *Low Threshold*. It is set to 0 so that all words, regardless of their recognition confidence, are added to the training set.

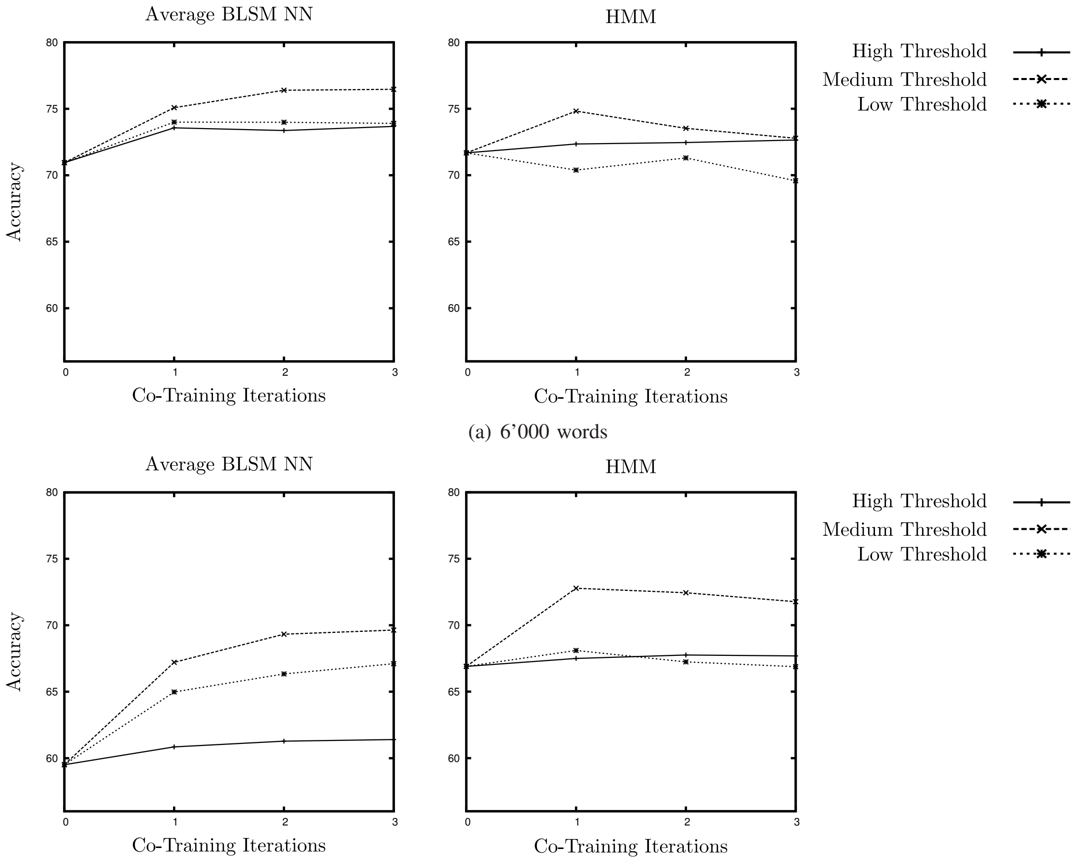
V. Experimental Evaluation

A. Setup

Experiments are conducted using all instances of the 4'000 most frequent words of the IAM offline database[14]. All correctly segmented words among the 4'000 most frequent words according to the LOB corpus [20] are considered. The three set, a working set (38'127 words), validation set (5'590 words) and training set (5'342 words) are writer disjunct, thus any person who contributed words to one of the three sets did not contribute to any of the other set.

To investigate the performance of Co-Training, two experiments are conducted using 2'000 labeled training words and 6'000 labeled training words, respectively. These sets are randomly sampled from the working set. The remaining 32'127, resp. 36'127 words act as the set of unlabeled data.

In each iteration, both recognizers decode the entire set on unlabeled data, the validation set, and the test set. The recognition results matching the retraining threshold of the HMM system are added to the training set of the BLSTM neural network and vice versa. To keep computational costs within reasonable bounds, the experiments are limited to 3 Co-Training iterations. The transformation parameters $\theta_v =$



(a) 6'000 words

(b) 2'000 words

Figure 2. Co-Training for Single Words

15, $f = 500$, and $b = 20$ were set according to preliminary experiments on the validation set.

B. Results

The plots in Fig. 2, show the results of these experiments. The left plots show the accuracy on the test set of the BLSTM neural networks trained on the output of the hidden Markov models, the right column show the accuracy of the hidden Markov models trained on the output of the BLSTM neural networks.

If retraining is done with only those elements whose correctness can be guaranteed as it is done using the *High Threshold*, the retraining set does not change significantly and the classifier may remain nearly the same. Enlarging the retraining set, on the other hand, is only possible at the cost of increasing noise, i.e. adding mislabeled words to the training set. With only few correctly recognized words and large amounts of possible mis-recognitions, the challenge of successful Co-Training lies in finding the optimal trade-off between data quality and data quantity for retraining. One can see that the *Medium Threshold* retraining rule constantly outperforms the other approaches. A statistically significant increase on the $\alpha = 0.05$ level is achieved. A reference system trained on the entire labeled working set achieved an accuracy of 84.39% (BLSTM NN), resp. 73.62% (HMM). After the first iteration, the HMM trained on 6'000 elements even surpassed the reference system. We assume that this happens because of badly written words within the 38'127 words that are filtered by the neural networks and lead to a poorer performance when included in the training for the HMM. Applying the *Medium Threshold* retraining rule, the HMM is trained in the first iteration with 30'011 words (including 6'000 words from the original training).

VI. Conclusion

In this paper we investigate the applicability of using Co-Training is investigated to increase the accuracy of weakly trained recognizers for handwritten word images. We demonstrate that it is possible to apply Co-Training to improve a Hidden Markov Model and a BLSTM neural network with unlabeled data. The performance increase of both systems are substantial. The experiments show that a system trained on 6'000 labeled elements only can be increased, using unlabeled data, to perform nearly as well as a system trained on 38'127 labeled words.

We also demonstrated the importance of selecting good elements for retraining. A conservative selection rule that focuses on very few, but likely correct elements does not influence the training set substantially. A selection rule that picks too many elements increases the noise in the training data, which might degrade the recognizers performance. A well balanced trade-off between data quantity and data quality is crucial for the success of Co-Training.

In the future, several improvements and subsequent research directions are possible. The most obvious one is to improve the recognition confidence measure, especially for the HMM based system. With a more sophisticated measure, a better improvement is possible. From the practical point of view, it is interesting to experiment with even less labeled data to make existing, large collections of historical data without any ground truth easily available.

Acknowledgment

The authors would like to thank Alex Graves for kindly providing us with the BLSTM Neural Network source code. This work has been partially supported by the EU project FP7-PEOPLE-2008-IAPP: 230653, by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03, CONSOLIDER-INGENIO 2010 (CSD2007-00018), the José Castillejo mobility research grant JC2010-0112, as well as by the Swiss National Science Foundation project CRSI22_125220.

References

[1] A. Brakensiek and G. Rigoll, *Reading and Learning*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2004, vol. 2956, ch. Handwritten Address Recognition Using Hidden Markov Models, pp. 103–122.

[2] R. Palacios, A. Gupta, and P. S. Wang, "Handwritten Bank Check Recognition Of Courtesy Amounts," *Int'l Journal of Image and Graphics*, vol. 4, no. 2, pp. 1–20, 2004.

[3] M. Ye, P. A. Viola, S. Raghupathy, H. Sutanto, and C. Li, "Learning to Group Text Lines and Regions in Freeform Handwritten Notes," in *Ninth Int'l Conf. on Document Analysis and Recognition*. IEEE Computer Society, 2007, pp. 28–32.

[4] V. Govindaraju and H. Xue, "Fast Handwriting Recognition for Indexing Historical Documents," in *First Int'l Workshop on Document Image Analysis for Libraries*. IEEE Computer Society, 2004, pp. 314– 320.

[5] H. Bunke, "Recognition of Cursive Roman Handwriting - Past, Present and Future," in *Proc. 7th Int'l Conf. on Document Analysis and Recognition*, vol. 1, Aug. 2003, pp. 448–459.

[6] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[7] X. Zhu, "Semi-Supervised Learning Literature Survey," Computer Science, University of Wisconsin-Madison, Tech. Rep. 1530, 2005, http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf.

[8] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," in *COLT' 98: Proc. of the 11th annual Conference on Computational Learning Theory*. New York, NY, USA: ACM, 1998, pp. 92–100.

[9] G. R. Ball and S. N. Srihari, "Prototype Integration in Off-Line Handwriting Recognition Adaptation," in *Proc. Int'l. Conf. on Frontiers in Handwriting Recognition*, 2008, pp. 529–534.

[10] ——, "Semi-supervised Learning for Handwriting Recognition," in *10th Int'l Conf. on Document Analysis and Recognition*, 2009, pp. 26–30.

[11] V. Frinken and H. Bunke, "Evaluating Retraining Rules for Semi-Supervised Learning in Neural Network Based Cursive Word Recognition," in *10th Int'l Conf on Document Analysis and Recognition*, 2009, pp. 31–35.

[12] ——, "Self-Training Strategies for Handwritten Word Recognition," in *9th Industrial Conference on Data Mining*, ser. Lecture Notes in Artificial Intelligence, 2009, pp. 291–300.

[13] S. A. Goldman and Y. Zhou, "Enhancing Supervised Learning with Unlabeled Data," in *17th Int'l Conf. on MAchine Learning*, 2000, pp. 327–334.

[14] U.-V. Marti and H. Bunke, "The IAM-Database: An English Sentence Database for Offline Handwriting Recognition," *Int'l Journal on Document Analysis and Recognition*, vol. 5, pp. 39–46, 2002.

[15] ——, "Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System," *Int'l Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, pp. 65–90, 2001.

[16] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequential Data with Recurrent Neural Networks," in *23rd Int'l Conf. on Machine Learning*, 2006, pp. 369–376.

[18] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[19] R. Bertolami, M. Zimmermann, and H. Bunke, "Rejection Strategies for Offline Handwritten Text Line Recognition," *Pattern Recognition Letters*, vol. 27, no. 16, pp. 2005–2012, 2006.

[20] S. Johanson, G. N. Leech, and H. Goodluck, "Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with Digital Computers," Department of English, University of Oslo, Norway, Tech. Rep., 1978.