Comparison of Bernoulli and Gaussian HMMs using a vertical repositioning technique for off-line handwriting recognition

Patrick Doetsch, Mahdi Hamdani and Hermann Ney
Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University
Aachen, Germany
{*doetsch,hamdani,ney*}*@i6.informatik.rwth-aachen.de*

Adrià Giménez, Jesús Andrés-Ferrer and Alfons Juan
DSIC
Universitat Politècnica de València, Camí de Vera s/n
València, Spain
{*agimenez,jandres,ajuan*}*@dsic.upv.es*

***Abstract*—In this paper a vertical repositioning method based on the center of gravity is investigated for handwriting recognition systems and evaluated on databases containing Arabic and French handwriting. Experiments show that vertical distortion in images has a large impact on the performance of HMM based handwriting recognition systems. Recently good results were obtained with Bernoulli HMMs (BHMMs) using a preprocessing with vertical repositioning of binarized images. In order to isolate the effect of the preprocessing from the BHMM model, experiments were conducted with Gaussian HMMs and the LSTM-RNN tandem HMM approach with relative improvements of 33% WER on the Arabic and up to 62% on the French database.**

***Keywords*-handwriting recognition; vertical distortion; center of gravity; recurrent neural networks; Bernoulli HMMs**

I. Introduction

According to the current state of the art [9], off-line handwriting recognition systems is still a challenging task with room for improvement. The choice of feature extraction and classification techniques is a very important step in the design of the recognizer. Hidden Markov Models (HMMs) are successful in handwriting recognition systems [11]. In particular, Bernoulli HMMs and Gaussian HMMs (GHMMs) had recently reported very good results on Arabic handwriting recognition [4], [11], [12]. Results reported for BHMMs were obtained using a novel feature extraction process in which input images were binarized and afterwards a vertical repositioning of a sliding window was applied. In contrast, the results reported by GHMMs were obtained in combination with a special type of Recurrent Neural Networks: Long Short Term Memory (LSTM); instead of using the vertical repositioning. Therefore, the main objective of this paper is to determine whether the good results given by the BHMMs are due to the use of the Bernoulli mixtures, the binarization of input images or the vertical repositioning of features.

In order to achieve such isolation, we compare three models: BHMM, GHMM and GHMM/LSTM classifiers. The same feature extraction processes was applied to each classifier. We compare the effect of vertical repositioning, binarization and both. Due to the nature of BHMMs employed features for BHMMs are always binary.

This paper is organized as follows, Section 2 presents the used repositioning method for preprocessing. The different systems are described in Section 3. Finally, a comparison of the results is given in Section 4 followed by the conclusions.

II. Center of Gravity Repositioning (CoG)

Given a (binary) image normalized in height to H pixels, we may think of a feature vector $\mathbf{o}_t$ as its column at position t or, more generally, as a concatenation of columns in a window of W columns in width, centered at position t. This generalization would be very helpful to better capture the image context at each horizontal position of the image. However, HMMs for image modeling are somewhat limited when dealing with vertical image distortions, and this limitation might be particularly strong in the case of feature vectors extracted with significant context. To overcome this limitation, we first compute the center of gravity (CoG) of each extracted window. Afterwards we reposition each window for each center to be to vertically aligned to the center of gravity. A synthetic example of feature extraction is shown in Figure 1 in which the the standard method (no repositioning) is compared with the vertical repositioning method.

Previous to the proposed feature extraction the images are scaled to a fixed height while respecting the original aspect ratio. Finally, if a binary input is expected, i.e. BHMMs, then they are binarized using Otsu's method.

III. Bernoulli HMMs

A Bernoulli HMM (BHMM) is an HMM specifically defined to deal with binary data [4], in which the emission probability function in each state is modeled as a Bernoulli mixture model as follows

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} \tau_{jk} \prod_{d=1}^{D} p_{jkd}^{o_{td}} \left(1 - p_{jkd}\right)^{1-o_{td}}, \qquad (1)$$

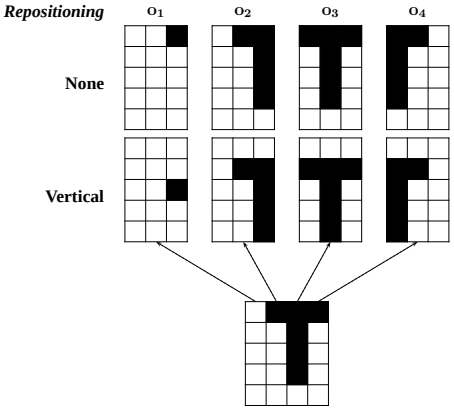

Figure 1. Example of transformation of a 4×5 binary image into a sequence of four 15-dimensional binary feature vectors $O = (\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4)$ using a window of width 3. No repositioning (top) is compared with the vertical repositioning (bottom).

where $\mathbf{o}_t \in \{0,1\}^D$ is the observation at t, and τ_{jk} and $\mathbf{p}_{jk}$ are, respectively, the prior and prototype of the k-th mixture component in state j. As conventional Gaussian HMMs, BHMMs can be trained using the MLE criterion by means of the Baum-Welch algorithm [4]. However, γ-MMI is reported to obtain better results in the literature [13], [15]. We will refer to BHMMs trained using MMI as Discriminative BHMMs. Given a collection of samples $\{(O_n, S_n)\}_{n=1}^N$, the γ-MMI criterion is defined as follows

$$F_{\gamma\text{-MMI}}(\lambda) = \frac{1}{\gamma} \sum_{n=1}^{N} \log \left(\frac{\left[\exp(\lambda^T f(O_n, S_N)) \right]^\gamma}{\sum_R \left[\exp(\lambda^T f(O_n, R)) \right]^\gamma} \right). \tag{2}$$

The optimization is performed by gradient descend using the RPROP algorithm [14]. In order to avoid overfitting, a L2 regularization term is added to the original criterion $F_{\gamma\text{-MMI}}(\theta)$.

IV. LSTM Tandem HMMs

Artificial neural networks (ANNs) in a tandem HMM approach combine the discriminative parameter estimation of the ANN with the sequence modeling ability of the HMM [8]. Training the ANN requires each observation $\mathbf{o}_t \in \mathbb{R}^D$ at time step t in the training data to be aligned to a character label of its transcription. In order to obtain this labeling a previously trained GHMM applied to the training data in the forced alignment mode. Then the ANN is trained on the labeled observations. Recurrent ANN architectures (RNNs) provide a natural way to deal with contextual information over time [3]. In the presented experiments we use bidirectional Long-Short-Term-Memory (LSTM) RNNs, which lead to significant improvements in handwriting recognition [10]. The LSTM RNN is trained in a frame-based approach with a softmax output layer using Backpropagation through time (BPTT).

The trained LSTM RNN it is used to calculate a posterior distribution over the character labels for each observation. In a tandem HMM approach the posterior estimates are considered as observations to train a new Gaussian HMM (GHMM) in order to perform the sequence modeling. See Figure 2 for an illustration.

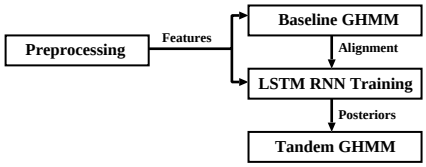

Figure 2. The three steps of the LSTM Tandem HMM approach: An alignment obtained by a baseline HMM is used to train the LSTM RNN. Afterwards the posterior estimates are used as observations to train the Tandem HMM.

V. Experimental Results

Experiments were conducted on corpora with Arabic and French handwriting using BHMMs, GHMMs and the LSTM tandem HMM approach.

The RIMES database [1] consists of $5,605$ fictional French letters by more than $1,300$ writers. Each word is built from 82 symbols containing upper- and lowercase characters, ligatures, typographical symbols, punctuation marks and a white-space model. In our experiments we used the training and validation corpus of the ICDAR 2011 competition for isolated word recognition. A closed vocabulary containing $5,340$ words was used to estimate a unigram language model with a perplexity of 45.2. The validation corpus was used as test corpus in the ICDAR 2009 competition. The training corpus contains $51,738$ words and the validation corpus contains $7,464$ words.

The IFN/ENIT database [11] contains $32,492$ Arabic handwritten Tunisian town names by about $1,000$ writers with a vocabulary size of 937. A whitespace character and position dependent length modeling of the 28 base characters leads to 121 different character labels [2]. The database is divided in five disjoint sets, where in the presented experiments the sets a-d were used for training and set e for testing. This setup results in 335 singletons.

A. LSTM Tandem HMM

The images of the RIMES database were scaled to a fixed height of 40 pixels keeping the aspect ratio. Afterwards the vertical repositioning method was applied and the features were reduced by PCA to 35 components using a sliding window of size 14. The baseline GHMM was composed of ten states with five separate Gaussian mixture models. With the alignment provided by the GHMM the LSTM RNN was trained with two hidden layers containing 100 and 200 nodes respectively resulting in about 785k weights. A separate validation set containing 20% of the training data was used

Table I
COMPARISON OF GHMMS ON THE RIMES DATABASE WITH AND WITHOUT THE VERTICAL REPOSITIONING.

repositioning	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
GHMM	36.6	24.4	23.5	15.5
+ LSTM	25.8	17.2	9.7	5.2

Table II
COMPARISON OF GHMMS ON THE RIMES DATABASE BEFORE AND AFTER BINARIZING THE FEATURES.

binarization	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
GHMM	23.5	15.5	24.7	13.8
+ LSTM	9.7	5.2	10.6	5.6

Table III
COMPARISON OF GHMMS ON THE IFN/ENIT DATABASE WITH AND WITHOUT THE VERTICAL REPOSITIONING.

repositioning	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
GHMM	13.1	10.6	6.7	5.2
+ LSTM	7.2	5.6	4.8	3.7

Table IV
COMPARISON OF GHMMS ON THE IFN/ENIT DATABASE BEFORE AND AFTER BINARIZING THE FEATURES.

binarization	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
GHMM	6.7	5.2	6.4	4.6
+ LSTM	4.8	3.7	5.0	3.9

to detect convergence of the RNN training. The posterior estimates of the LSTM RNN were reduced by PCA to 72 components and used to train a tandem GHMM with the same topology as the baseline GHMM.

Table I compares the results of GHMMs with and without vertical repositioning method on the validation data of the RIMES database. Vertical repositioning improves the GHMM system absolutely by 12.2% in terms of word error rate (WER) and 9.8% in terms of character error rate (CER). With the LSTM tandem GHMM an absolute improvement of 8.3% WER and 7.2% CER can be observed. The relative improvement of the LSTM tandem GHMM compared to the baseline GHMM decreases from 29.5% WER to 25.8% WER. In order to make a clear comparison to BHMMs, additional experiments were conducted using the same features after binarizing them with the Otsu's method. The results of the experiments with and without the additional binarization step are shown in Table II. Both the GHMM and the LSTM tandem GHMM show an increase of the WER and the CER.

On IFN/ENIT a scaling to 30 pixels height was performed keeping the aspect. Then, the vertical repositioning method was applied and the features were reduced by PCA to 35 components using a sliding window of size six. A 12-state baseline GHMM With six separate Gaussian mixture Models was trained on the features and used to generate the alignment for the RNN training. The LSTM RNN again consisted of two hidden layers with 100 and 200 nodes respectively resulting in about 800k weights. Convergence was detected on a separate validation set containing 20% of the training data. A tandem GHMM with the same topology as the baseline GHMM was trained on the 121 posterior estimates of the LSTM RNN, which were reduced by PCA to 64 components.

Table III shows the results of the systems with and without vertical repositioning. The preprocessing method improves the baseline GHMM by 6.4% WER and 5.4% CER absolutely. With the LSTM tandem approach an absolute improvement of 2.6% WER and 1.9% CER can be observed. The relative improvement of the LSTM tandem HMM compared to the baseline GHMM decreases from 45% WER to 28.8% WER. As on the RIMES database, additional experiments were conducted using the same features after binarizing them with the Otsu's method. Table IV compares the results with and without the additional binarization step. Only a small absolute improvement of 0.2% WER and 0.6% CER can be observed in the baseline GHMM. In the LSTM tandem GHMM the WER and CER increase through the binarization step.

B. BHMM

For the BHMM classifier all images were first scaled to a given height H, and then binarized using the Otsu's method. The CoG repositioning is then applied to the binarized images using a sliding window of a given width W. As a result, original images are transformed into sequences of $(H \times W)$-dimensional binary feature vectors.

Regarding to the model topology we used BHMM with a left to right topology without skip transitions and with a fixed number of states per character. MLE parameter estimation was carried out using a typical incremental strategy. That is, for $K = 1$ mixture components per state, BHMMs were initialized by first segmenting the training set with a "neutral" model analogous to that in [16], and then using the resulting segments to perform a Viterbi initialization. For $K > 1$, the BHMMs were initialized by splitting the mixture components of the models trained with $K/2$ mixture components per state. In each case, we performed 4 EM iterations after the initialization.

We tried different values for the sliding window width, $W \in \{1, 3, 5, 7, 9\}$, different heights $H \in \{20, 30, 35, 40\}$, number of states per character $Q \in \{4, 6, 8, 10\}$ and several number of mixture components per state $K \in$

Table V
COMPARISON OF BHMMS ON THE RIMES DATABASE WITH AND WITHOUT THE VERTICAL REPOSITIONING.

repositioning	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
BHMM	26.5	17.0	21.3	12.9
+ MMI	-	-	16.9	9.8

Table VI
COMPARISON OF BHMMS ON THE IFN/ENIT DATABASE WITH AND WITHOUT THE VERTICAL REPOSITIONING.

repositioning	no		yes	
	WER[%]	CER[%]	WER[%]	CER[%]
BHMM	13.7	10.3	6.2	5.2
+ MMI	-	-	6.2	5.2

$\{1, 2, 4, 8, 16, 32, 64\}$. In both corpora the parameter tuning was carried out over a special train-validation partition. In order to tune the number of states Q, window width W, height H and number of mixture components per state K, we carried out experiment over a special train-validation sets. On the IFN/ENIT database we performed a cross-validation over the sets a,b,c and d. In RIMES the train set was randomly split into train ($\approx 80\%$) and validation ($\approx 20\%$). In IFN/ENIT the best results were obtained using $H = 30$, $W = 9$, $Q = 6$ and $K = 32$, while in the case of RIMES the best configuration was $H = 40$, $W = 9$, $Q = 8$ and $K = 64$.

With the previous parameters we carried out experiments with and without vertical repositioning on the standard partitions of both corpus. The results for IFN/ENIT and RIMES are shown respectively in the top row in Table VI and Table V. As expected, repositioning clearly outperforms the use of a sliding window without repositioning. We are obtaining an absolute improvement of 7% WER on IFN/ENIT and an absolute improvement of 5% WER on RIMES.

A last experiment was carried out in order to try to improve the previous results with repositioning by applying the γ-MMI criterion. We initialized the training process by transforming the best MLE models from previous experiments into equivalent Log-Linear HMMs (LLHMMs) for binary data. Then we used RPROP for optimizing the training criterion. And finally, the resulting LLHMMs were transformed again into equivalent BHMMs classifiers. Despite the best generative results are obtained with $K = 64$ and $K = 32$, some works reported [5] that the best classifier obtained using MMI training requires less mixture components than its generative counterpart. For this reason, and for the required computational cost by the discriminative training, we reduce the number of mixture components to $K = 26$ and checked that similar results were obtained to those obtained increasing the value of K. A comparison of the conventional BHMMs with discriminatively trained BHMMs is shown in the second column in Table VI and Table V. For the IFN/ENIT database no improvement was obtained using discriminative training. In fact, without regularization we quickly observed overfitting over the validation set. However, on the RIMES database we obtained an absolute improvement of 4% WER absolutely.

VI. CONCLUSIONS

We examined a method to overcome the limitations of HMMs to deal with vertical image distortion and evaluated it for different HMM systems on databases with Arabic and French handwriting. In order to remove the vertical distortion the CoG is calculated for a window of the image data. Afterwards the window is repositioned to be vertically aligned to its CoG. For BHMMs a final binarization step is required to make the data suitable for the Bernoulli mixture model used as emission probability function.

Our experiments show that vertical repositioning is able to augment the information given to an HMM, which can not be discovered by the HMM itself due to its inability to deal with vertical distortions. The same is true for the LSTM RNN because they are also trained on a one-dimensional sequence of fixed size pixel columns, such that the pixels of each row are always associated with the same unit in the input layer. Multidimensional RNNs exist [6], [7], but without further heuristics they enlarge the number of time steps in a magnitude that offline training with BPTT becomes infeasible for large network architectures.

The relative improvement of the LSTM tandem GHMM compared to the baseline GHMM remains roughly the same on the RIMES database, while it decreases by more than 16% WER on the IFN/ENIT database which in general shows a better recognition performance. However, the final binarization step required for BHMMs leads to no improvement in GHMM models as shown in the experiments. The binarization step discards valuable information for GHMM and LSTM RNN. Finally, BHMMs show a superior performance compared to the GHMM approach. In combination with discriminative training their performance on the RIMES database could be improved by 4.4% WER absolutely.

ACKNOWLEDGMENT

Work supported by the EC (FEDER/FSE/FP7) (Translectures project 287755), and the Spanish MICINN (MIPRCV "Consolider Ingenio 2010", iTrans2 TIN2009-14511, MITTRAL TIN2009-14633-C03-01 and erudito.com TSI-020110-2009-439).

REFERENCES

[1] E. Augustin, M. Carré, G. E., J. M. Brodin, E. Geoffrois, and F. Preteux. Rimes evaluation campaign for handwritten mail processing. In *Proceedings of the Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 231–235, October 2006.

[2] P. Dreuw, S. Jonas, and H. Ney. White-space models for offline Arabic handwriting recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 1–4, Tampa, Florida, USA, December 2008.

[3] J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[4] A. Giménez, I. Alkhoury, and A. Juan. Windowed Bernoulli Mixture HMMs for Arabic Handwritten Word Recognition. In *ICFHR' 10*, pages 533–538, Kolkata (India), 2010.

[5] A. Giménez, J. Andrés-Ferrer, A. Juan, and N. Serrano. Discriminative Bernoulli Mixture Models for Handwritten Digit Recognition. In *ICDAR' 11*, pages 558–562, Beijing (China), 2011.

[6] A. Graves, S. Fernndez, and J. Schmidhuber. Multidimensional recurrent neural networks. In *in Proceedings of the 2007 International Conference on Artificial Neural Networks*, 2007.

[7] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Neural Information Processing Systems*, pages 545–552, 2008.

[8] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature stream extraction for conventional HMM systems. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1635–1638, 2000.

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

[10] M. Liwicki, A. Graves, H. Bunke, and J. Schmidhuber. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007.

[11] V. Märgner and H. El Abed. ICFHR 2010 - Arabic handwriting recognition competition. *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 0:709–714, 2010.

[12] V. Märgner. and H. El Abed. ICDAR 2011 - arabic handwriting recognition competition. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1444 –1448, September 2011.

[13] D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, Cambridge University Engineering Dept, March 2003.

[14] M. Riedmiller and H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

[15] R. Schluter and W. Macherey. Comparison of discriminative training criteria. In *ICASSP' 98*, volume 1, pages 493–496 vol.1, May 1998.

[16] S. Young et al. *The HTK Book*. Cambridge University Engineering Department, 1995.