Comparing Natural and Synthetic Training Data for Off-line Cursive Handwriting Recognition

Tamás Varga and Horst Bunke
Institut für Informatik und angewandte Mathematik, Universität Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
varga,bunke@iam.unibe.ch

Abstract

In this paper, a perturbation model for the generation of synthetic textlines from existing cursively handwritten lines of text, produced by human writers, is presented. The goal of synthetic textline generation is to improve the performance of an off-line cursive handwriting recognition system by providing it with additional, synthetic training data. In earlier papers, it has been shown that it is possible to improve the recognition performance by using such synthetically expanded training sets. In this paper, we investigate the suitability of synthetically generated handwriting when enlarging the training set of a handwriting recognition system in a more rigorous way. In particular, the improvements achieved with synthetic training data are compared to those achieved by expanding the training set using natural, i.e. human written, textlines.

Keywords: off-line cursive handwriting recognition, training set expansion, synthetic training data, perturbation model, hidden Markov model (HMM).

1. Introduction

Handwriting recognition systems need to be trained. It is well known that the performance of a handwriting recognition system is strongly affected by the size and quality of the training data [1]. As a rule of thumb says, the classifier that is trained on the most data wins. This was empirically confirmed in a number of experiments [2, 8, 10, 15].

However, usually it is not possible to arbitrarily enlarge the training set using natural, i.e. human written, texts. One promising way to overcome this dilemma is to use synthetic data. The synthetic generation of new training samples can be achieved in many different ways, e.g. through perturbation of, or interpolation between, the original samples. Several methods for synthetic handwriting generation have already been reported in the literature [2, 3, 4, 5, 8, 11]. Nevertheless, those works where synthetic training data was successfully used to train recognizers are mainly related to the field of isolated character recognition [2, 3, 5, 8, 12].

For the problem of general, off-line cursive handwritten textline recognition, a perturbation model to generate synthetic textlines from existing cursive handwritten text was proposed in [13]. The basic idea of the approach is to use continuous nonlinear functions that control a class of geometrical transformations applied on an existing handwritten textline. Besides geometrical distortions, thinning and thickening operations are also used. It was shown that by expanding the training set using such additional, synthetically generated textlines, it is possible to improve the recognition performance, even when the original training set is large and the textlines are provided by many different writers [14].

In the present paper, we quantitatively investigate the suitability of synthetically generated handwriting when enlarging the training set of a handwriting recognition system. In particular, the improvements achieved by expanding the training set using synthetic textlines are compared to those improvements where the training set is expanded using only natural, i.e. human written, textlines. For character and digit recognition tasks, similar comparisons were carried out in [2] and [8]. But for the domain of unconstrained handwriting recognition, no similar studies have been published before to the knowledge of the authors.

The paper is organized as follows. Section 2 briefly describes the perturbation model. In Section 3, an overview of the handwriting recognition system used for the experiments is given. Experimental results are presented in Section 4. Finally, in Section 5 conclusions are drawn and suggestions for future work are provided.

2. Perturbation model

Variation in human handwriting is due to two major sources. The first is letter shape variation, and the second

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

comes from the large variety of writing instruments. In this section, a perturbation model for the distortion of cursive handwritten textlines is presented, where these sources of variation are modeled by geometrical transformations, including thinning and thickening operations.

The perturbation model incorporates some parameters with a range of possible values, from which a random value is picked each time before distorting a textline. To distort a textline, a series of geometrical transformations are applied, followed by possible thinning or thickening operations. There is a constraint on the textlines to be distorted: they have to be skew and slant corrected, because of the nature of the applied geometrical transformations. This constraint is not severe, because skew and slant correction are very common preprocessing steps found in almost any handwriting recognition system.

Each geometrical transformation is controlled by a continuous nonlinear function, which determines the strength of the considered transformation at each horizontal or vertical coordinate position of the textline. These functions are called underlying functions, and their creation is based on the $cosine$ function, involving some random parameters.

There are four classes of geometrical transformations. Their purpose is to change properties, such as slant, horizontal and vertical scaling, and the position of characters with respect to the baseline. In the following the geometrical transformations will be defined and illustrated by figures. Note that the figures are only for illustration purposes, and weaker instances of the distortions are actually used in the experiments described in Section 4.

Shearing: The underlying function of this transformation defines the tangent of the shearing angle for each x coordinate. Shearing is performed with respect to the lower baseline. An example is shown in Fig. 1. In this example and the following ones, the original textline is shown at the bottom, the underlying function in the middle, and the result of the distortion on top.

Horizontal scaling: Here the underlying function defines the horizontal scaling factor for each x coordinate. This transformation is performed through horizontal shifting of the pixel columns. An example of this operation is shown in Fig. 2.

Vertical scaling: The underlying function determines the vertical scaling factor for each x coordinate. Scaling is performed with respect to the lower baseline. An example is shown in Fig. 3.

Baseline bending: This operation shifts the pixel columns in vertical direction according to the value of the underlying function for each x coordinate. An example is shown in Fig. 4.

The perturbation model also includes transformations, similar to the ones described above, on the level of connected components, rather than on the level of complete

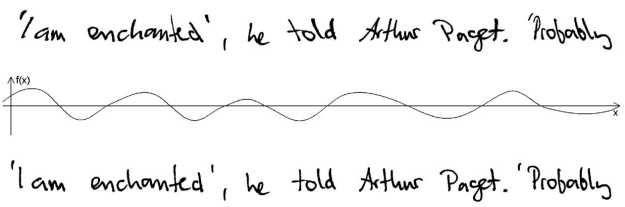

Figure 1. Illustration of shearing (original textline at the bottom, underlying function in the middle, and result of distortion on top).

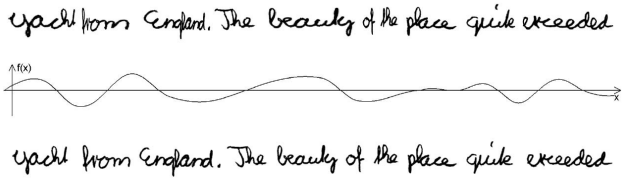

Figure 2. Illustration of horizontal scaling.

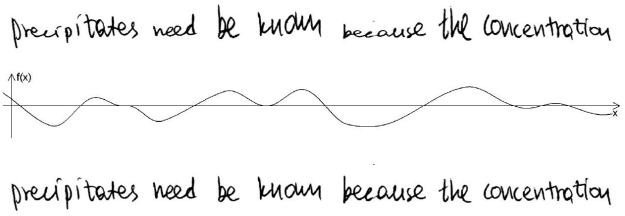

Figure 3. Illustration of vertical scaling.

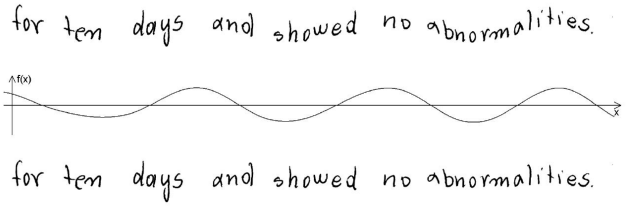

Figure 4. Illustration of baseline bending.

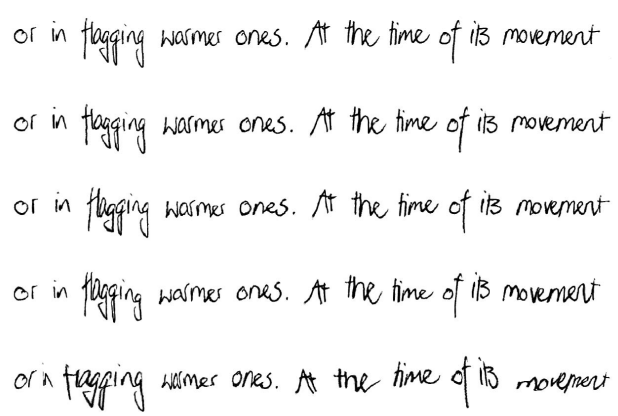

Figure 5. Illustration of distortion strengths. Original textline on top, then distortion strength gradually increases from top to bottom.

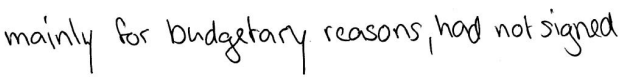

Figure 6. Example of an input textline, before normalization, to be recognized by the system.

lines of text. These transformations change the structure of the writing in a local context, i.e. within connected components. After the application of these transformations, the resulting connected components are scaled in both horizontal and vertical direction, so that their bounding boxes regain their original size. For any further details the reader is referred to [13].

The distortion strength can be controlled by changing the intervals of the possible amplitude values of the underlying functions. An illustration of the distortion strengths is shown in Fig. 5. Here the line on top is the original one, rendered by a human writer. This textline was distorted using increasing distortion strengths. Note that due to the random nature of the perturbation method, virtually all generated textlines, at any strength, will be different. This enables us to generate multiple distorted instances of the same natural textline, at any given strength.

3. Handwriting recognition system

The application considered in this paper is the off-line recognition of cursively handwritten textlines. The recognizer used in this paper is similar to the Hidden Markov Model (HMM) based cursive handwritten textline recognizer described in [6]. The recognizer takes, as a basic input unit, a complete line of text, which is first normalized with respect to skew, slant, baseline location and writing width and produces, as the output, an ASCII transcription of the input textline. An example of an input textline is shown in Fig. 6.

For feature extraction, a sliding window of one pixel width is moved from left to right over the input textline, and nine geometrical features are extracted at each window position (see [6] for further details). Thus an input textline is converted into a sequence of feature vectors in a 9-dimensional feature space. For each character, an HMM is built. In all HMMs the linear topology is used, i.e. there are only two transitions per state, one to itself and one to the next state. In the emitting states, the observation probability distributions are estimated by mixtures of Gaussian components. In other words, continuous HMMs are used. The number of Gaussian mixtures, Ga, is the same in all HMMs. Concrete values of parameter Ga will be provided in Section 4. The character models are concatenated to represent words and sequences of words.

For training, the Baum-Welch algorithm [9] is applied. In the recognition phase, the Viterbi algorithm [9] is used to find the most probable word sequence. As a consequence, the difficult task of explicitly segmenting a line of text into isolated words is avoided, and the segmentation is obtained as a byproduct of the Viterbi decoding applied in the recognition phase. The output of the recognizer is a sequence of words.

4. Experiments

The purpose of the experiments was to compare the improvements achieved by expanding the training set using synthetic textlines to those improvements where the training set was expanded using natural textlines only. For the experiments, subsets of the IAM-Database were used [7]. This database includes over $1,500$ scanned forms of handwritten text from more than 600 different writers. In the database, the individual textlines of the scanned forms are extracted already, allowing us to perform off-line handwritten textline recognition experiments directly without any further segmentation steps.

To examine the system performance as a function of the training set size, four different training set sizes were considered: 160, 320, 638 and $1,275$ textlines, where each smaller set was a subset of all larger sets. These four sets consisted of natural textlines only. This means that at each training set expansion, i.e. when going from one set to the next larger one, the size was approximately doubled. Furthermore, at each expansion, the additional textlines came from writers who had not yet been represented in the training set. The numbers of writers in the four training sets

were 32, 64, 128 and 256, respectively. (So the number of writers were also doubled at each training set expansion.)

To evaluate the system performance as a function of the training set size, a test set of 398 textlines from 80 writers was used. All the experiments were writer independent, i.e. the population of writers who contributed to the training sets were disjoint from those who produced the test set. The underlying lexicon contained $6,012$ words.

For distorted textline generation, the distortion strength that turned out to be optimal in preliminary experiments was used (the test set of this paper was not involved in those experiments). In Fig. 5, this strength corresponds to the second textline from bottom. To expand a training set by synthetically generated textlines, five distorted textlines per given natural training textline were generated and added to the training set. So the synthetically expanded training set was always six times larger than the original one.

In [14], it was pointed out that the capacity, i.e. the number of free parameters, of the recognizer is very important and needs to be tuned appropriately when dealing with synthetically expanded training sets. Particularly, the optimal capacity of the recognizer is expected to be higher when the training set is expanded. Since we wanted to compare the highest possible recognition performance achieved with natural data with that obtained with a mixture of natural and synthetic data, optimization in terms of capacity was performed. This means that, for each training set, we selected the number of Gaussian mixture components, Ga, for which the recognition rate was maximal (see also Section 3).

In the experiments described in the following, the recognition rate will always be measured on the word level. In Table 1, the recognition results on the test set for the four different training set sizes as well as their synthetically expanded counterparts can be seen. In each row, results for a specific training set size are shown. For example, in row *Size=160* it can be seen that the optimal recognition rate using the training set of 160 textlines was 62.86%, at $Ga = 6$. Furthermore, when this training set of 160 natural textlines was expanded by synthetically generated textlines (for each natural textline, 5 perturbed versions were generated and added to the training set, resulting in a total training set size of 960 textlines), the optimal recognition rate was 70.58%, at $Ga = 24$.

This recognition rate of 70.58% is comparable to 70.44% we could achieve using the training set of 638 natural textlines (see row *Size=638*). So the synthetic expansion of 160 training textlines had a similar effect as if the number of natural textlines in the training set had been increased by a factor of four. The improvement from 62.86% to 70.58%, achieved by adding synthetic textlines to the training set is quite substantial. For training set sizes of 320 and 638 natural textlines, synthetic expansions also yielded substantial improvements, from 68.59% to 73.05% and from 70.44% to 74.66%, respectively. We observe that these improvements are higher than those achieved by doubling the size of the training set using additional natural textlines. For size $1,275$, synthetic expansion also improved the recognition rate, but there was no larger natural training set in our experiments to which this improvement could be compared.

Table 1. Recognition rates (in %) on the test set for natural and synthetically expanded training sets of different sizes.

	Natural text only		Synth. expanded set	
	Rec. rate	Opt. Ga	Rec. rate	Opt. Ga
Size=160	62.86	6	70.58	24
Size=320	68.59	9	73.05	18
Size=638	70.44	9	74.66	24
Size=1275	73.96	18	75.98	27

The results of the experiments reported in Table 1 clearly show that synthetically augmented training sets can lead to improved performance of a handwriting recognition system. Since the optimal number of Gaussians was always greater for the synthetically expanded training set than for its natural counterpart, the results also confirmed that increasing capacity has a beneficial effect when augmenting the training set by synthetic data. Furthermore, the recognition rates in Table 1 suggest that in terms of recognition performance, the acquisition of a remarkable amount of new natural textlines can be substituted by generating synthetic textlines from the available ones. We also note that according to the results in [14], similar phenomena can be expected when using other distortion strengths, too. Of course, generation of synthetic textlines requires less effort and time than collection of natural handwriting samples.

5. Conclusions and future work

A method for training set expansion by generating randomly perturbed versions of natural textlines rendered by human writers was presented. It was demonstrated that adding synthetically generated textlines to the training set improves the recognition performance of our off-line cursive handwritten textline recognition system. The aim of the experiments was to compare the improvements achieved by expanding the training set by synthetic textlines to those improvements where the training set was expanded using natural, i.e. human written, textlines only. For this purpose, four different training set sizes were used. The results suggest that, in terms of recognition performance, a remarkable amount of additional human written training textlines can be substituted by synthetic textlines generated from the available training set of natural textlines.

In the future, further investigations on this issue will be conducted involving larger training sets of natural textlines. Optimizations will also be carried out to generate synthetically expanded training data of better quality. These include the optimization of the number of synthetic textlines to generate, as well as the optimization of the distortion strength for different training set sizes. Another idea is not to add all the generated texts to the training set, but perform a kind of pre-selection, i.e. exclude (reject) badly distorted ones. Applying style dependent distortions may also be useful. The current paper makes use of hidden Markov models for handwritten textline recognition. However, similar effects can be expected when dealing with other types of recognizers, for example, neural nets.

6. Acknowledgements

This work was supported by the Swiss National Science Foundation program Interactive Multimodal Information Management (IM)2 in the Individual Project Scene Analysis, as part of NCCR. The authors also thank Urs-Viktor Marti and Matthias Zimmermann for providing the recognition system and the IAM-Database.

References

[1] H. Baird. The State of the Art of Document Image Degradation Modeling. In *Proc. 4th IAPR Workshop on Document Analysis Systems (DAS 2000)*, Rio de Janeiro, Brasil, December 2000.

[2] J. Cano, J. Pérez-Cortes, J. Arlandis, and R. Llobet. Training Set Expansion in Handwritten Character Recognition. In *Proc. 9th SSPR / 4th SPR*, pages 548–556, Windsor, Ontario, Canada, 2002.

[3] H. Drucker, R. Schapire, and P. Simard. Improving Performance in Neural Networks Using a Boosting Algorithm. In S. Hanson et. al., editor, *Advances in Neural Information Processing Systems 5*, pages 42–49. Morgan Kaufmann, San Mateo CA, 1993.

[4] I. Guyon. Handwriting synthesis from handwritten glyphs. In *Proc. 5th Int. Workshop on Frontiers in Handwriting Recognition*, pages 309–312, Essex, England, 1996.

[5] J. Mao and K. Mohiuddin. Improving OCR performance using character degradation models and boosting algorithms. *Pattern Recognition Letters*, 18:1415–1419, 1997.

[6] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artifical Intelligence*, 15(1):65–90, 2001.

[7] U.-V. Marti and H. Bunke. The IAM-Database: an English Sentence Database for Off-line Handwriting Recognition. *Int. Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[8] M. Mori, A. Suzuki, A. Shio, and S. Ohtsuka. Generating new samples from handwritten numerals based on point correspondence. In *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, pages 281–290, Amsterdam, The Netherlands, 2000.

[9] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[10] H. Rowley, M. Goyal, and J. Bennett. The effect of large training set sizes on online Japanese Kanji and English cursive recognizers. In *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 36–40, Niagara-on-the-Lake, Ontario, Canada, 2002.

[11] S. Setlur and V. Govindaraju. Generating manifold samples from a handwritten word. *Pattern Recognition Letters*, 15:901–905, 1994.

[12] P. Simard, Y. Cun, and J. Denker. Efficient pattern recognition using a new transformation distance. In S. Hanson et. al., editor, *Advances in Neural Information Processing Systems 5*, pages 50–58. Morgan Kaufmann, San Mateo CA, 1993.

[13] T. Varga and H. Bunke. Generation of Synthetic Training Data for an HMM-based Handwriting Recognition System. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 618–622, Edinburgh, Scotland, August 2003.

[14] T. Varga and H. Bunke. Off-line Handwritten Textline Recognition Using a Mixture of Natural and Synthetic Training Data. In *Proc. 17th International Conference on Pattern Recognition*, Cambridge, United Kingdom, August 2004, to appear.

[15] O. Velek and M. Nakagawa. The Impact of Large Training Sets on the Recognition Rate of Off-line Japanese Kanji Character Classifiers. In *Proc. 5th Int. Workshop on Document Analysis Systems (DAS 2002)*, pages 106–109, August 2002.