Proceedings of the 12th International
Conference on Pattern Recognition,
Jerusalem (Israel), October 1994,
pages 596-598, IEEE Computer Society Press

Combining Bitmaps with Dynamic Writing Information for On-Line Handwriting Recognition

Stefan Manke, Michael Finke, and Alex Waibel

University of Karlsruhe
Computer Science Department
D-76128 Karlsruhe, Germany

Carnegie Mellon University
School of Computer Science
Pittsburgh, PA 15213-3890, USA

Abstract

Writer independent, large vocabulary on-line handwriting recognition systems require robust input representations, which make optimal use of the dynamic writing information, i.e. the temporal ordering of the sampled data points. In this paper we describe an input representation for cursive handwriting, which combines this dynamic writing information with static bitmaps used in optical character recognition. This input representation is used with a connectionist recognizer, which is well suited for handling temporal sequences of patterns as provided by this kind of input representation. Our system has been tested on different cursive handwriting recognition tasks with vocabulary sizes up to 20000 words. We achieve recognition rates up to 99.5% on writer independent, single character recognition tasks and up to 98.1% on writer dependent, cursive handwriting tasks.

1 Introduction

Several different preprocessing techniques both for optical character recognition (OCR) and on-line character recognition (OLCR) have been developed during the last decades. Robust preprocessing has great influence on subsequent processing (recognition and postprocessing) and the recognition rate. While in OCR the preprocessing is usually based on static bitmaps (scanned text), in OLCR the dynamic writing information, i.e. the temporal ordering of sampled data points, can be recorded and used for recognition. But if coordinates are observed only as a function of time, the spatial context and proximity of the strokes of characters is distorted or lost.

In this paper we propose an input representation for on-line cursive handwriting, which benefits both from the advantages of static bitmaps used in OCR and the dynamic writing information available in OLCR. In this input representation characters and words are represented as a temporal sequence of so called context bitmaps, which are basically low resolution descriptions of the coordinate's neighborhood. By using this sequence of context bitmaps as input representation for our connectionist recognizer [1] both the temporal information and the spatial context is preserved and no important information is lost. We compare this input representation to an representation, which considers for each coordinate only a small temporal context, i.e. for each coordinate a set of local features is calculated, which describe the curvature, writing direction, pressure, speed, and position in this coordinate [3].

The following two sections describe the complete preprocessing consisting of normalization and feature extraction, followed by a short description of the connectionist recognizer, in which this input representation is used (section 4). Recognition results both for different single character recognition tasks and cursive handwriting recognition tasks with vocabulary sizes up to 20000 words are presented in section 5.

2 Normalization

Normalization is performed to remove variability occurring in the raw coordinate sequence. To compensate for different sampling rates, varying writing speeds of different writers, and of a single writer within a single word or character, the coordinate sequence is resampled from temporal equidistance to spatial equidistance. Then the resampled coordinate sequence is smoothed, using a moving average window, which mainly removes sampling noise. Finally a two stage baseline correction is performed, using a linear regression through all data points to get a rough baseline correction and a linear regression through all local minima of the curvature for a final correction [7].

3 Context bitmaps and local features

The second step of our preprocessing is the extraction of features along the pen trajectory yielding a sequence of time-ordered feature vectors, preserving the dynamic writing information. The basic idea of our feature extraction is to refer to low level topological features of the trajectory only and leave the extraction of high level features to the connectionist recognizer.

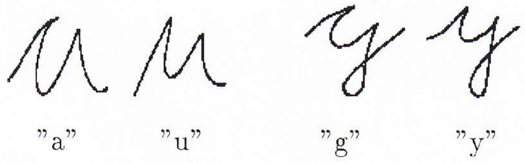

Figure 1: Hard to detect differences between cursive characters

First we started with a set of strictly local features similar to those in [3, 2]. Each time frame consisted of information on the pen position (x, y coordinates), directional features $(\delta x, \delta y)$, curvature, speed and pen-up/pen-down indicator. But an inspection of the confusion matrices of networks trained on these features revealed significant problems in discriminating between cursive letters like ”a” and ”u” or ”g” and ”y”, which look very similar and differ only in small regions of the characters (see figure 1 for examples). These problems arise due to the fact that the features are strictly local, which means that they are local both in space and time. Therefore they are inadequate for modeling temporal long range context dependencies occurring in the pen trajectory.

The basis for the new set of features we use now is a bitmap representation of the digitizer input. After normalizing the input we map the sequence of points (x_t, y_t) to a grey scale bitmap $B = \{b(i,j)\}$, where $b(i,j)$ indicates the number of points (x_t, y_t) falling into pixel (i,j).

Unlike the setting of optical character recognition, where bitmaps are the only source of information, we also have the temporal sequence of the points. The idea is to combine these two sources in the following way: Assume (x_t, y_t) falls into bitmap pixel (i,j). Consider a local $d \times d$ section of bitmap B centered around (i,j) (figure 2b) and derive a 3×3 grey scale bitmap L_t by averaging this section (figure 2c). That means, we derive a temporal sequence of low resolution bitmaps L_t centered around (x_t, y_t) (figure 2a). These bitmaps plus directional information $(\delta x, \delta y)$ and the pen-up/pen-down feature form the new set of input features we use for recognition.

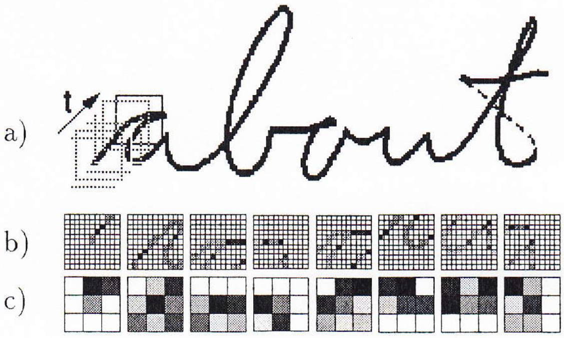

Figure 2: Calculation of context bitmaps

These features are still local in space but no longer local in time. Each point of the trajectory is visible from each other point of the trajectory in a small neighborhood. Therefore, we call the local bitmaps L_t context bitmaps. Another way of interpreting these context bitmaps is to view at them as low resolution long term memory. They seem to be appropriate for modeling temporal long range and spatial short range phenomena as observed in pen trajectories. Compared to the set of strictly local features we achieved a 50% error reduction by using the new feature set.

4 Using context bitmaps in a connectionist recognizer

The input representation described in this paper is used in a connectionist recognizer [1], which is well suited for handling temporal sequences of patterns as provided by this kind of input representation. This recognizer, a Multi-State Time Delay Neural Network (MS-TDNN) [4], integrates the segmentation and recognition of words in a single framework. The MS-TDNN architecture, which was originally proposed for continuous speech recognition tasks [4, 5], combines shift invariant high accuracy pattern recognition capabilities of a TDNN [6] with a non-linear time alignment procedure (dynamic time warping) for aligning strokes into character sequences.

5 Experiments and results

We have tested the proposed input representation both on single character and cursive (continuous) handwriting recognition tasks using the MS-TDNN architecture. The handwriting databases used for training and testing of the MS-TDNN were collected at the

Table 1: Results for different writer dependent/writer independent handwriting recognition tasks

Task	Vocabulary Size	Training Patterns	Test Patterns	Recognition Rate Local Features	Recognition Rate Context Bitmaps
0_9	10	1600	200 (20 writers)	97.9%	99.5%
A_Z	26	2000	520 (20 writers)	92.5%	95.9%
a_z	26	2000	520 (20 writers)	89.9%	93.7%
*msm*_400_a	400	2000 (writer *msm*)	800 (writer *msm*)	94.7%	98.1%
*msm*_400_b	400	– ” –	– ” –	93.2%	96.7%
*msm*_1000	1000	– ” –	2000 (writer *msm*)	90.5%	94.8%
*msm*_10000	10000	– ” –	– ” –	82.1%	86.6%
*msm*_20000	20000	– ” –	– ” –	79.9%	83.0%
*multi*_400	400	3000 (15 writers)	2500 (10 writers)	-	85.0%

University of Karlsruhe. All subjects had to write a set of single words from a given vocabulary, covering all lower case letters, and at least one set of isolated lower case letters, upper case letters, and digits. The data is preprocessed as described in sections 3 and 2. Table 1 summarizes results for different writer independent/writer dependent, single character recognition/cursive handwriting recognition tasks, comparing the local feature representation with the new context bitmap representation.

The network used for the writer dependent results in table 1 is trained with approx. 2000 training patterns from a 400 word vocabulary (*msm*_400_a) and tested without any retraining on different vocabularies with sizes from 400 up to 20000 words. Vocabularies *msm*_400_b, *msm*_1000, *msm*_10000, and *msm*_20000 are completely different from the vocabulary the network was trained on and were selected randomly from the Wall Street Journal vocabulary.

6 Conclusions and future work

In this paper we have proposed an input representation, which combines dynamic writing information with context bitmaps for achieving high recognition performance using a connectionist recognizer. This input representation has been shown to be appropriate for modeling temporal long range and spatial short range phenomena typically observed in pen trajectories. It is superior to other representations, which consider only coordinates as a function of time (see table 1). Using the input representation in conjunction with the MS-TDNN architecture, we can achieve high recognition performances both on single character and cursive handwriting tasks.

Work is in progress to improve the baseline correction and height normalization and to add a slant correction [7]. Currently we are investigating, if the 3×3 context bitmaps should be replaced by 5×5 context bitmaps and reducing the number of input features by using a principal component analysis. Work is also in progress to apply our recognition system to larger writer independent handwriting tasks with vocabulary sizes up to 25000 words.

References

[1] S. Manke and U. Bodenhausen, "A Connectionist Recognizer for Cursive Handwriting Recognition", *Proceedings of the ICASSP-94*, Adelaide, April 1994.

[2] M. Schenkel, I. Guyon, and D. Henderson, "On-Line Cursive Script Recognition Using Time Delay Neural Networks and Hidden Markow Models", *Proceedings of the ICASSP-94*, Adelaide, April 1994.

[3] I. Guyon, P. Albrecht, Y. Le Cun, W. Denker, and W. Hubbard, "Design of a Neural Network Character Recognizer for a Touch Terminal", *Pattern Recognition*, 24(2), 1991.

[4] P. Haffner and A. Waibel, "Multi-State Time Delay Neural Networks for Continuous Speech Recognition", *Advances in Neural Network Information Processing Systems (NIPS-4)*, Morgan Kaufman, 1992.

[5] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving Connected Letter Recognition by Lipreading", *Proceedings of the ICASSP-93*, Minneapolis, April 1993.

[6] A. Waibel, T. Hanazawa, G. Hinton, K. Shiano, and K. Lang, "Phoneme Recognition using Time-Delay Neural Networks", *IEEE Transactions on Acoustics, Speech and Signal Processing*, March 1989.

[7] W. Guerfali and R. Plamondon, "Normalizing and Restoring On-Line Handwriting", *Pattern Recognition*, 16(5), 1993.