A full English sentence database for off-line handwriting recognition

U.-V. Marti and H. Bunke
Institute of Informatics and Applied Mathematics
University of Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
E-Mail: {marti,bunke}@iam.unibe.ch

Abstract

In this paper we present a new database for off-line handwriting recognition, together with a few preprocessing and text segmentation procedures. The database is based on the Lancaster-Oslo/Bergen(LOB) corpus. This corpus is a collection of texts that were used to generate forms, which subsequently were filled out by persons with their handwriting. Up to now (December 1998) the database includes 556 forms produced by approximately 250 different writers. The database consists of full English sentences. It can serve as a basis for a variety of handwriting recognition tasks. The main focus, however, is on recognition techniques that use linguistic knowledge beyond the lexicon level. This knowledge can be automatically derived from the corpus or it can be supplied from external sources.

Keywords: handwriting recognition, database, unconstrained English sentences, corpus, linguistic knowledge

1 Introduction

Standard databases have become very important in handwriting recognition research[2]. They are an essential requirement for the development, evaluation and comparison of different character recognition techniques. Examples of widely used databases in the field of handwriting recognition are CEDAR[4], NIST[16], CENPARMI[15], UNIPEN[3], ETL9(Japan)[13], and PE92(Korea)[9]. However, these databases contain mostly isolated characters or single words. To the knowledge of the authors, no database containing large amounts of general unconstrained handwritten English text exists. (The NIST database contains many instances of the Preamble of the American Constitution, but the vocabulary is too small to do conclusive experiments.)

The exploitation of contextual knowledge is a key to successful OCR. In speech recognition the importance of contextual and linguistic constraints have been recognized long ago [7, 1]. Also in machine printed character recognition, linguistic knowledge has been successfully applied [6, 5]. However, in handwriting recognition many systems have been developed for isolated characters, where no contextual knowledge is available at all [14, 15]. In other applications, such as bank check or postal address reading, contextual constraints can be applied, but only to a limited degree. Only few works in handwriting recognition address applications where broad linguistic knowledge is applicable. An example is the work by Kim, Govindaraju and Srihari [10] or of Oh, Ha and Kim [12] on unconstrained English sentence recognition. Our own research is directed towards similar goals [11]. The problem under study is the application of linguistic knowledge beyond the word level in the context of reading handwritten unconstrained English sentences.

In this paper we describe the first version of a database that contains full English sentences. As present the database consists of 43751 instances of handwritten words distributed over 4881 lines of text produced by approximately 250 writers. The underlying lexicon includes 6625 different words. The database can be used to train and test word recognizers with a particular focus on techniques that apply linguistic knowledge beyond the lexical level. The database described in this paper will be made available to others upon request.

In the next section, we describe the database acquisition procedure. Section 3 is concerned with the task of defining ground truth for the database. In Section 4 further characteristics of the data collection are listed. Finally some conclusions are presented in Section 5.

2 Corpus and Forms

In the domain of linguistics, large collections of texts, called corpora, exist. These corpora have different appearance and contents. In some only plain text is included. In more elaborated versions the words are tagged. That means for every word in the text there is a tag, which marks the word as a noun, a verb, or another word class. Often also the tempus, modus a.s.o. of a word is included.

A	Press: reportage
B	Press: editorial
C	Press: reviews
D	Religion
E	Skills, trades and hobbies
F	Popular lore
G	Belles letters, biography, essays
H	Miscellaneous
J	Learned and scientific writings
K	General fiction
L	Mystery and detective fiction
M	Science fiction
N	Adventure and western fiction

Table 1. Some text categories in the corpus

For our database we have decided to use the Lancaster - Oslo/Bergen corpus (LOB)[8], a collection of 500 English texts, each consisting of about 2000 words. The texts in this corpus are of quite diverse nature. The different categories of texts in the corpus are listed in Tab. 1. Using a corpus as the foundation of the database rather than collecting text from "random" sources has the advantage that linguistic knowledge can be automatically extracted in a more systematic and easy fashion. It was our goal to acquire a database of handwritten sentences contained in the corpus. For this purpose, we split the texts in the corpus into fragments of about 3 to 6 sentences with at least 50 words each. These text fragments were copied onto forms and we asked different persons to write the text on one or more of the forms by hand.

The forms were automatically generated. We extracted the sentences of each text fragment from the corpus and generated a LaTeXdocument containing the text and the structure of the form. Thus all forms were processed in the same way. The form consists of four parts (see Fig. 1). The first part comprises the title "Sentence Database" and a number assigned to the text. The first character of this number shows which category the text belongs to, and the following two digits identify the text number. For example, M01 indicates that the text of the form is extracted from text "01" in the text category "Science fiction". The next three digits show with which sentence the text starts. In the second part of the form, the text the individual persons were asked to write is printed. The first part of the form is separated from the printed text by a horizontal line. The third part of the form is a blank zone where the writers have to put in their handwriting. Also this part of the form is separated from the others by a horizontal line. In the last part, the writer can voluntarily enter his or her name.

As the main focus of our research is the application of high level linguistic knowledge in handwriting recognition, we wanted to make image preprocessing as easy as possible.

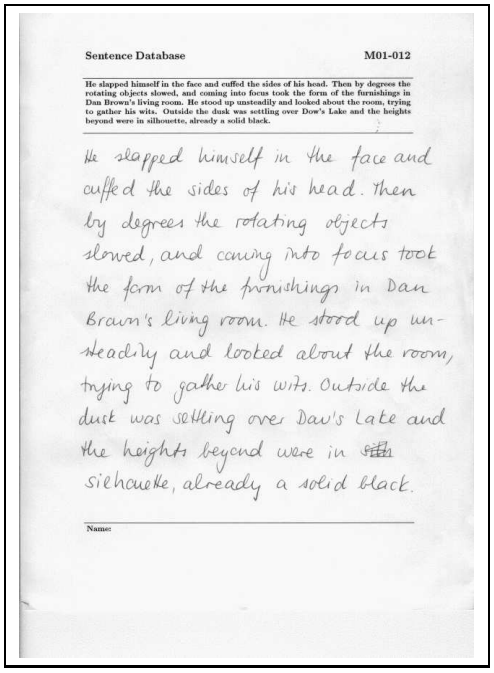

Figure 1. Filled form

Therefore, we decided that the writers had to use rulers. These guiding lines were printed on a separate sheet of paper which was put under the form. The writers were asked to use their every day writing in order to get the most natural and unconstrained way of writing. We also told the writers to stop to write, if there was not enough space left to write the whole text. This way we wanted to avoid to get pressed and deformed words. For the pencil we did not make any constraints. So we got all kinds of writing instruments represented in our database. The filled forms were scanned with a HP-Scanjet 6100 which is connected to a Sun Ultra 1. The software we used to scan the data is *xvscan* version 1.6. It is an add on to the well known image tool *xv*. The resolution was set to 300 dpi at a grey level resolution of 8 bit. The images were saved in TIFF-format with LZW compression. Each form was completely scanned, including the printed and handwritten text. (Thus it is also possible to do experiments with the printed text, for example to distinguish between handwritten and printed text.)

3 Labeling

The labeling of data is a prerequisite for recognition experiments. Because labeling data is expensive, time consuming and error prone, we decided to do as much as possible automatically. The sources of all the forms printed (and subsequently filled by writers) were saved on disk. So it was an easy task to generate the correct labels for the printed text on the forms.

For labeling the handwritten text, a number of image preprocessing and text extraction algorithms are applied. First, a form is segmented into its four main parts (see Section 2). This is a relatively easy task because these four parts are separated by very long horizontal lines, which are easy to detect by horizontal projection. To make the projection algorithm more robust we not only consider the horizontal projection profile, but also search for long horizontal black areas. A horizontal line separating two parts from each other is characterized by a high peak in the horizontal grey value projection histogram and a high peak in the horizontal length histogram of black areas. The separating lines are also used to correct the skew of the document. Once the left and right end of the first horizontal separating line has been found, its skew angle is used to correct the skew of the whole document.

After all three horizontal lines have been found, we are able to extract the part of the form that contains the handwriting. Over the whole database we could extract the handwritten zone without any error using the method described above.

The next step in preprocessing is to cut the text into lines. For this purpose we use a histogram of the horizontal black/white transitions. In this histogram we look for local minima. If the value at a local minimum is zero, a cut has been found that doesn't touch any word. If the value is greater than zero, we have found a position where we can cut the image with a minimal number of intersections with words of the previous or the following text line. To handle these kind of intersections we use a method based on the center of gravity. If the center of gravity of the touching connected component is above the cutting line, we assume that the cut part belongs to the line above. Otherwise it is supposed it belongs to the line below. With this method we could extract almost all text lines correctly. Only in about 0.6% (29 of 4881) of the lines small errors occurred.

After the handwritten text has been segmented into individual lines, the labels of the printed part were copied (because the labels for the handwritten part should be the same as for the printed part) and the line feeds were filled in manually. In some cases corrections were necessary, because the handwritten text did not correspond exactly to the printed text, for example, because the writer left out some words. These corrections were done manually, but they didn't take a long time. Figure 2 gives an overview of the different steps involved in the generation of the database. The result of the labeling is an ASCII file, which contains for each printed and handwritten line of text the labels. In this file no positional information for the lines or words is included. It is clear that the corpus used here can be replaced by any other collection of texts.

In our database there is no splitting of complete handwritten text lines into individual words. The reason is that we aim at applying a segmentation-free recognition method that does not only avoid the segmentation of individual words into characters, but also of text lines into single words [11]. The primary goal of the preprocessing and segmentation procedures is to support the labeling of the text. However, these procedures can be integrated in any recognizer as well.

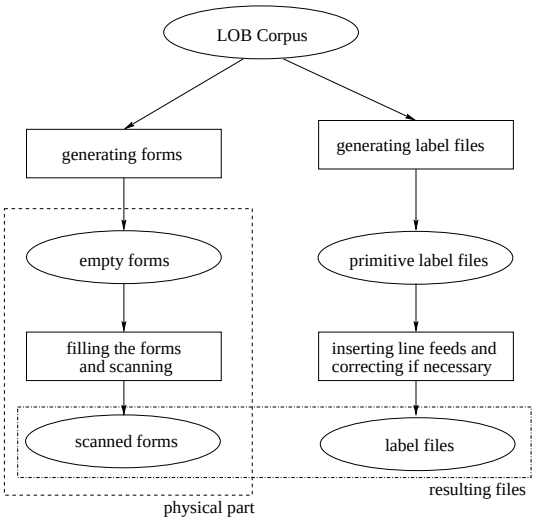

Figure 2. Database acquisition

4 Further Characteristics of the Database

At the moment (December 1998) the database consists of a total of 556 filled forms. The distribution over the different text categories and over the text fragments can be seen in Table 2. There are 4881 lines of handwritten text all together. A total of 43751 word instances out of a vocabulary of 6625 words occur in the database. There are between 8 and 9 words per line of text on the average. There are some subsets of forms in the database that include multiple instances of the same text fragment written by different writers, or the same text fragment written by the same writer; see Table 3. These subsets can serve as a starting point for further research, for example, on writer dependent recognition systems.

For the rest of the database there are between one and three forms from each writer. If it is more than one form, then these forms contain different text fragments, sometimes from different text categories. Moreover, in most of the cases there is exactly one handwritten instance of a text fragment.

5 Conclusion and Further Work

A database consisting of handwritten English sentences has been described in this paper. It is built upon the Lancaster-Oslo/Bergen corpus. The database can serve as a basis

Text category	Text number 1	2	3	total
A	66	16	18	100
B	25	6	8	39
C	3	15	60	78
D	13	0	2	15
E	18	12	0	30
F	12	11	5	28
G	21	4	10	35
H	10	13	0	23
J	6	0	0	6
K	1	13	9	23
L	20	0	3	23
M	23	17	9	49
N	8	16	12	36
P	4	20	23	47
R	0	22	2	24
total	130	165	161	556

Table 2. Distribution of the forms

subset	vocabulary	number of forms	number of lines
x	483	17	155
u	765	29	287
a	411	10	91
b	411	10	89
c	411	10	87
d	411	10	97
e	411	10	97
f	411	9	94

Table 3. Overview of the data sets

for research in handwriting recognition. In particular, it is potentially useful for recognition of general unconstrained English text utilizing knowledge beyond the lexicon level. Linguistic knowledge can be either supplied from external sources, or directly derived from the underlying corpus, which is available in electronic form. A few preprocessing and segmentation procedures have been developed together with the database. Their primary aim was to aid in automatic labeling of the database, but they can be integrated in any recognition system as well.

Currently we continue our effort to further enlarge the database. The version described in this paper (the images of the forms, the labelfiles and the segmentation algorithms) can be made available to other researchers upon request.

References

[1] C. Chelba and F. Jelinek. Refinement of a structured language model. In *Proceedings of the Int. Conf. on Advances in Pattern Recognition*, pages 275–284. Springer Verlag, 1998.

[2] I. Guyon, R. Haralick, J. Hull, and I. Phillips. Database and benchmarking. In H. Bunke and P. Wand, editors, *Handbook of Character Recognition and Document Image Analysis*, chapter 30, pages 779–799. World Scientific, 1997.

[3] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. Unipen project of on-line data exchange and benchmarks. In *Proc. of the 12th IAPR Int. Conf on Pattern Recognition*, pages 29–33, Jerusalem, Israel, Oct. 1994.

[4] J. Hull. A database for handwritten text recognition research. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(5):550–554, May 1994.

[5] J. Hull. Incorporating language syntax in visual text recognition with statistical model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(12):1251–1256, Dec. 1996.

[6] J. Hull and S. Srihari. Experiments in text recognition with binary n-gram and viterbi algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 4(5):520–529, Sept. 1982.

[7] F. Jelinek. Self-organized language modeling for speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers, Inc., 1990.

[8] S. Johansson, G. 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, Oslo, 1978.

[9] D. Kim, Y. Hwang, S. Park, E. Kim, S. Paek, and S. Bang. Handwritten korean character image database PE92. In *Proc. of the Second Int. Conf. on Document Analysis and Recognition*, pages 470–473, 1993.

[10] G. Kim, V. Govindaraju, and S. Srihari. Architecture for handwritten text recognition systems. In *Proceedings of IWFHR'98, Taejon, South Korea*, pages 113–122, 1998.

[11] U.-V. Marti and H. Bunke. Towards general cursive script recognition. In *Proceedings of IWFHR'98, Taejon, South Korea*, pages 379–388, 1998.

[12] S.-C. Oh, J.-Y. Ha, and J.-H. Kim. Context dependent search in interconnected hidden markov model for unconstrained handwriting recognition. *Pattern Recognition*, 28(11):1693–1704, Nov. 1995.

[13] T. Saito, H. Yamada, and K. Yamamoto. On the data base ETL 9 of handprinted characters in JIS chinese characters and its analysis. *IEICE Transactions*, J68-D(4):757–764, 1985.

[14] J. Schürmann. *Pattern Classification*. John Wiley and Sons, Inc., 1996.

[15] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam. Computer recognition of unconstrained handwritten numerals. *Proc. of the IEEE*, 7(80):1162–1180, 1992.

[16] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson. The first census optical character recognition systems conf. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.