Sixth International Conference on Document Analysis and Recognition (ICDAR), Seattle (USA), pages 566-570, 2001

Two On-Line Japanese Character Databases in Unipen Format

Stefan Jaeger, Masaki Nakagawa
Department of Computer Science
Tokyo University of Agriculture and Technology
2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan
email: stefan@hands.ei.tuat.ac.jp, nakagawa@cc.tuat.ac.jp

Abstract

This paper presents the UP_Kuchibue and UP_Nakayosi databases containing on-line handwritten Japanese characters. These databases are the international versions of two databases, Kuchibue and Nakayosi, collected in the Nakagawa Laboratory at the University of Agriculture & Technology in Tokyo. They contain more than 3 million characters written by 283 Japanese writers. UP_Kuchibue and UP_Nakayosi are stored in the common Unipen format. Unipen is a western plain ASCII format, which allows easy access for international researchers and facilitates international benchmarks.

1. Introduction

Eastern and western on-line handwriting recognition have long been lacking in appropriate benchmark data. This fact, which was already criticized more than ten years ago [5], has only gradually improved during the last decade. Compared to collecting off-line data, the collection of on-line data is more expensive and time-consuming. One of the main reasons is that off-line applications typical deal with huge masses of data scanned from paper; e.g., postal automation, while on-line applications require special hardware to collect data. With the advent of pen computers and the gaining interest in pen-based interfaces, however, this situation is now beginning to change. Benchmarks are, of course, essential for handwriting recognition. They allow evaluation of recognizers, measurement of improvements, and provide a common basis of comparison. For instance, the introduction of the Japanese off-line database ETL9 [7] led to a significant boost in performance of Japanese off-line handwriting recognition [6]. Due to the Unipen project [1] and new emerging databases widely-recognized benchmarks have just began to establish themselves in western on-line handwriting recognition [1, 3, 9]. On-line handwriting databases may also benefit off-line recognition since pictorial off-line data can be derived from on-line data [8]. Generally speaking, databases containing any sort of graphical input data are essential to develop high performance systems in pen-based computing.

This paper presents two databases, UP_Kuchibue and UP_Nakayosi, intended for benchmarking Japanese character recognizers. Both databases together contain more than 3 million on-line characters and thus provide an appropriate set for benchmark tests. UP_Kuchibue and UP_Nakayosi are plain ASCII versions of two databases, Kuchibue and Nakayosi, collected in the Nakagawa Laboratory at the University of Agriculture and Technology in Tokyo. The original Kuchibue database has already been adopted as a benchmark by several research groups in Japan since its introduction a few years ago [6]. Today, it is fair to say that the state of the art for the Kuchibue database provides recognition rates between 80% and 90%. To support international cooperation between researchers, Kuchibue and Nakayosi have now been made available in the western Unipen format. Unipen is a plain ASCII format that facilitates international data exchange and recognition experiments. This allows international research groups to apply their existing algorithms to Japanese character recognition without any coding problems. This paper is a brief description of the Unipen format of Kuchibue and Nakayosi, which were originally coded in SJIS. The international Unipen versions of Kuchibue and Nakayosi are named UP_Kuchibue and UP_Nakayosi respectively.

2. Data Collection

Kuchibue and Nakayosi have been collected by asking writers to copy sentences from the 1993 year edition of the Japanese Asahi Shinbun newspaper. Rare or special characters were put together and copied at the end of the newspaper text. Various LCD tablets with different sizes and resolutions have been utilized to capture the data. No restriction was imposed on the style of writing, though boxes (frames) were provided for each character. Writing into boxes is not

	Kuchibue	Nakayosi
#Writers	120	163
#Classes	3356	4438
#Kanjis per Writer	11962	10403
$\sum$ Kanjis	1435440	1695689

Table 1. Data content of Kuchibue and Nakayosi.

as much a restriction as it is for the English language since Japanese people are used to writing into boxes, or imaginary boxes, on their manuscript papers. The main difference between Kuchibue and Nakayosi is the type of captured coordinates: For the Kuchibue database, trajectories were written by pen but captured by tracking mouse events under a MS-Windows environment; i.e., the pen just replaces the mouse as input device. In Nakayosi, coordinates are genuine tablet coordinates provided by the sensors and drivers of the LCD tablets. Though mouse events generally entail a low resolution and a low sampling rate, we think they do not compromise the usefulness of Kuchibue as a testbed for Japanese character recognition.

An elaborate verification process and a thorough label check within Kuchibue and Nakayosi ensure reliable ground truth data for both databases. More information about the data and the setup for data collection is given in [6] for the Kuchibue database and in [4] for the Nakayosi database.

3. Data

Table 1 shows the data content collected for Kuchibue and Nakayosi. Kuchibue contains about 1.4 million characters donated by 120 writers. (Please note that the number of writers has been extended compared to an earlier version presented in [6]). Nakayosi comprises almost 1.7 million characters from 163 writers. The sets of writers for Kuchibue and Nakayosi do not overlap. Each writer of Kuchibue donated 11962 characters covering 3356 Kanji categories. In Nakayosi, each writer donated 10403 characters covering 4438 categories, which include more than 1000 special Kanji characters for naming. Altogether, this sums up to more than 3 million characters donated by 283 writers.

4. Unipen format

The Unipen format is a plain ASCII format designed for on-line handwriting recognition research and development applications. In contrast with binary formats, the Unipen format is not efficient for data storage or for real time data transmission and not designed to handle ink manipulations. However, it does have data annotation capabilities to encode information about recording conditions, writers, segmentation, data layout, data quality, labeling and recognition results [1].

The public-research distribution of collected Unipen data is safeguarded by the International Unipen Foundation. The main goals of the Unipen Foundation according to their statutes are:

- to provide an independent platform for research groups in the area of on-line handwriting recognition and pen-computing technology,
- to collect and distribute large databases containing on-line handwriting data,
- to organize open benchmark rounds, in which performance profiles for existing handwriting recognition algorithms are determined.

More information and a detailed description of the Unipen format can be found on the web pages of the International Unipen Foundation [2].

The Unipen format allows international researchers to easily make experiments with Kuchibue and Nakayosi since the plain ASCII coding requires no major changes to their existing algorithms. Reference [2] contains some links to programs for processing Unipen format files, such as tools for visualization.

5. Organization of UP_Kuchibue and UP_Nakayosi

The Unipen formats of UP_Kuchibue and UP_Nakayosi are almost identical. Both databases contain two files for each writer: data file and segment file. While data files contain the captured data together with a short description, segment files contain information about the corresponding labels and ground truth data.

A data file begins with a header followed by the captured data. The header contains information about the tablet capturing the data, the writer, and the data. Tablets are described by their size, resolution and sampling rate. Writers are characterized by their age, sex, and writing hand; i.e., left-handed or right-handed. The description of the data contains information about the frames provided for each character, which are arranged as a rectangular grid. This includes the upper left coordinate of the first frame, the horizontal and vertical distance between frames, the number of frames for each column and row, and the height and width for each frame. This information may facilitate addressing of single characters in practical experiments. Accord-

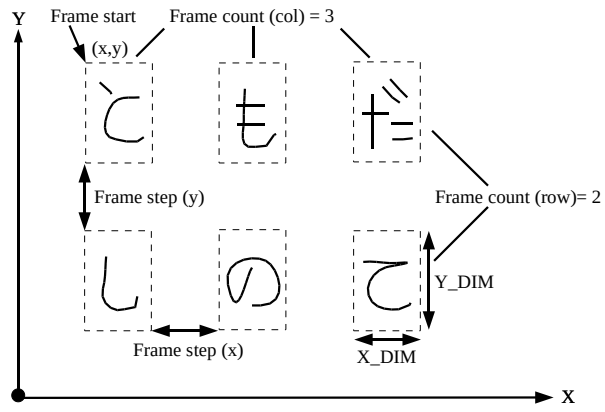

Figure 1. Data collection setup.

ing to the Unipen format, the origin of the coordinate system is in the lower left-hand corner. Figure 1 illustrates these features for a rectangular grid of frames as it is characteristic for data collection. The captured trajectories for each character follow directly after the header in the data file. They are basically sequences of (x,y)-coordinates interrupted by pen-downs and pen-ups, where the beginning of a new character in a new frame is indicated by the keyword ".START_BOX". The position of the pen is not captured when the pen is lifted. Pen-up segments are just represented by one single point indicating the position of the pen-lift. The following listing shows an exemplary excerpt of a data file from the Kuchibue database, illustrating the structure of data files according to the Unipen format:

```
.VERSION 1.0
.INCLUDE GENERAL.HDR
.INCLUDE KUCHIBUE.TXT

.HIERARCHY SCREEN LINE CHARACTER

.INCLUDE MDB0001.SEG

.COMMENT ### Tablet Information ####

.PAD
     Input Resolution: 1280 960
     Display Resolution: 1280 960
     Tablet Size
     (width[mm] x height[mm]): 281 211

.COORD X Y
.POINTS_PER_SECOND 60
.X_POINTS_PER_MM 4.55
.Y_POINTS_PER_MM 4.54

.COMMENT ### Writer Information ###

.HAND R
.SEX M
.AGE 46


.COMMENT ### Data Information ###

.START_SET
.DATA_INFO

     Frame start (x/y): 20 920
     Frame  step (x/y): 65 105
     Frame count (columns/rows): 19 8

.X_DIM 60
.Y_DIM 60

.COMMENT ### First Screen    ###
.COMMENT ### First Character ###
.START_BOX
.PEN_DOWN
40 910
40 911
  .
  .
  .
69 909
69 908
.PEN_UP
69 911
.PEN_DOWN
44 918
44 916
  .
  .
  .
44 884
45 882
.PEN_UP
46 881

.COMMENT ### First Screen     ###
.COMMENT ### Second Character ###
.START_BOX
.PEN_DOWN
107 905
107 903
   ...
```

The first lines of the data file indicate the version of the Unipen format and include some additional headers containing more information, such as the original Japanese text copied. The hierarchy of the data is defined according to the Unipen format and the corresponding segment file is included. In the exemplary data file, the name of the included segment file is MDB0001.SEG, which contains the labels for the first writer.

A segment file generally contains the corresponding ground truth data for its data file. It is organized hierarchically. Entities on the top level are called screens since they represent sets of characters displayed and captured simultaneously on a tablet screen. Every screen is composed of several lines, which in turn contain several characters written into single frames. The segment file begins with the description of the first screen followed by the description of every contained line together with its characters. According to the Unipen format, we describe screens, lines, and characters by their corresponding stroke numbers in the data file, the quality of the data, and their assigned label. The label of a screen is the label of its first line. The label of a line is simply the concatenation of all labels of each character on the line. We have transformed the originally SJIS-coded labels of Kuchibue into the corresponding hexadecimal coding to ensure a plain ASCII format. Since SJIS is a two-byte code, every character is now labeled by four hexadecimal digits in the Unipen format. The following excerpt shows an exemplary segment file corresponding to the data file shown before. It contains the description of the first screen with 8×19 characters followed by the first line of the second screen.

```
.COMMENT ### First Screen =      ###
.COMMENT ### 8 x 19 Frames       ###
.SEGMENT SCREEN 0-839 OK
"SJISx8175977082aa82a082c182bd82e78cfb
934a82f0908182a282c4817682cc96bc82b9
82e882d3...."

.COMMENT ### First Line ###
.SEGMENT LINE 0-103 OK
"SJISx8175977082aa82a082c182bd82e78cfb
934a82f0908182a282c4817682cc96bc82b9
82e882d3"

.COMMENT ### 19 Characters ###
.SEGMENT CHARACTER 0-3 OK
"SJISx8175"
.SEGMENT CHARACTER 4-13 OK
"SJISx9770"
.SEGMENT CHARACTER 14-19 OK
"SJISx82aa"
   .
   .
   .
.SEGMENT CHARACTER 88-93 OK
"SJISx82b9"
.SEGMENT CHARACTER 94-97 OK
"SJISx82e8"
.SEGMENT CHARACTER 98-103 OK
"SJISx82d3"


.COMMENT ### Second Line ###
.SEGMENT LINE 104-259 OK
"SJISx82c682c682e082c9817792458f6f8178
82c58f6f89ef82a281418c8b82ce82ea82bd
96a39866"

.COMMENT ### 19 Characters ###
.SEGMENT CHARACTER 104-105 OK
"SJISx82c6"
.SEGMENT CHARACTER 106-107 OK
"SJISx82c6"
.SEGMENT CHARACTER 108-113 OK
"SJISx82e0"
  .
  .
  .
.SEGMENT CHARACTER 204-211 OK
"SJISx82bd"
.SEGMENT CHARACTER 212-237 OK
"SJISx96a3"
.SEGMENT CHARACTER 238-259 OK
"SJISx9866"

            .
            .
            .

.COMMENT ### Last (8th) Line ###
.SEGMENT LINE 744-839 OK
"SJISx97a782c482c488a58e4182f08cf082ed
82b5814182b782ce82e782b582ad91a782cc
8d8782c1"

.COMMENT ### 19 Characters ###
.SEGMENT CHARACTER 744-747 OK
"SJISx97a7"
.SEGMENT CHARACTER 748-749 OK
"SJISx82c4"
.SEGMENT CHARACTER 750-751 OK
"SJISx82c4"
  .
  .
  .
```

```
.SEGMENT CHARACTER 828-829 OK
"SJISx82cc"
.SEGMENT CHARACTER 830-837 OK
"SJISx8d87"
.SEGMENT CHARACTER 838-839 OK
"SJISx82c1"



.COMMENT ### Second Screen ###
.SEGMENT SCREEN 840-1909 OK
"SJISx82bd82c682b182eb82f08ca982b982c4
82ad82ea82e9814282bb82b582c481418175
8c9d8f65...."

.COMMENT ### First Line ###
.SEGMENT LINE 840-929 OK
"SJISx82bd82c682b182eb82f08ca982b982c4
82ad82ea82e9814282bb82b582c481418175
8c9d8f65"

.COMMENT ### 19 Characters ###
.SEGMENT CHARACTER 840-845 OK
"SJISx82bd"
.SEGMENT CHARACTER 846-847 OK
"SJISx82c6"
.SEGMENT CHARACTER 848-851 OK
"SJISx82b1"
  .
  .
  .
```

6. Access

Both databases are now ready to be shipped. We offer two different licenses for both databases: an academic license for research purposes only and a commercial license allowing the licensee to exploit both databases for commercial products. The academic license is priced considerably lower than the commercial version. Please contact one of the authors to get more information on the pricing or to order UP_Kuchibue and UP_Nakayosi directly.

7. Summary

We have presented the UP_Kuchibue and UP_Nakayosi databases in this paper. Both databases are the corresponding Unipen versions of the originally SJIS-coded Kuchibue and Nakayosi databases collected at the Tokyo University of Agriculture and Technology. Both databases together contain more than 3 million characters written by 283 writers. The Unipen format is a common plain ASCII format utilized by western recognizers. UP_Kuchibue and UP_Nakayosi allow western researchers to experiment with Japanese character recognition and apply their algorithms without any coding problems. We expect that UP_Kuchibue and UP_Nakayosi will facilitate cooperation between western and eastern researchers and establish both databases as a widely-recognized benchmark also outside Japan. UP_Kuchibue and UP_Nakayosi are now available. Please contact the authors for more information on accessing both databases.

8. Acknowledgment

We would like to thank Lambert Schomaker and Louis Vuurpijl from the International Unipen Foundation for their assistance in transforming Kuchibue and Nakayosi into the Unipen format.

References

[1] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proc. of the 12th International Conference on Pattern Recognition*, pages 29–33, Jerusalem, Israel, 1994.

[2] International Unipen Foundation. http://www.unipen.org.

[3] S. Jaeger, S. Manke, J. Reichert, and A. Waibel. Online Handwriting Recognition: The Npen++ Recognizer. *International Journal on Document Analysis and Recognition*, 3:169–180, 2001.

[4] K. Matsumoto, T. Fukushima, and M. Nakagawa. Collection and Analysis of On-Line Handwritten Japanese Character Patterns. In *6th International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, 2001.

[5] M. Nakagawa. Non-Keyboard Input of Japanese Text: On-Line Recognition of Handwritten Characters as the Most Hopeful Approach. *Journal of Information Processing*, 13(1):15–34, 1990.

[6] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, and K. Akiyama. On-Line Handwritten Character Pattern Database Sampled in a Sequence of Sentences without Any Writing Instructions. In *Fourth International Conference on Document Analysis and Recognition (ICDAR)*, pages 376–381, Ulm, Germany, 1997.

[7] T. Saito, H. Yamada, and K. Yamamoto. On the Database ETL9 of Handprinted Characters in JIS Chinese Characters and its Analysis (in Japanese). *Transactions of IECEJ*, J.68-D(4):757–764, 1985.

[8] O. Velek, C.-L. Liu, and M. Nakagawa. Generating Realistic Kanji Character Images from On-Line Patterns. In *6th International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, 2001.

[9] C. Viard-Gaudin, P. Lallican, S. Knerr, and P. Binter. The Ireste On-Off (Ironoff) Handwritten Image Database. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 455–458, Bangalore, India, 1999.