UNIPEN project of on-line data exchange and recognizer benchmarks

Isabelle Guyon*, Lambert Schomaker**,
Réjean Plamondon***, Mark Liberman**** and Stan Janet*****

* AT&T Bell Laboratories, USA, isabelle@research.att.com
** Nijmegen Institute for Cognition and Information, The Netherlands
*** IAPR, TC11, Ecole Polytechnique de Montreal, Canada
**** Linguistic Data Consortium, University of Pennsylvania, USA
***** National Institute of Standards and Technologies, USA

Abstract

We report the status of the UNIPEN project of data exchange and recognizer benchmarks started two years ago at the initiative of the International Association of Pattern Recognition (Technical Committee 11). The purpose of the project is to propose and implement solutions to the growing need of handwriting samples for on-line handwriting recognizers used by pen-based computers. Researchers from several companies and universities have agreed on a data format, a platform of data exchange and a protocol for recognizer benchmarks. The on-line handwriting data of concern may include handprint and cursive from various alphabets (including Latin and Chinese), signatures and pen gestures. These data will be compiled and distributed by the Linguistic Data Consortium. The benchmarks will be arbitrated the US National Institute of Standards and Technologies. We give a brief introduction to the UNIPEN format. We explain the protocol of data exchange and benchmarks.

1 Introduction

Pen computers and pen communicators have drawn a lot of interest from the public in the past few months, as the first hand held machines appeared on the market. Pen interfaces are attractive to users, but the handwriting recognition is still disappointing. Researchers are therefore challenged to find better recognition techniques.

Handwriting recognition for pen computers is known as "on-line" handwriting recognition. It addresses the problem of recognizing handwriting from data collected with a sensitive surface which provides the discretized pen trajectory. In contrast, "off-line" handwriting recognition addresses the problem of recognizing optical images of handwriting.

In other pattern recognition fields, such as speech and optical character recognition, significant progresses have been made since large corpora of training and test data are publicly available and public competitions are organized to compare recognition techniques on a fair basis [1, 2, 3, 4, 5, 6, 7, 8].

Funding agencies have not been interested, so far, in funding a similar effort to collect data for on-line handwriting recognition and organize competitive tests. Over the years, companies and universities working in the field have collected their own databases. Obviously, the data is so valuable that everyone is reluctant to make it publicly available. To remedy to these problems, we propose a strategy of data exchange which has received the approbation of leading industry players, including Apple, AT&T, HP, IBM, Microsoft, NCR and Philips and by many university researchers.

2 The UNIPEN format

To facilitate data exchange, a common data format, the UNIPEN format[1], was agreed upon. Users should be able to convert their own format to and from the UNIPEN format easily and hopefully collect new data directly in the UNIPEN format.

[1] This section provides an introduction to the UNIPEN format, but is not a complete description or a manual. Data donors should conform to the instructions which can be obtained by ftp (see section 4)

We developed and tested the format in collaboration with a work group of 14 experts in on-line handwriting recognition. The format incorporates features from the internal formats of IBM, Apple(Tap), Microsoft, Slate (Jot), HP, AT&T, NICI, GO, CIC. It is an ASCII format designed specifically for data collected with any touch sensitive, resistive or electromagnetic device providing discretized pen trajectory information. The minimum number of signal channels is two: X and Y, but more signals are allowed (e.g., pen angle or pressure information). In contrast with binary formats, such as Jot [9], the UNIPEN format is not optimized for data storage or real time data transmission and it is not designed to handle ink manipulation applications involving colors, image rotations, rescaling, etc. However, in the UNIPEN format, there are provisions for data annotation about recording conditions, writers, segmentation, data layout, data quality, labeling and recognition results. To obtain realistic training and benchmarking conditions, UNIPEN data should be as faithful as possible to the original data. Because of the broadly differing methods of data collection, a rich annotation is necessary for UNIPEN to become a success.

Our design efforts focused on making the format:

- human intelligible without documentation (keywords are explicit English words)
- easily machine readable
- compact (few keywords)
- complete (enough keywords)
- expandable

The format is a successions of instructions consisting of a keyword followed by arguments:

- Keywords are reserved words starting with a dot in the first column of a line.
- Arguments are strings or numbers, separated indifferently by spaces, tabulations or new-lines. The arguments relative to a given keyword start after that keyword and end with the apparition of the next keyword or the end of file.

Almost everything is optional, so that simple data sets can be described in a simple way. All variables are global: declared variables values hold until the next similar declaration.

Databases written in the UNIPEN format may optionally be organized in different files and directories, but all the data can also be concatenated into a single file.

The format is thought of as a sequence of pen coordinates, annotated with various information, including segmentation and labeling. The pen trajectory is encoded as a sequence of components[2] $.PEN\_DOWN$ and $.PEN\_UP$, containing pen coordinates (e.g. XY or XYT as declared in $.COORD$). The instruction $.DT$ permits precising the elapsed time between two components. The database is divided into one or several data sets starting with $.START\_SET$. Within a set, components are implicitly numbered, starting from zero.

Segmentation and labeling are provided by the $.SEGMENT$ instruction. Component numbers are used by $.SEGMENT$ to delineate sentences, words, characters. A segmentation hierarchy (e.g. $SENTENCE\ WORD\ CHARACTER$) is declared with $.HIERARCHY$. Because components are referred by a unique combination of set name and order number in that set, it is possible to separate the $.SEGMENT$ from the data itself.

The format also provides a unified way of encoding recognizer outputs to be used for benchmark purpose.

Typical data files written in the UNIPEN format are appended.

3 Benchmark protocol

The participation to the benchmarks organized in the framework of the UNIPEN project is subjected to providing a minimum amount of handwriting data. This ensures that a large database containing a wide variety of samples will progressively be constituted.

A first test is currently organized jointly by the Technical Committee 11 of the International Association of Pattern Recognition, the Linguistic Data Consortium (LDC) [10] and the US National Institute of Standards and Technologies (NIST). This test concerns only the Latin alphabet, isolated characters,

[2] To keep the format as general as possible with respect to future developments and extensions as well as to be consistent with previous psychophysical studies on handwriting generation and perception, we use here the terminology proposed in Plamondon and Maarse (1989). We avoid using the term "stroke" defined as a fundamental unit of handwriting movement, with respect to a specific generation model. We use instead the term "component" for any continuous trace of the pen recorded by the digitizer. A pen-down component is a trace recorded when the pen is in contact with the surface of the digitizer. A pen-up component is a trace recorded when the pen is at proximity of the digitizer without touching it. Components are not necessarily delimited by pen-lifts and may or may not coincide with strokes.

words and sentences. Other tests for other alphabets, gestures, signatures, etc. will be organized in the future.

The minimum amount of data required for participation in this test is a set of sentences, words or isolated characters containing at least 12,000 characters. From this set, 2000 characters will be used as development test set, immediately made available to other participants, and 10,000 will be held by NIST to serve as benchmark test set.

Donation of large sets of training data is not mandatory but encouraged. Participants providing more than 50,000 characters will receive a money compensation from LDC. The data received from participants will be complemented with training data collected by LDC.

The data is sent to a designated FTP site at the Linguistic Data Consortium. NIST reserves the right of rejecting participants whose data is of too poor quality. A data browser for DOS or X Windows systems is provided upon request to check the data quality.

Benchmark test data is kept confidential and reviewed by NIST before release. Deprived of class labels, the test sets will be made available by NIST, at the LDC FTP site. NIST will compute statistics from the recognition results. The results will be reported at a workshop to which all participants of the benchmark will be invited.

After the benchmark, the data will be open to the public and distributed by LDC.

4 The ftp site at LDC

An FTP site has been set up at the Linguistic Data Consortium for data and software exchange. Currently, most of the directories can be read only by registered benchmark participants. After the benchmark, more directories will be open.

To access the directories that are publicly available, proceed as follows:

```
ftp ftp.cis.upenn.edu
Name: anonymous
Password: [use your email address]
ftp> cd pub/UNIPEN-pub/documents
ftp> get call-for-data.ps (benchmark instruc)
ftp> cd ../definition
ftp> get unipen.def       (format definition)
ftp> quit
```

Acknowledgments

The handwriting recognition experts who participated to the first UNIPEN workshop and to the successive evaluations of intermediate releases of the data format and benchmark protocol are very gratefully acknowledged. Special thanks to Donnie Henderson for using the format at an early stage in his data collection program, to Wes Hunter for using the format in the internal benchmark organized by NCR and to Gebern Abbink for developing the Upview library.

References

[1] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA speech recognition research database: specifications and status. In *Proceedings of the DARPA speech recognition workshop*, 1986.

[2] L. Lamel, R. Kassel, and S. Seneff. Speech database development: design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA speech recognition workshop*, 1987.

[3] P. J. Price, W. M. Fisher, and J. Bernstein. The DARPA 1000-word resource management database for continuous speech recognition. In *Proceedings of ICASSP*, 1988.

[4] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The ATIS spoken language system pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, 1990.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, 1992.

[6] J. Hull. Large database organization for document images. In *Proceedings of NATO-ASI summer school on fundamentals in handwriting recognition*, 1993.

[7] R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. Wilson. The first census optical character recognition systems conference. Technical Report NISTIR-4912, NIST, US Department of Commerce, 1992.

[8] I. T. Phillips, J. H. R. Haralick, and D. Dori. The implementation methodology for a CD-ROM

English database. In *Proceedings of ICDAR*, Tsukuba, Japan, October 1993. IAPR-IEEE.

[9] D. Gerrety. JOT, a specification for an ink storage and interchange format. Technical Report draft version 0.99, Slate corporation, San Mateo, California, USA, 1993.

[10] M. Liberman and V. Zue. Introduction to the Linguistic Data Consortium. Technical report, LDC, University of Pennsylvania, Philadelphia, USA, 1993.

An example of a typical UNIPEN set.

```
================== File ATT.DOC ==================

.VERSION                0.6

.DATA_SOURCE            ATT
.DATA_ID                SAMPLE

.COMMENT                ###################
                        ### INFORMATION ###
                        ###################
.DATA_CONTACT
        Name: Dept 11359 - Adaptive Systems Research
        Affiliation: AT&T Bell Laboratories
        Address: 101 Crawfords Corner Rd
        Holmdel, NJ, 07733 , USA
        Phone: 1 (908) 949-2783 (secty.)
        Fax: 1 (908) 949-7722
        Tech Contacts: Don Henderson   908-949-4591
        Isabelle Guyon  415-442-5937
.DATA_INFO
        Alphabet: Latin, uppercase, lowercase, punct
        Lexicon: The text from "Alice In Wonderland"
        Quantity: very small sample
        Quality: ?
        Distribution: Given by .LEXICAL_FREQ
        Number of writer(s): 1
        Writing style: Natural - no constraints....
        Segmentation: Words written in boxes in
        upper area of form. Last line on
        the form is a unseg'd sentence.
.SETUP
        Site: AT&T Bell Laboratories
        Holmdel cafeteria, New Jersey, USA
        Time:
        Conditions: Supervised recording conditions
        People sitting at a desk
        Writers: Volunteer staff members
        Form layout: 1 form containing 1 segmented
        word area (1-12 words) and a second
        area having a sentence made up of these
        words.
.PAD
        Machine name: WACOM HD-648A LCD Digitizer &
                      386 PC
        Display: Backlit LCD
        Matrix 640  x 480 points
        VGA standard, black on white, monochrome
        Sensor: Electromagnetic resonance sensor
        Pen: Untethered pen, tip switch and
             side button
        Driver: PC driver developed in house..

.COMMENT                ###################
                        ### DATA LAYOUT ###
                        ###################
.X_DIM                  4160
.Y_DIM                  3200
.H_LINE                 450 1900 2300

.COMMENT                ###################
                        ### UNIT SYSTEM ###
                        ###################
.X_POINTS_PER_INCH      508
.Y_POINTS_PER_INCH      508
.POINTS_PER_SECOND      200

.COMMENT                ####################
                        ### DECLARATIONS ###
                        ####################
.COORD                  X Y
.HIERARCHY              PAGE TEXT_CHUNK WORD

=============== End file ATT.DOC ==================


=============== File ATT.DAT ======================

.COMMENT  Most of the point have been
                        removed to shorten the
                        example.

.INCLUDE                ATT.DOC

.DATE                   9 20 93

.WRITER_ID              08171408_14804
.STYLE                  MIXED

.START_BOX
.SEGMENT PAGE           0-58
.SEGMENT TEXT_CHUNK     0-29 ?
                      "that nothing more happened ,"
.SEGMENT WORD           0-6 ? "that"
.PEN_DOWN
707 2417
707 2424
590 2319
.PEN_UP
.DT 151
.PEN_DOWN
588 2377
586 2377
695 2393
.PEN_UP
.DT 231
.PEN_DOWN
772 2456
771 2456
745 2343
.PEN_UP
.DT 231
.PEN_DOWN
```

```
827 2384
826 2384
827 2340
.PEN_UP
.DT 201
.PEN_DOWN
818 2362
819 2362
871 2361
.PEN_UP
.DT 231
.PEN_DOWN
929 2411
929 2411
902 2340
.PEN_UP
.DT 151
.PEN_DOWN
882 2370
882 2370
995 2384
.PEN_UP
.SEGMENT WORD           7-15 ? "nothing"
.PEN_DOWN
1778 2366
1778 2366
1778 2366
.PEN_UP

.COMMENT For more examples, please ftp the data.
=============== End file ATT.DAT ===================

=============== File NADANET.REC ===================

.VERSION 0.6

.REC_SOURCE     ATT_HO4G

.REC_ID         NADANET

.REC_CONTACT    Name: Isabelle Guyon
                Affiliation: ATT Bell Labs
                Address: 50 Fremont street,
                         40th floor
                San Francisco, CA 94105, USA
                Phone: (415) 442 5937
                       (510) 524 5328
                Fax: (415) 442 5967
                Email: isabelle@research.att.com

.REC_INFO       Recognizer: Time Delay Neural Net
                Designer name: Nada Matic
                Creation date: 1992
                Last update: October 1992
                Preprocessing: Fixed-length-convert
                               7 features
                Number of frames in input: 90
                Network name: allclass.fp
                Alphabet: Latin, all keyboard char.
                Postprocessing: none

.IMPLEMENT      Lisp and C code, run under UNIX,
                on host globi (SUN sparc-station)

.TRAINING_DATA  GO-training
```

```
.TEST_DATA      NCR

================= End file NADANET.REC =============

================= File NADANET.RES =================

.INCLUDE        NADANET.REC

.DATE           9 13 93

.START_SET

.REC_LABELS 0-2 ? "A" "i" "O" "1" "I"
.REC_SCORES 0-2 0  0.49  0.30  0.28  0.27  0.26

.REC_LABELS 3 ? "B" "D" "5" "y" "3"
.REC_SCORES 3 0  0.45  0.31  0.27  0.25  0.25

.REC_LABELS 4 ? "C" "e" "c" "6" "{"
.REC_SCORES 4 0  0.35  0.30  0.30  0.25  0.25

.REC_LABELS 5 ? "D" "b" "N" "B" "3"
.REC_SCORES 5 0  0.43  0.27  0.26  0.26  0.26

.REC_LABELS 7-9 ? "E" "8" "I" "?" "H"
.REC_SCORES 7-9 0  0.36  0.35  0.29  0.27  0.26

.REC_LABELS 10 ? "E" "5" "I" "F" "%"
.REC_SCORES 10 0  0.36  0.33  0.30  0.30  0.29

.REC_LABELS 12 ? "G" "a" "x" "d" "h"
.REC_SCORES 12 0  0.38  0.31  0.26  0.25  0.25

.REC_LABELS 14-16 ? "H" "A" "\"" "P" "p"
.REC_SCORES 14-16 0  0.41  0.32  0.28  0.27  0.26

.REC_LABELS 17 ? "I" "1" "|" "l" "@"
.REC_SCORES 17 0  0.33  0.32  0.29  0.27  0.27

.REC_LABELS 18 ? "J" "j" "S" "s" "d"
.REC_SCORES 18 0  0.39  0.33  0.30  0.28  0.26

.REC_TIME 0-18 1.150000

============== End file NADANET.REC ================
```