The IRESTE On/Off (IRONOFF) Dual Handwriting Database

Christian VIARD-GAUDIN♣, Pierre Michel LALLICAN♣, Stefan KNERR♦, Philippe BINTER♣

♣ : IRESTE, laboratoire SEI/EP CNRS 063,
Rue Christian PAUC – BP 60603 – 44087 Nantes Cedex 03 – France

♦ : VISION OBJECTS
11 Rue de la Fontaine Caron, 44300 Nantes – France

Abstract

Databases for Character Recognition algorithms are of fundamental interest for the training of statistics based recognition methods (Neural Networks, Hidden Markov Models) as well as for benchmarking existing recognition systems. Such databases currently exist, but none of them gives access to the online data (pen trajectory) and offline data (digital images) for the same writing signal. We have developed such a dual on/off database, named IRONOFF. Currently, it contains a large number of isolated characters, digits, and cursive words written by French writers. We have designed this database so that, given an online point, it can be mapped at the correct location in the corresponding scanned image, and conversely, each offline pixel can be temporally indexed. Since we think this database is of interest for a large part of the research community, we make it publicly available.

I. Introduction

Publicly available data bases are important for the research community in order to test new ideas and algorithms and to perform benchmarks and thereby measure progress and general tendencies.

Several data bases containing printed or handwritten characters, words or documents have been made available in recent years. All of them are devoted to one of the following three specific character recognition problems.

1. Off-line machine print recognition

The historically first domain of character recognition is the recognition of machine printed characters (OCR, for Optical Character Recognition). The UW-I and UW-II image databases have been designed to address the needs of researchers in page segmentation and machine printed text recognition [1]. They contain more than 2,000 pages of English and Japanese technical journal articles.

2. On-line handwriting recognition

Online handwriting is produced by the means of a digital tablet and an electronic pen. The data is acquired while the writing is in progress and consists of the discrete pen trajectory information. The output is a temporal sequence of (x, y) coordinates. Additional information such as pressure and tilt angles of the pen may also be available for each point of the trajectory. In the field of online handwriting, several non-public databases exist. Since 1993, the UNIPEN project has organized the collection of more than 5 million handwritten characters from several countries [2]. Unfortunately, up to now, this data base is not publicly available.

3. Off-line handwriting recognition

In the case of off-line handwriting recognition, a paper document including handwritten information, such as checks, forms, envelopes, etc., is scanned into a digital image. For this kind of problem, there also exist several data bases which contain examples of isolated handwritten characters, digits, isolated words, and phrases. For instance, the CEDAR data base [3] has been extracted from real mail stream and contains nearly 50,000 samples of isolated characters and digits. Another popular data base is the CENPARMI data base which contains 17,000 isolated digits extracted from postal ZIP codes [4]. The largest publicly available data base is the NIST SD3 data set, which has been extracted from census forms and contains over 300,000 character images.

These databases contain (i) mainly English vocabulary written by people used to the American writing style which is somewhat different from European and other countries, and (ii) concerning handwritten data sets, they have been designed specifically either for online recognition (UNIPEN) or for offline recognition (CEDAR, CENPARMI, NIST) ; basically because the acquisition tools and the data formats are different. However, there is a need in the research community for a handwriting data base (isolated character or words) for which both the off-line image and the on-line trajectory is available. One interest concerns the evaluation of skeleton algorithms. Here, the online data could provide a way to compare the skeleton points (offline image) to an objective trajectory (online coordinates). It also becomes possible to study the correlation that could exist between the pressure or the speed of the pen and the gray level distribution or the width of the corresponding strokes.

Another field of interest is directly linked to offline character recognition. Many approaches attempt to represent a handwritten character or word as a sequence of small segments, the granularity of which is variable from pixel slices to character. However, unlike speech recognition, the ordering of these segments (essentially a 2D information) is not trivial, and the need for a good segmentation algorithm cannot be overstressed. Currently, no satisfactory representation exists for offline handwriting. If the online data is jointly accessible with the offline images, it can be used to recover the temporal order of strokes from the offline images and thereby guide and train the segmentation to provide a relevant frame description [5]. In that sense, such approaches bridge the gap between online and offline character recognition methods [6], [7] which is very attractive since it has been shown that online handwriting exhibits superior results compared to offline recognition [8].

This paper presents a methodology for the construction of a dual on/off database which has been intended for research on the use of online information for the design and training of an offline handwriting recognition system. However, we are confident that it will enable many other experiments. A large number of samples of isolated characters, digits and French words have already been collected. We briefly present the content of the resulting database (IRONOFF database = IRESTE ON/OFF).

II. Online versus offline data

In order to have access to the online trajectory as well as to the digital image for the same handwriting signal, some important points have to be considered.

Of course, there is no transformation that allows to reconstruct one type of data from the other. Each of them, has specific attributes not available in the dual representation. For instance, online data do not convey any information about the width of the strokes and of its possible variation along the tracing. Conversely, offline images have lost all the dynamical relationships among its basic constituents. As a result, strokes which are superimposed in the spatial domain are not easy to segment, while they can be easily separated if the time dimension is considered.

However, going from the online data to the offline image appears to be much easier than the reverse transformation. One solution is to synthesize a 2D image by interpolating the tracing, using a spline function for instance, between the digitized online points. Next, some morphological operators can be applied to enlarge the tracing to the desired stroke width. Furthermore, to have a more realistic representation, some controlled degradations can be added to the character images [9].

Although this way of generating a dual online/offline database is appealing because of its simplicity, we have preferred to work on real digital images as coming from a scanner. These images show typically defects, noise and deformation which are not present in the online data. Consequently, we have performed two steps in the acquisition stage. One uses a digital tablet to which we attach a paper to be filled with an electronic ink-pen. Then, in a second step, this form is scanned to provide the offline image, figure 1.

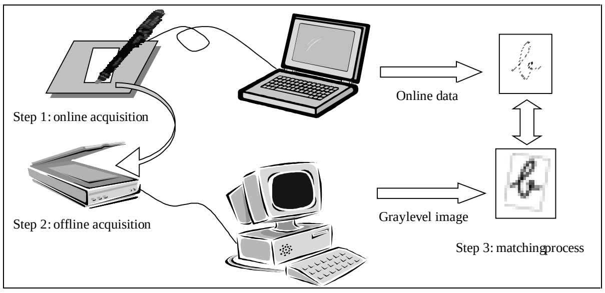


Figure 1 : Overall scheme for the database construction

So, for each sample in the database, two complementary files are available. One of them, produced during the online acquisition, contains the list of the coordinates of the pen trajectory, the other one is the digital image of this piece of handwriting produced by the scanner. Of course, these two types of information should be available within the same coordinate system, with the same origin, the same resolution and orientation. But, as the two acquisition systems (tablet and scanner) are processing the writing separately and with their own parameters, this assumption is not satisfied directly. Geometrical transformations have to be applied to compensate for these differences.

III. Contents and technical specifications

We have collected data of isolated characters, digits and cursive words. We use A4-forms with predefined boxes printed in a very light yellow which facilitates their removal from the image later on. Above each box, the ground truth of the character or word to be written is provided in machine print. A human operator verifies each filled form to ensure that it has been correctly filled. Presently three types of forms, FormB, C and D, have been used. FormB contains the 26 letters of the alphabet, in lower and in upper case character, the 10 digits and the Euro symbol, and at last the words involved in French check processing. FormC (Figure 2) and D are forms that contain French cursive words.

Currently, approximately thousand forms have been collected (32,000 isolated characters and 50,000 cursive words). Each of them has been filled by a different writer. Data collection is still in progress. In the online domain, the forms have been sampled with a spatial resolution of 300 dpi and a sampling rate of 100 points/s (Wacom UltraPadA4) and are stored using the UNIPEN format [10]. An example of such a file is given in figure 3. In addition to the x and y coordinates, the pressure of the pen and time have also been included.

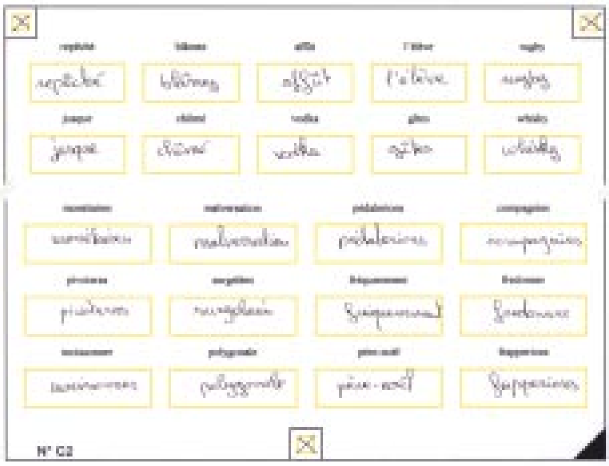


Figure 2 : FormC

```
.DATE           10 23 1998
.KEYWORD .CALIBRATION
.COMMENT  Computed_points_for_calibration
.COMMENT  Strokes'_bounding box

.COMMENT Declarations
.X_DIM          3600
.Y_DIM          3600
.X_POINTS_PER_INCH 300
.Y_POINTS_PER_INCH 300
.POINTS_PER_SECOND 100
.COORD   X Y P T

.COMMENT Data
.WRITER_ID      unknown
.COUNTRY France
.HAND           R
.AGE            36
.SEX            M
.CALIBRATION   <-----  Coordinates of the landmarks
809     215
2959    245
1818    3372
.PEN_DOWN
884     407     38      0
884     406     79      12
885     406     113     22
.............................
876     412     204     96
872     417     202     106
.PEN_UP
.PEN_DOWN
...............................
```

Figure 3 : Example of UNIPEN file

To collect the data, a friendly Graphical User Interface (GUI) has been developed on a PC/NT window environment. The GUI has been developed in Visual C++ and based on the Microsoft Foundation Classes (MFC). In order to drive the tablet with high level functions, it uses the WinTab library [11].

In the offline domain, the images are scanned with a resolution of 300 dpi with 8 bits per pixel (256 gray levels). Another tool has been developed to segment the form into its data-fields and to display simultaneously the offline image of the fields and the corresponding online points. Other interesting tools have been included which allow to dynamically follow the tracing. Figure 4 illustrates the GUI.

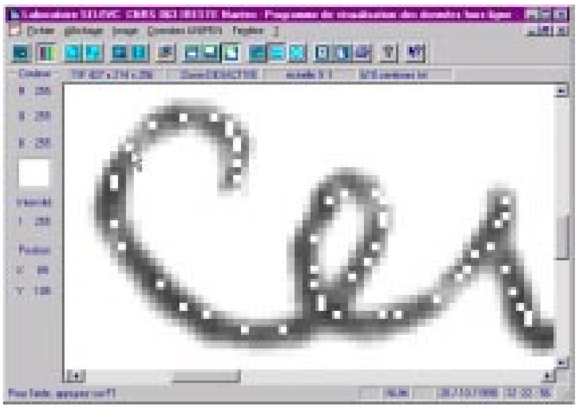


Figure 4 : GUI of the visualization tool of the database

IV. ON/OFF mapping procedure

IV. 1 Global transformation

The data coming from the tablet (online data) is mapped to the same coordinate system than the data coming from the scanner (offline data). In order to achieve this mapping, it is necessary to know the coordinates of some landmarks in both original coordinate systems. From these points, parameters of the mapping can be computed.

We assume that the global transformation is composed of three basic transformations : a translation of the coordinate system origin, a rotation of the axis's and a scaling transformation. These transformations are affine transformations and can be represented in matrix notation :

$$(x_T, y_T)^T = T \bullet H \bullet R \, (x_S, y_S)^T \qquad \text{equation (1)}$$

where (x_T, y_T) are the coordinates of a point in the Tablet system, (x_S, y_S) are the coordinates of the same point in the Scanner system, T is a translation vector, H contains the scaling coefficients and R is a rotation matrix :

$$x_T = z_x \, (\cos \alpha \, x_S + \sin \alpha \, y_S) - T_x$$

$$y_T = z_y \, (-\sin \alpha \, x_S + \cos \alpha \, y_S) - T_y$$

To estimate the parameters (α, z_X, z_Y and T_x, T_y), it is required to know at least two corresponding points in the two coordinate systems. The difference in the orientation of the segment formed by these two points gives the α angle, the ratio of the length between the two segments allows to compute the z parameters, and the position of one point in both coordinate systems determines the translation vector.

IV.2 Landmark definition and extraction

The landmarks are specific points which should be automatically extracted both in the online domain and in the offline domain. As the position and orientation of the form on the tablet is not known in advance (it would be a too severe constraint to position the form exactly at the same location every time), it is the writer who defines the

location of the landmarks. We have introduced three points (instead of the minimum of 2) in order to have a more robust parameter estimation. They are defined by 3 handwritten crosses which are drawn in an area reserved for this purpose, cf. figure 2. The landmark position is the central point of a cross.
The crosses are allowed to be drawn with very variable shapes. Examples are displayed in figure 5.

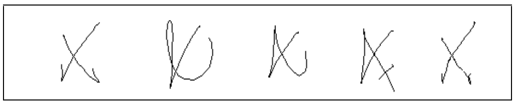
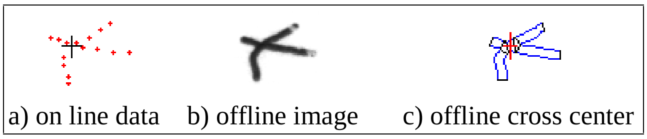


Figure 5 : Variability of the cross shapes

The central point of the crosses are extracted both in the online domain and offline image, figure 6.



a) on line data b) offline image c) offline cross center

Figure 6 : Landmark extraction

IV.4 Online to offline matching

Given the parameter values (z, α and T_x,T_y) of the global transformation, the online points are projected into the offline image (or reversibly a pixel from the offline image can be projected in the online coordinate system) using equation (1). This transformation is performed globally for every point of the online file. Due to some computational approximations that have been made during the modeling of the global transformation and also due to limitations on hardware accuracy (tablet and scanner), the two types of data do not perfectly superpose everywhere in the form image, figure 7. To enhance the matching process, a local refinement of the translation vector has been developed which is performed on a field by field basis.

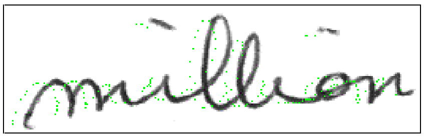
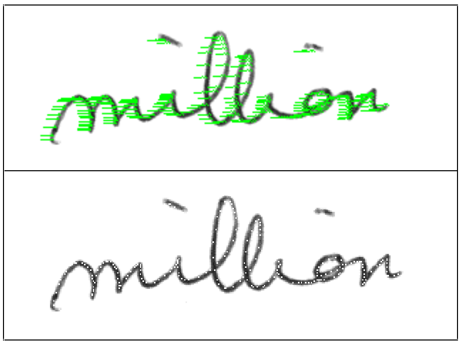


Figure 7 : Online points projected in the offline image

First, the gradient of the gray-level offline image is computed. For each online point mapped into this gradient image, we estimate the gradient direction. Averaging all these directions provides the direction in which the online points have to be moved in order to converge towards the offline handwriting signal (Figure 8-a). This procedure is repeated until a stable position of the set of online points is encountered (white dots in Figure 8-b). Figure 9 shows an example of an image and its corresponding online points.

V. Conclusions

Although several databases dedicated to handwriting recognition exist throughout the scientific community, none of them contains both the online data and the offline image for the same samples of writing. We have defined a methodology that allows the construction of such a database and presented the current content of the first release of this database. The main limitation regarding this database is that the same inking pen, although with two different colors (blue and black), has been used for all writers. We believe that the availability of this dual database will open new perspectives and push further the frontiers of handwritten character recognition.



a) Displacement of the online points

b) terminal position of the online points in the image

Figure 8 : The local refinement matching

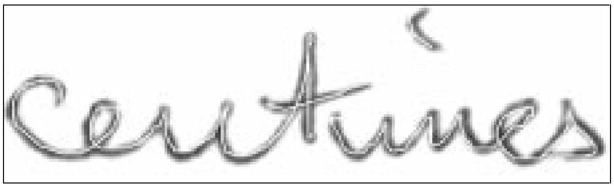


Figure 9 : An image and the online trajectory

(The IRONOFF database can be obtained by contacting :
Christian VIARD-GAUDIN : cviard@ireste.fr)

References :

[1] I.T. Philips, S. Chen, R.M. Haralick, “CD-ROM Document Database Standard”, ICDAR’93, Tsukuba, Japan, Oct. 1993, 484-487.

[2] I. Guyon et al., “Data sets for OCR and document image understanding Research”, Handbook of character recognition and document proc., Word Scientific, 779-799.

[3] J.J. Hull, “A database for handwritten text recognition research”, IEEE Trans. On PAMI, 16,5 (1994) 550-554.

[4] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, “Computer recognition of unconstrained handwritten numerals”, Proc. of the IEEE, 80, 7, 1992, 1162-1180.

[5] P.M. Lallican et al., “An offline handwriting recognition system trained with online data”, paper in preparation.

[6] S. Jäger, “Recovery dynamic information from static, handwritten word images”, Ph D.Thesis , Daimler-Benz AG Research and Tech., Verlag Dietmar Fölbach, 1998.

[7] P.M. Lallican, C. Viard-Gaudin, “Off-line handwriting modeling as a trajectory tracking problem”, IWFHR’6, Taejon, Korea, Aug. 1998, 347-356.

[8] R. Seiler, M. Schenkel, F. Eggimann, “Off-line cursive handwriting recognition compared with On-line recognition”, ICPR’96, Vienna, 505-509.

[9] H.S. Baird, “Document image defect models”, Structured Doc. Image Analysis, Springer-Verlag,, 1992, 546-556.

[10] http://hwr.nici.kun.nl/unipen/unipen-history.html

[11] http://www.pointing.com