A new Cursive Basic Word Database for Bank-check Processing Systems

S. Impedovo IAPR Fellow, IEEE S.M., G. Facchini, F.M. Mangini

Department of Computer Science
University of the Studies of Bari "Aldo Moro"
Via Edoardo Orabona, 4 – 70125 Bari – Italy
impedovo@di.uniba.it

Interfaculty Centre "Rete Puglia"
University of the Studies of Bari "Aldo Moro"
Via Giulio Petroni, 15/F.1 – 70126 Bari – Italy
http://www.retepuglia.uniba.it/Impedovo/index.htm

Abstract—**In this paper is presented a new database for handwritten cursive basic words recognition. The database is devoted to research on bank-check processing. In fact, for security reasons, the banks rarely allow the treatment of checks handled by them. On the other hand, the reasons for the English basic words choice lie in the fact that English language is the world most commonly used for bank-check drawing. The database realised includes a considerable number of instances of basic words. Pattern images are stored using a standard image format that will be available to all researchers by Internet. The importance of this work lies in the fact that the database is queried by the network, giving the possibility to grow with the contribution of others researchers. The tagging has been generated by using the XML language that allows recovering also information on the writers. Furthermore, the information handled not only could allow the semantic recognition of the specimen but also the research development on the author identity of the manuscript. The database is an open one to be increased with the contribution of all others researchers in the world.**

Keywords - Pattern Recognition; Handwriting Recognition; Basic Word; Tagging; Database

I. INTRODUCTION

The development of standard database is crucial for advancing research in handwriting recognition. In fact, for security reasons the bank system does not make available their databases for researches grand. On the contrary, the production of large databases is undoubtedly necessary to evaluate the new recognition algorithms and to compare them [1]. In other words, standard databases for handwriting recognition are an essential requirement for the development, training, evaluation and comparison of different bank-check recognition system.

Several databases for research in handwriting recognition have been realised so far [2]. Some of them concerns on-line handwriting, in this case the hand-written patterns are acquired during the writing process by a graphic tablet or an integrated tablet-display device [3]. Other databases concerns off-line handwriting, in this case the hand-written patterns are acquired by optical scanners or cameras after that the writing process has been completed. Several databases of off-line hand-written patterns have been developed so far. Some of them contains isolated characters and digits, others contains isolated words or entire phases, but doesn't exist a free database for testing bank check recognition systems. We are producing a free database usable by Internet also to be enlarged with contribution of all others researchers everywhere in the world.

The database product, named SIDB, at the moment includes 95,760 specimens realised by 380 writers, 36 basic words of the legal English amounts that allow composing any other amount. This database in our experiments is used to support the research towards the development of recognition systems based on HMM in word recognition. In order to present our database, in section 2 is reported an overview of widely used databases in the field of handwriting recognition. In section 3, the data acquisition procedure is presented. Section 4 highlights the design and structure of the database. Section 5 illustrates various pre-processing algorithms to support efficient processing. Section 6 and Section 7 show, respectively, the data labels of the writers and the discovery of these and other data relating to patterns. Finally, in section 8 are given the conclusions and future works are proposed.

II. OVERVIEW ON DATABASE TO HANDWRITING RECOGNITION

The CENPARMI database contains about 17,000 isolated digits extracted from ZIP code images. The ZIP code images were scanned at 200 dpi in one-bit gray scale and segmented manually by an human operator which also assigned a true value to each segmented digit [4].

The NIST database ST3 contains more than 300,000 character images extracted from filled forms. The forms were scanned at 300 dpi in one-bit gray scale and automatically segmented. Since the way in which the form should be filled in is known, the label of the digit that had to be written in each position of the form has been used as a truth value for the image extracted from that position. The NIST database also contains instances of running text (SD3) and phases (SD11-SD13) [5].

The CEDAR database contains about 28,000 instances of isolated hand-written characters and digits extracted from the address images on mail pieces. The address images were scanned at 300 dpi in 8-bit gray scale and successively converted to one-bit images. A semi-automated process has been used to segment the address into connected components which have been then treated by a human operator. From the ZIP codes, an additional set of about 21,000 digits has been extracted by an automated segmentation procedure. The CEDAR database also contains the images of about 5,000 ZIP codes, 5,000 city names and 9,000 state names [6].

The ETL Japanese database ETL9 contains more than 607,000 character images written by 4,000 authors [7].

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.5

The BERN database from Bern University of full English sentences consists of more than 43,700 instances of handwritten words. The underlying lexicon includes about 6,600 different words. The database is based on the Lancaster-Oslo/Bergen corpus, a collection of 500 English texts each consisting of about 200 words. 556 forms filled in by about 250 writers have been scanned at 300 dpi in eight-bit gray scale [8].

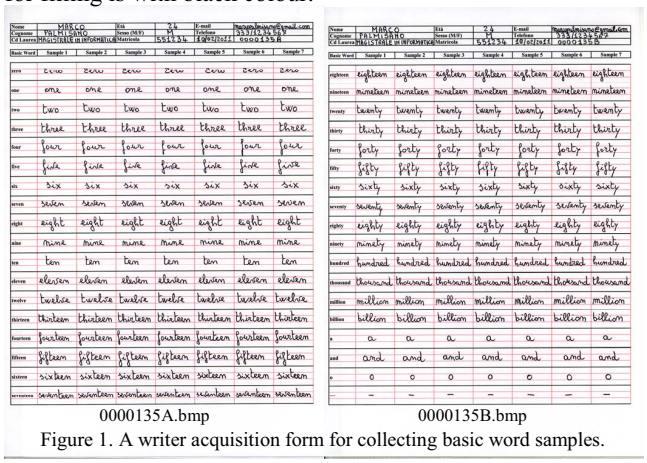
The UNIPEN database has been produced in collaboration by large number of academic institutions and companies, among which Apple Computer, Hewlett Packard, IBM Research Center, University of Genoa and Princeton University. The project's goal is to create a collection of images on a large scale for use in optical character recognition. The database has more than a million specimens, from different organizations who participated in the work. The huge amount of data is organized in a clear hierarchy: a more general level, the available images are grouped according to whether they represent characters, words or entire portions of text. Each subset is further divided based on additional characteristics of the samples. The result of the classification is arranged in eleven subsets, which contain a number of images ranging from about 15,000 to more than 120,000 [9].

The databases considered are oriented to work and a sharing of the results on a larger scale, often the result of the collaboration of several institutions and companies. Nevertheless, many other databases have been made, to meet specific needs, based on western and eastern languages and alphabets. Among the western databases, it mention: IAM database, SRTP Database, IRESTE ON-OFF (IRONOFF) Database, AWS-1334 Single Writer Database; among the eastern databases, it mention: KAIST Hangul Database, KAIST Hanja DB1, KAIST Hanja DB2, KAIST-POWORD, KAIST KMail98, KAIST OP 2 DB.

In particular, the IAM Handwriting Database contains forms of unconstrained handwritten English text, which were scanned at a resolution of 300 dpi and saved as PNG images with 256 gray levels. The IAM Handwriting Database 3.0 is structured as follows: 657 writers contributed samples of their handwriting, 1,539 pages of scanned text, 5,685 isolated and labeled sentences, 13,353 isolated and labeled text lines, 115,320 isolated and labeled words. The words have been extracted from pages of scanned text using an automatic segmentation scheme and were verified manually [10].

III. DATA ACQUISITION

The acquisition form proposed to the writers for data acquisition, consist of two sheets, see figure 1. The figure shows a scheme already drawn by a writer. The scheme consists of 36 basic words for 7 samples of each. 380 individuals (writers) of various age, sex and cultural level have been engaged for this purpose. The basic words 36 are: "one, two, tree, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, hundred, thousand, million, billion, a, and, o, -". The schemes have been filled in without any limitation in the duration of the process; while the type of the pen to be used for filling is with black colour.



0000135A.bmp 0000135B.bmp

Figure 1. A writer acquisition form for collecting basic word samples.

The policy adopted so that the acquisition schemes could be scanned, and hence the patterns to be included in the database, consists of a review by the database administrator to ensure that the basic word produced did not contain grammatical errors, smudges, omissions, exit out of the box.

Then, each acquisition schema have been digitized by Epson Perfection 3170 Photo, scanned at resolution of 300 dpi in 24-bit colours scale, retailed in the format A4 (dimension 210 × 297 mm) and stored in two file, one for each sheet, in bitmap format with no compression, into the same folder. The images of size 2480 (width) x 3507 (height), they are 24-bit RGB colour (R8 G8 B8). The storage space of each image is 25,481 KB. Then, they are converted in one gray scale.

The name assigned to each file is a 7-digit serial number that identifies the writer, divided into "A" for the first sheet and "B" for the second sheet. As a result, every file in the folder is automatically processed with the "Threshold" tool and then use the "Retail" tool. It is implemented the "Threshold" tool with which you can make visible or invisible pixel level according to their colour intensity value of the same points. The operator "Threshold" setting the intensity value 211, allows doing away with the horizontal and vertical obtaining a gray level image. It is implemented the "Retail" tool, which is a procedure of automatic retailing of the 252 patterns, resulting in 252 files that are sorted as BMP file in the right accommodation in the database. The name assigned to each file contains different information:

- The first 7 digits represent the serial number for the author, and thus the schema of acquisition;
- The next 2 digits represent the serial number for the basic word;
- The last digit represents the number of samples of basic words.

The size of each pattern is 304 × 132 pixels. Figure 2 shows a pattern called 0000010-08-5.bmp, which is the fifth sample of the basic word "eight" product by eleventh writer. Figure 3 shows a pattern called 0000015-15-1.bmp, which is the one sample of the basic word "fifteen" product by fifteenth writer.

Figure 2. The pattern 0000010-08-5.bmp.

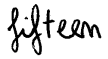


Figure 3. The pattern 0000015-15-1.bmp.

Successively the patterns are automatically converted to one-bit images. Personal data are collected manually and loaded into a text file. This file is used when using XML markup.
The amount will be formulated according to the following syntax rules:

<AMOUNT_CHECK>::=<AMOUNT_BILLION>+<AMOUNT_MILLION>+<AMOUNT_THOUSAND>+<AMOUNT_HUNDREDS>/<AMOUNT_BILLION>+<AMOUNT_THOUSAND>+<AMOUNT_HUNDREDS>/<AMOUNT_BILLION>+<AMOUNT_MILLION>+<AMOUNT_HUNDREDS>/<AMOUNT_BILLION>+<AMOUNT_MILLION>+<AMOUNT_THOUSAND>/<AMOUNT_BILLION>+<AMOUNT_HUNDREDS>/<AMOUNT_BILLION>/<AMOUNT_MILLION>+<AMOUNT_THOUSAND>+<AMOUNT_HUNDREDS>/<AMOUNT_MILLION>+<AMOUNT_THOUSAND>/<AMOUNT_MILLION>+<AMOUNT_HUNDREDS>/<AMOUNT_MILLION>/<AMOUNT_THOUSAND>+<AMOUNT_HUNDREDS>/<AMOUNT_THOUSAND>/<AMOUNT_HUNDREDS>
<AMOUNT_BILLION>::=<AMOUNT_HUNDREDS>+billion/<AMOUNT_COMPOSITE>+billion
<AMOUNT_MILLION>::=<AMOUNT_HUNDREDS>+million/<AMOUNT_COMPOSITE>+million
<AMOUNT_THOUSAND>::=<AMOUNT_HUNDREDS>+thousand/<AMOUNT_COMPOSITE>+thousand
<AMOUNT_HUNDREDS>::=<AMOUNT_COMPOSITE>+hundred
<AMOUNT_COMPOSITE>::=<AMOUNT_TWO_DIGITS>/<AMOUNT_TWO_DIGITS_SIMPLE>+<AMOUNT_ONE_DIGIT>/<AMOUNT_TWO_DIGITS_SIMPLE>/<AMOUNT_ONE_DIGIT>
<AMOUNT_TWO_DIGITS>::=ten/eleven/twelve/thirteen/fourteen/fifteen/sixteen/seventeen/eighteen/nineteen
<AMOUNT_TWO_DIGITS_SIMPLE>::=twenty/thirty/forty/fifty/sixty/seventy/eighty/ninety
<AMOUNT_ONE_DIGIT>::=one/two/three/four/five/six/seven/eight/nine

Therefore, any English legal amount of bank checks is based on the lexicon of basic words required by the acquisition schedule.

IV. Database Structure

The database has been realised by extracting hand-written patterns from filled forms. It is structured in a folder tree to three levels. The samples after being identified and cut from the acquisition module are automatically placed in its subfolders, following the granting of a name to the files created. The database contains 95,760 digital images (pattern) of 380 writers.

A. Data Storage

All module images of the basic words are provided as BMP files and the corresponding module label files, including a variety of personal writer's data as meta-information in XML format which is described in XML file and XML file format (DTD). Figure 4 shows the Entity-Relationship model in the design phase of the database.

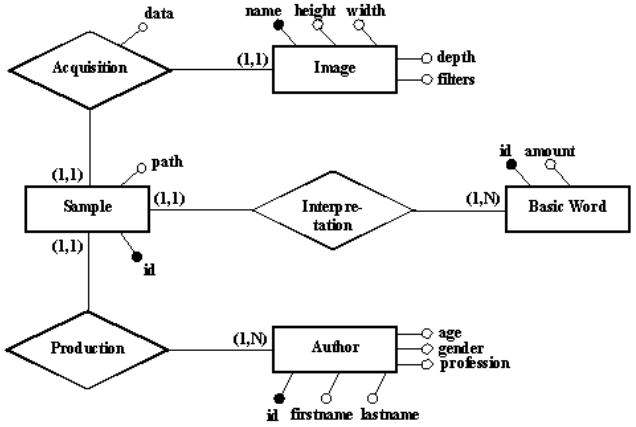


Figure 4. Entity-Relationship model.

Figure 5 shows the tree structure of folders where stored patterns. The dataset is partitioned into training and testing subsets.

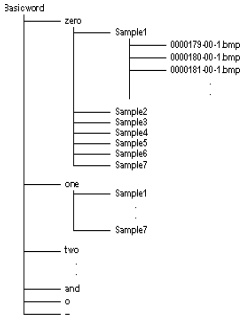


Figure 5. Tree structure of the folder.

V. Preprocessing

The pre-processing is to treat the raw data in order to achieve a form suitable for use in Pattern Recognition. The removal of the track in the background, rows and similar traits, it is often necessary when the word is extracted from the completed forms. A binarization is useful when the words are stored in image gray level. The next step is the normalization. The slope of the non-zero form (skew), or the slant of a word can be caused during the acquisition of the module or the writing styles of the writers. The results of these two of correction factors, which consist in the adjustment of the inclination in a basic word, the image is unchanged, hence the name of normalization. In general terms, the result of preprocessing must be an image containing the word to be recognized without any other disturbing element.

In this section is presented several pre-processing procedure accompanying database.

A. Binarization

Many algorithms handle only two-level (binary) images. Binarization (thresholding) is a major step in such algorithms to convert gray-scale images into binary images. In binarization algorithms, a threshold is usually computed first and then if a pixel has a higher intensity than that threshold, it is labeled as background; otherwise it is labeled as foreground (stroke). Due to the fact that the binarization is usually applied in the primary steps of a recognition problem, and its result greatly influences on the performance of the whole system, much attention is devoted to this task. Binarization is challenging for gray-level images with poor contrast, strong noise and variable modalities in histograms, and it is still a difficult problem in pattern recognition. The colour images of every acquisition form of basic words were binarized at a threshold fixed by using a binarization method at fixed threshold. It is to make use of a fixed threshold, which was set by default to S = 211, but can be changed, so the binarization is performed through the following transformation:

$$y(x) = \begin{cases} 0 & se\ x < S \\ 255 & se\ x \geq S \end{cases}.$$

The histogram of the image to binarize exhibits a distinctly bimodal, in practice there are two distinct peaks representing the background and the writing on the form of acquisition. In this case, the threshold is set at the minimum point between two peaks. The first of the two peaks coincides with the gray level at maximum value in the histogram, while the second

peak, however, does not necessarily coincide with the second largest value in the histogram. A pretty effective trick is to find the second peak, after multiplying the histogram values by the square of the distance from the first peak:

$$sm = \max\left[(k-j)^2 h(k)\right],$$

where 0 <= k <= 255, j is the gray level of the first peak and h(k) is the value of the histogram in each k. In this way, benefits the peaks that are not close to the maximum.

Another method of binarization based on the Otsu algorithm has been implemented [13].

B. Noise Reduction

The noise is the degradation of the image depends on several factors: the type of data acquisition, deterioration or wear of the paper. The reduction of noise in the image is intended to eliminate these irregularities that may have been introduced by the scanning operation, for example. The binarized image often produces spurious segments that have been removed by a 3x3 median filter. The median filter is a non-linear filter that provides in output the median value of the pixels around the pixel to be converted.

C. Skew Correction

The skew correction is very important. The skew of a document, in our case the acquisition module of basic words, is a distortion that is often introduced during scanning or photocopy of the document and it is often unavoidable. For the rotation of images was developed an application based on the algorithm of rotation, through successive inclinations, said shear algorithm. It consists in the inclination of the source image along the axis, a number of times, to produce a rotated image, proportionally and dimensionally identical to the original image. The algorithm, while fast, lacks the interpolation between pixels, so it was appropriate to change it.

D. Slant Correction

Slant is the deviation of average near-vertical strokes from the vertical direction. Slant correction is an attempt to reduce the range of variations of handwritten basic words. In handwritten words, the slant is due to the specific writing style. The slant is non-informative, also slanted words may considerably degrade the performance of the whole system, so this normalization step must be performed before of the feature extraction, training and recognition, to remove or reduce the slant influence as much as possible. The uniform slant correction techniques perform successfully when all near-vertical strokes have the same slant angle. In the handwritten basic words, the slant angle usually varies within each of them, and hence a uniform slant correction is not optimum. The method that we have implemented fits a minimum bounding parallelogram to each connected component of the binarized basic word image, such that top and bottom sides of each parallelogram are parallel to x-axis, each of the two oblique lines is determined by identifying pixels of the sample by scrolling through the image from left to right, then the slant angle is chosen as the median value of all parallelogram angles. Before the average slant angle is estimated and then a shear transformation in the horizontal direction is applied at the word image to correct its slant.

E. Smoothing

The image smoothing operation is to smooth the contour of the track of the handwritten word, removing all those pixels that are not significant for image morphological or probabilistic analysis, and eroding all those traits that do not belong to the track of the word until to their elimination.
The smoothing is obtained by iterated application of an operation of filtering is to change pixel based on the value of the neighborhood, which consists of all points adjacent to the observed pixels. Conceptually, the identity of a filter is based on the mask associated with it, that is, a grid of m x n pixels iteratively applied to each image point. The observed point is then modified in accordance with the configuration of the mask, and on the basis of some parameters, if any. For smoothing and noise reduction has implemented a filter based on the arithmetic mean, which is to replace the value of each pixel with the average of pixel values in the neighborhood, reducing the intensity variation of gray between a pixel and the other. It is usually used a mask of size 3 x 3 or 5 x 5. The result is the elimination of image points that are not representative of your surroundings.

F. Control Box

The control box is the smallest rectangle that contains the image of the handwritten character. It has implemented a procedure for trimming the control box. The image of the basic word is loaded into the matrix, then row x column is scanned in each pixel, the points of minimum, maximum, to the left and far right of the foreground are located. At this point, the sample is cut into these four points identified, and is downloaded into a new image file format standard with similar name.

VI. Data Markup

The personal information of the writers is integrated in the database to facilitate testing on groups of writers with different backgrounds such as age, occupation, gender, and education. This will help to investigate some characteristics of the writing styles by different groups of writers.

The database incorporates the operations provided by the mechanism of XML. It allows generating a text file in which to specify relevant characteristics of any conceptual entity, in this case the images of the basic words in the database provided. It was designed a Document Type Definition (DTD), i.e. a document that describes the tags that can be used in an XML document, their mutual relationship towards the structure of the document and other information about the attributes of each tag, in the other words a sort of header that allows to establish the criteria for the classification of elements and their attributes of interest, from which it generated the XML file. The tree structure of the same discriminates images first by author and then by the basic word; for each of them there are seven samples, which represent the lowest hierarchical level of the distribution of elements in the file. The information recorded in it, as

attributes of elements, including the full path of the stored images. The software bundle of the database includes predefined queries, made in language C, which deal with the search for certain classes of basic word exploiting the indexing mechanism just in XML. Each query thus returns the locations of images that meet the search criteria associated with it, as strings organized in data structures. In this way the utilizer in the final product will have the path by which date back to the image associated to use in their applications.

The directory hierarchy that describes the physical layout of the patterns has been translated into an XML document, which stores all relevant information regarding the images and their respective authors. See Figure 6.

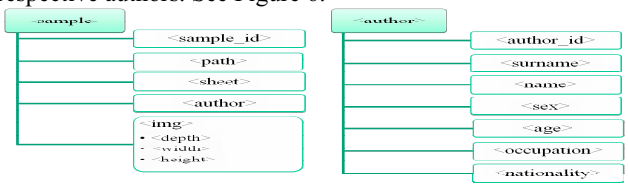


Figure 6. Data structure on images and writers.

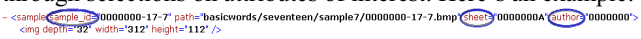
Here is presented a slice of the XML document:

```
<?xml version = "1.0"?>
<!-- A simple XML document -->
<data_set sub_set_name = "Training" samp_amount = "2">
  <basic_words>
    <basic_word word = "nineteen">
      <samples>
        <sample id = "3">
          <path>c:\IntSysDB\nineteen\sample3\0000001-19-3.bmp</path>
          <sheet>0000371B</sheet>
          <author>0000371</author>
        </sample>
      </samples>
    </basic_word>
    <basic_word word = "twenty">
      <samples>
        <sample id = "1">
          <path>c:\IntSysDB\twenty\sample1\0000001-20-1.bmp</path>
          <sheet>0000371B</sheet>
          <author>0000371</author>
        </sample>
      </samples>
    </basic_word>
  </basic_words>
</data_set>
```

The XML document is valid because it meets the structural requirements defined in the DTD file.

VII. Data Retrieval

The choice of data storage in the database Intelligent Systems II fell on XML with the specific goal of achieving optimal results in terms of reading and updating of stored information. The use of XML represents an innovation in the field of basic word collections: the classic approach of organization of data is limited to the simple arrangement of the same into directories that represent certain classes of membership of the entries of the databases. In this light, formal organization of the content is almost absent, and the characteristics of XML are ideally suited to provide an alternative solution to the issue. Firstly, XML is based on Unicode, an encoding system of texts independent from keyboard, and enables presence of all languages at the same time. This allows overcoming any frontier for the spread of the output file. In addition, the compactness of a simple text file, free from any complication due to indexing and processing of data, makes it the ideal tool for the distribution network of a data set. Finally, a critical factor in the choice of XML was the versatility for which it stands. It can produce code, in fact, to query the archive in any of the popular programming languages, thanks to the many tools for scripting. The tree structure of XML document created, allows everyone to make intelligent query, selecting specific subsets of specimens through selections on attributes of interest. Here's an example:



Similarly, the section devoted to the metadata of several writers, allows the extraction of pattern sets, produced by certain categories of individuals. Here's an example:

VIII. Conclusions and Future Work

A single word written by different authors, or by the same author at different times, produces a different result. A database in handwriting recognition should therefore be expanded, updated and amended over time, as each pattern is related to a human, and writers change over time and change their cultures and knowledge of the people. The SIDB database can be used to train and test handwritten legal amount recognizers and to perform writer identification and verification experiments. In order to expand the database, researchers interested to cooperate are invited to link and operate to the address: http://www.web-learning.uniba.it/html/scienze_mm_ff_nn_.html.

References

[1] I. Guyon, R. Haralick, J. Hull, and I. Phillips, *Database and benchmarking*, in Handbook of Character Recognition and Document Image Analysis, H. Bunke and P. Wand eds., World Scientific, pp. 779-799, 1997.

[2] J.J. Hull and R.K. Fenrich, *Large Database Organization for Document Images*, in Fundamentals in Handwriting Recognition, S. Impedovo ed., NATO ASI Series, Series F: Computer and System Sciences, Vol. 124, Springer-Verlag, Berlin, pp. 397-414, 1992.

[3] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, *Unipen project of on-line data exchange and recognizer benchmarks*, Proc. of the 12th ICPR, Jerusalem, Israel, pp. 29-33, October 1994.

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

[5] R. Wilkinson, et al., *The first census optical character recognition systems*, Conf. #NISTIR 4912, the U.S. Bureau of Census and the NIST, Gaithersburg, MD, 1992.

[6] J. Hull, *A database for handwritten text recognition research*, IEEE Trans. On PAMI, 16(5):550-554, May 1994.

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

[8] U.-V. Marti and H. Bunke, *A full English sentence database for off-line handwriting recognition*, Proc. of ICDAR '99, Bangalore, India, pp. 705-708, 1999.

[9] http://www.unipen.org/, UNIPEN Database.

[10] http://www.iam.unibe.ch/fki/databases, IAM Handwriting Database.

[11] *Automatic Bankcheck Processing*, Special issue of IJPRAI, S. Impedovo, H. Bunke and P.S.P. Wang eds., Vol. 28, WS, 1997.

[12] D.S. Lee and S.N. Srihari, *Handprinted digit recognition: A comparison of algorithms*, Proc. of the 3rd IWFHR, 25-27 May, Buffalo, NY, pp. 153-164, 1993.

[13] N. Otsu, *A threshold selection method from gray-level histogram*, IEEE Transactions on SMC, Vol. SMC-9, No. 1, pp. 62-66, 1979.

[14] G. Dimauro, S. Impedovo, R. Modugno, and G. Pirlo, *A New Database for Research on Bank-check Processing*, Proc. of the Eighth IWFHR, IEEE Computer Society Washington, DC, USA, pp. 524-528, 2002.

[15] A.W. Paeth, *A Fast Algorithm for General Raster Rotation*, in Graphics Gems, edited by Andrew Glassner, published by Academic Press, pp. 179-195, 1990.