A Benchmark Kannada Handwritten Document Dataset and its Segmentation

Alireza Alaei
Department of Studies in Computer Science, University of Mysore
Mysore, 570006, India
alireza20alaei@yahoo.com

P. Nagabhushan
Department of Studies in Computer Science, University of Mysore
Mysore, 570006, India
pnagabhushan@hotmail.com

Umapada Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata–108, India
umapada@isical.ac.in

***Abstract*—Research towards Indian handwritten document analysis achieved increasing attention in recent years. In pattern recognition and especially in handwritten document recognition, standard databases play vital roles for evaluating performances of algorithms and comparing results obtained by different groups of researchers. For Indian languages, there is a lack of standard database of handwritten texts to evaluate performance of different document recognition approaches and for comparison purpose. In this paper, an unconstrained Kannada handwritten text database (KHTD) is introduced. The KHTD contains 204 handwritten documents of four different categories written by 51 native speakers of Kannada. Total number of text-lines and words in the dataset are 4298 and 26115, respectively. In most of text-pages of the KHTD contains either an overlapping or a touching text-lines and the average number of text-lines in each document on the database is 21. Two types of ground truths based on pixels information and content information are generated for the database. Providing these two types of ground truths for the KHTD, it can be utilized in many areas of document image processing such as sentence recognition/understanding, text-line segmentation, word segmentation, word recognition, and character segmentation. To provide a framework for other researches, recent text-line segmentation results on this dataset are also reported. The KHTD is available for research purposes.**

Keywords: Handwritten document; Kannada handwritten recognition; Kannada handwritten dataset; Ground truth

I. INTRODUCTION

Many researchers have been working on the recognition of handwritten documents and many standard datasets have been developed to help researchers sharing their results on the same datasets and comparing performances of their classifiers [1–17]. For some scripts such as Roman, Chinese, and Korean many standard datasets are available in literature [1–9].

The NIST [1], MNIST [2], CENPARMI [3], CEDAR [4] are some examples of widely used English databases for numerals, characters, words and text-pages. The NIST standard dataset SD3 [1] contains one of the largest selections of isolated digits and characters. Altogether, it contains over 300000 character images that were extracted from data collection forms filled out by 2100 individuals. The images on SD3 were provided as a training set. A separate dataset (TD1) was developed for testing. The quality of the data varied more widely in TD1 than in SD3. A significant amount of running English text is also available in the NIST SD3 and TD1 datasets. Three other NIST datasets (SD11-SD13) contain examples of phrases. Altogether, 91500 handwritten phrases were scanned at 200 dpi in binary mode. The CENPARMI dataset contains 17000 isolated digits that were extracted from images of about 3400 postal ZIP Codes [3]. The isolated digits were extracted from the ZIP Codes by a manual segmentation. The CEDAR dataset as one of the available English dataset contains nearly 28000 isolated handwritten characters and digits that were extracted from images of US postal addresses [4]. Individual characters and digits were segmented from the address images by a semi-automatic process. Approximately 5000 ZIP Codes, 5000 city names, and 9000 state name images are included in [4]. Although there are many available databases for isolated characters and digits, to the best of our knowledge only a few benchmark datasets are available for handwritten text-lines [5–7]. Recently, a benchmark dataset of 300 text-pages have been made available for text-line and word segmentation purpose [5]. The ground truths for text-line and word segmentation have manually been annotated. This dataset has been used for ICDAR2009 segmentation contest. 100 document images and associated ground truths have been considered for training dataset and 200 images and associated ground truth have been used for testing dataset. The documents used in order to build the training and test datasets did not include any non-text elements and the documents have been written by 50 individuals in English, French, German and Greek languages [5]. Because of easy availability of different datasets in English, different types of research starting from numeral/character to sentence recognition using natural language processing [18] are available in literature.

In Persian and Arabic languages, there exists a few handwritten datasets for characters, numerals, words as well as text-lines [10–13]. The IFN/ENIT dataset [11] consists of 26459 images of the 937 city/town names in Tunisia, and it was written by 411 different writers. The images were partitioned into different sets so that researchers can use them for training and testing [11]. In [12] authors have described formations of a standard handwritten databases including isolated digits, letters, numerical strings, Legal amounts (used for cheques), and dates. It consists of totally 18000 digit samples out of which 11000, 2000 and 5000 are used for training, verification and testing, respectively. Total of 12000 character data were also included in the dataset out of which 7140, 1360 and 3400 samples are used for training,

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.37

verification and testing, correspondingly. A dataset of Persian texts are introduced in [13]. It can be used for different tasks of Persian text-line segmentation, word segmentation, character segmentation/recognition etc.

In Indian languages domain the situation is worse compared to the other languages in terms of availability of standard datasets although India is a multilingual, multi-script country. Recently, researchers have provided some datasets for off-line handwritten Bangla and Devanagari numerals and characters [14–16]. An on-line handwritten dataset for Tamil and Kannada words (OHR) was also provided [17]. The database presented in [14] contains 22556 Devanagari numeral samples written by 1,049 persons and of 23392 handwritten Bangla numeral samples written by 1106 people. In [15] numeral recognition of six Indian scripts are presented and this dataset is also available. However, to the best of our knowledge no benchmark dataset is available for text-pages of Indian languages.

From the literature, it can be noted that most of the current non-Indian languages databases are contained only handwritten isolated digits, characters or extracted words. Only a small number of the current available databases include sentences or text-pages. Lack of standard databases can be considered as one of the reasons of the slower development of OCRs for Indian languages [19, 20] rather than the Latin based languages. Considering these issues, we propose to develop a Kannada handwritten texts database (KHTD) containing 204 text-pages of four different categories of stories and general news, sentences of sports news, movie and medical news written by 51 Kannada native speakers. Two types of ground truths based on pixels information and content (line-wise, word-wise and character-wise) information are generated in the proposed database.

The organization of rest of the paper is as follows: In Section II we briefly illustrate characteristics of Kannada scripts and Section III describes the proposed KHTD dataset. Ground truth extraction and some preprocessing are described in Section IV. In Section V, we demonstrate the results of two state-of-the-art text-line segmentation methodologies on the KHTD. Finally, we present conclusion in Section VI.

II. Characteristics of Kannada Scripts

Kannada is one of the twenty-two official languages of India, spoken by approximately 50 million people mainly in the state of Karnataka and to a good extent in its neighboring states such as Andhra Pradesh, Maharashtra, Tamil Nadu, Kerala and Goa. It is the 27th most spoken language in the world. Kannada is the official and administrative language of the Karnataka state of India. The Kannada language is written using the Kannada script and it is derived from Brahmi script. The Kannada literatures produced by the end of the nineteenth century and later are classified as Modern Kannada.

The Kannada language uses forty-nine phonemic letters, divided into three groups: vowels (thirteen letters), consonants (thirty-four letters), and two other letters. Table I shows Kannada vowels. The number of written symbols is far more than the forty-nine characters in the alphabet, because different consonants and vowels can be combined to form composite characters. These special characteristics and intrinsic features of Kannada scripts make the handwriting recognition of this language a challenging problem [20].

TABLE I. KANNADA VOWELS WITH ITS UNICODE VALUES

Kannada vowels	Unicode Hex-Value	Phenomena
ಅ	0C85	A
ಆ	0C86	AA
ಇ	0C87	I
ಈ	0C88	II
ಉ	0C89	U
ಊ	0C8A	UU
ಋ	0C8B	vocalic R
ಎ	0C8E	E
ಏ	0C8F	EE
ಐ	0C90	AI
ಒ	0C92	O
ಓ	0C93	OO
ಔ	0C94	AU

III. Overview of the KHTD

To start with, four different text categories: stories and general news, sentences of sports news, movie and medical texts of Kannada were considered. For data collection the texts are given to 51 individuals with different ages and educational backgrounds and asked them to write each of the given texts in a separate unruled A4 sheet without any stress and constrained. Participants have written the given text-pages by different types of pens but no restrictions were imposed on them for choosing the instrument. The handwritten texts are then scanned in gray-scales with the resolution of 300 dpi using a flat bed scanner. Using Otsu's method [21] binary images of the scanned documents are also provided. We have named and stored each binary file based on a sequence starting with the word "Kannada" following by the symbol "_" and continuing by a digit of 1, 2, 3 or 4 which shows the type of the document and again the symbol "_" and finally three-digit number starting from "001". Therefore, the first image of the first category of Kannada text database is named as "kannada_1_001". The images are stored in "TIF" format so the first image file of the database has the name of "kannada_1_001.TIF". A scanned gray-level image of handwritten Kannada text is shown in Figure 1. Since, we considered handwritten of 51 individuals, 204 text-pages of Kannada handwritten are gathered (each individual has written one page for each of the four categories of the documents).

Some statistics of the KHTD are tabulated in Table II. From Table 2, it can be noted that the KHTD contains 204 pages of Kannada handwritten text, and in the KHTD, the number of text-lines varies from 6 to 29 in the text-pages of the KHTD. The KHTD contains totally 4298 Kannada handwritten text-lines and 26115 words. On an average, each text-page includes 21.1 text-lines, 128.01 words, and 934.27

characters. In most of the text-pages of the KHTD either an overlapping or a touching text-lines exist. The KHTD can be used for Kannada text-line segmentation, sentence recognition/ understanding, word segmentation/recognition, segmentation the words into characters, word spotting, text-line extraction, and writer identification etc.

TABLE II. SOME STATISTICS OF THE KHTD

	Number of Text-lines	Number of words
Total	4298	26115
Average	21.1	128.01

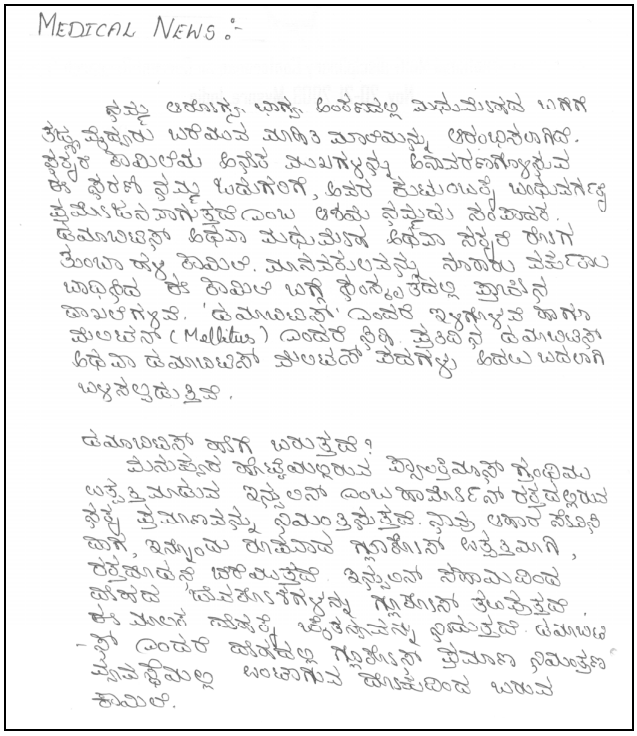

Figure 1. A scanned gray level image of handwritten Kannada text is shown here as sample.

IV. GROUND TRUTH

For automatic evaluation of results of any segmentation/recognition system, the ground truth (GT) information plays a significant role. However, it is an error-prone and time-consuming task. There are two types of GT information in the literature [5, 12]. In the first type, the GT is computed by labeling pixels of a text-page with respect to their belongingness in the background or foreground (text-lines or words). This GT is useful for text-line and word segmentations. The second type of the GT is generated by creating a text-file using Unicode for every character present in each text-page image with respect to its content. This GT is useful to evaluate the recognition results. To compute the first type of GT for a text-page, first we manually removed the noise from each text-page (binary document image). Then we draw manually some separating lines to segment individual text-lines. Manually drawn separating lines are shown in Figure 2. The background pixels of the binary document image will get values of zero and the pixels belonging to each segmented text-line will get a unique value greater than 0. For example, if we have 10 text-lines in a text-page, all the pixels belonging to the foreground part of the first text-line get value of 1, all the foreground pixels belonging to the second text-line get value of 2 and so on. Final GT information of the text-page shown in Figure 1 is given in Figure 3. As a result, we create a set of GT files and the number of GT files is equal to the number of text-pages presented in the proposed database. The GT files have the same name as the original files had and only an additional file extension of ".DAT" is added to the original filename. For example for the text image of "kannada_1_001.TIF" the GT filename will be "kannada_1_001.TIF.DAT".

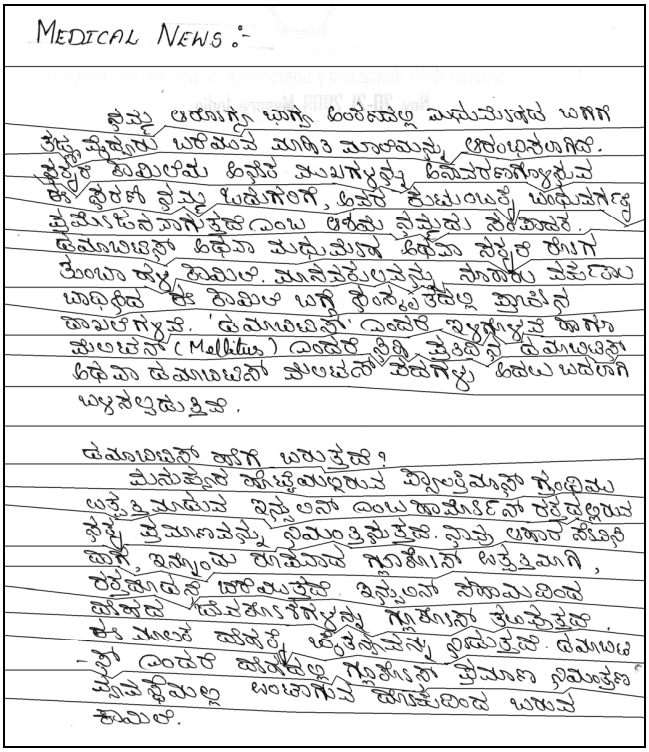

Figure 2. Manual segmentation of text-lines in the handwritten Kannada document (shown in Figure 1) for ground truth generation

The second type of the GT is generated by creating a text-file using Unicode for each text-page image with respect to its content. The text-file contains the same material in the text-image and also same content in each text-line and word. To generate a content-based GT file of a document, text version of that document (text file) is manually generated and considered as content-based ground truth. The file name for each content-based GT is chosen by considering only the name (without extension) of the original text-page image with "TXT" as extension. Therefore, for the "kannada_1_001.TIF" image the corresponding content-

based ground truth file "kannada_1_001.TXT" is obtained. The content of ground truth information of the handwritten document of Figure 1 is shown in Figure 4. As a result, two GT files (one for pixel-wise and the other one for content-based text) for each document image are generated in the proposed database that can be used by different evaluation strategies.

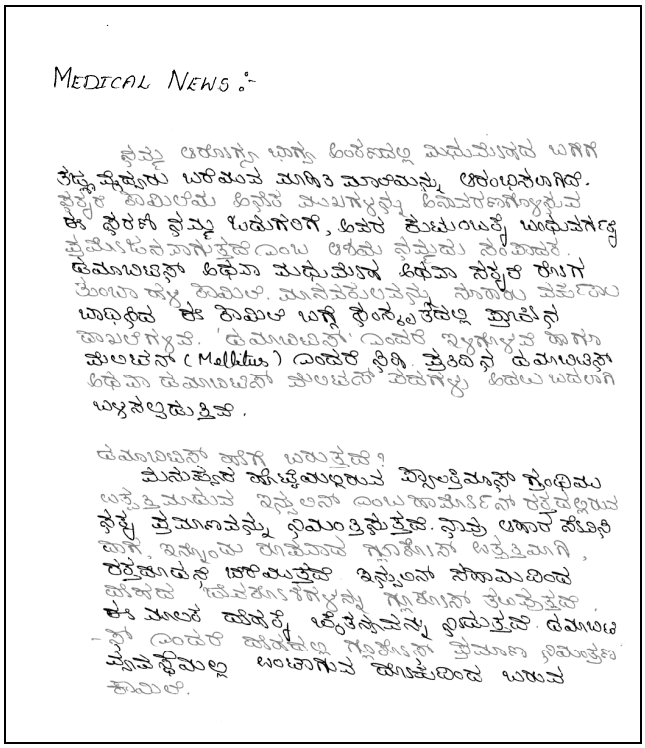

Figure 3. Pixel based GT information of the Kannada handwritten text-page shown in Figure 1

V. TEXT-LINE SEGMENTATION RESULTS OF THE KHTD

To provide a platform for other researchers regarding the comparison of their results of text-line segmentation techniques on the KHTD, we tested 2 different text-line segmentation techniques discussed in [5] and [22] on the KHTD.

In the text-line segmentation technique presented in [22], initially the input document is vertically decomposed into parallel pipe structures called stripes. Each row of a stripe is painted by a gray intensity, which is the average intensity value of gray values of all pixels present in that row-stripe. The painted stripes are then converted into two-tone and using some smoothing operations, the two-tone painted image is smoothed. A dilation operation is employed on the foreground portion of the smoothed image to obtain a single component for each text-line. Thinning of the background portion of the dilated image and subsequently some trimming operations are performed to obtain a number of separating lines, called candidate line separators. Using the starting and ending points of the candidate line separators and analyzing the distances among them, related candidate line separators are connected to obtain segmented text-lines [22]. In the PPSL technique [5] like the technique, proposed in [22], document is vertically crumbled into few stripes. In order to get Potential Piece-wise Separation Line (PPSL) between two consecutive text-lines, the white/black spaces in each stripe are analyzed. Next, such PPSLs are concatenated or extended in both the left and right directions to construct the complete segmentation lines based on distance analysis of each PPSL with left and right neighboring PPSLs.

MedicalNews :-

ನಮ್ಮ ಆರೋಗ್ಯ ಭಾಗ್ಯ ಅಂಕಣದಲ್ಲಿ ಮಧುಮೇಹದ ಬಗೆಗೆ
ತಜ್ಞ ವೈದ್ಯರು ಬರೆಯುವ ಮಾಹಿತಿ ಮಾಲೆಯನ್ನು ಆರಂಭಿಸಲಾಗಿದೆ.
ಸಕ್ಕರೆಯ ಕಾಯಿಲೆಯ ಅನೇಕ ಮುಖಗಳನ್ನು ಅನಾವರಣಗೊಳಿಸುವ
ಈ ಸರಣಿ ನಮ್ಮ ಓದುಗರಿಗೆ, ಅವರ ಕುಟುಂಬಕ್ಕೆ, ಬಂಧುವರ್ಗಕ್ಕೆ
ಪ್ರಯೋಜನವಾಗುತ್ತದೆ ಎಂಬ ಆಶಯ ನಮ್ಮದು- ಸಂಪಾದಕ
ಡಯಾಬಿಟಿಸ್ ಅಥವಾ ಮಧುಮೇಹ ಅಥವಾ ಸಕ್ಕರೆ ರೋಗ
ತುಂಬಾ ಹಳೆ ಕಾಯಿಲೆ. ಮಾನವಕುಲವನ್ನು ನೂರಾರು ವರ್ಷಕಾಲ
ಬಾಧಿಸಿದ ಈ ಕಾಯಿಲೆ ಬಗ್ಗೆ ಸಂಸ್ಕೃತದಲ್ಲಿ ಪ್ರಾಚೀನ
ದಾಖಲೆಗಳಿವೆ. 'ಡಯಾಬಿಟಿಸ್' ಎಂದರೆ ಇಳಿಗೊಳವೆ ಹಾಗೂ
ಮೆಲಿಟಸ್ (Mellitus) ಎಂದರೆ ಸಿಹಿ. ಪ್ರತಿದಿನ ಡಯಾಬಿಟಿಸ್
ಅಥವಾ ಡಯಾಬಿಟಿಸ್ ಮೆಲಿಟಸ್ ಪದಗಳು ಅದಲು ಬದಲಾಗಿ
ಬಳಸಲ್ಪಡುತ್ತಿವೆ.
ಡಯಾಬಿಟಿಸ್ ಹೇಗೆ ಬರುತ್ತದೆ?
ಮನುಷ್ಯರ ಹೊಟ್ಟೆಯಲ್ಲಿರುವ ಪ್ಯಾಂಕ್ರಿಯಾಸ್ ಗ್ರಂಥಿಯು
ಉತ್ಪತ್ತಿ ಮಾಡುವ ಇನ್ಸುಲಿನ್ ಎಂಬ ಹಾರ್ಮೋನ್ ರಕ್ತದಲ್ಲಿರುವ
ಸಕ್ಕರೆ ಪ್ರಮಾಣವನ್ನು ನಿಯಂತ್ರಿಸುತ್ತದೆ. ನಾವು ಆಹಾರ ಸೇವಿಸಿ
ದಾಗ, ಇನ್ನೊಂದು ರೂಪವಾದ ಗ್ಲೂಕೋಸ್ ಉತ್ಪತ್ತಿಯಾಗಿ,
ರಕ್ತದೊಡನೆ ಬೆರೆಯುತ್ತದೆ. ಇನ್ಸುಲಿನ್ ಸಹಾಯದಿಂದ
ದೇಹದ ಜೀವಕೋಶಗಳನ್ನು ಗ್ಲೂಕೋಸ್ ತಲುಪುತ್ತದೆ.
ಈ ಮೂಲಕ ದೇಹಕ್ಕೆ ಚೈತನ್ಯವನ್ನು ನೀಡುತ್ತದೆ. ಡಯಾಬಿಟಿ
-ಸ್ ಎಂದರೆ ದೇಹದಲ್ಲಿ ಗ್ಲೂಕೋಸ್ ಪ್ರಮಾಣ ನಿಯಂತ್ರಣ
ವ್ಯವಸ್ಥೆಯಲ್ಲಿ ಉಂಟಾಗುವ ದೋಷದಿಂದ ಬರುವ
ಕಾಯಿಲೆ.

Figure 4. GT information based on text content of the Kannada handwritten text-page shown in Figure 1 (text is saved in UNICODE)

The same evaluation technique and parameters used in [5, 22] were considered for our experimentation. A table called as MatchScore is created and its values are calculated based on the intersection of the black pixel sets of the result and the ground truth. Detection (Dec.) rate and recognition (Rec.) accuracy are calculated as follows.

$$Det = \frac{one2one}{N}, \qquad \text{Re}c = \frac{one2one}{M} \tag{1}$$

where one2one is a full match between a pair if the matching score for this pair is more than or equal to an acceptance threshold. The value of one2one is worked out from MatchScore table. Parameter N is defined as the number of ground truth text-lines in each text-page, M is the number of separating text-lines obtained from the segmentation

algorithm. More details about the evaluation technique can be found in [5, 22]

From the experimentation, 94.55% and 96.12% were obtained for detection rate and recognition accuracy, respectively from the technique proposed in [22]. Overall, 95.32% of text-lines in the KHTD were accurately segmented using the technique presented in [22].

While testing the PPSL technique [5] on the KHDT, the results of 95.73% for detection rate, 94.25% for recognition accuracy and overall accuracy of 94.98% were obtained. The text-line segmentation results obtained from experimentations on KHTD using the techniques presented in [5] and [22] are briefly tabulated in Table III.

TABLE III. TEXT-LINE SEGMENTATION RESULTS OF KHTD

Result / Algorithm	Det. Rate (%)	Rec. Acc. (%)	Overall Segmentation (%)
PPSL [5]	95.73	94.25	94.98
Alaei et al. [22]	94.55	96.12	95.32

VI. CONCLUSION

A database of handwritten text-pages of Kannada is introduced. It contains 204 handwritten text-pages of Kannada. The KHTD has also two different ground truths based on pixels and text-content information. This characteristic of the KHTD provides two different evaluation strategies for the researchers to evaluate the performances of their algorithms. This database can be used in many areas of document image processing such as sentence recognition/understanding, text-line segmentation, word segmentation/ recognition, segmentation of the words into characters, word spotting, writer identification, discrimination between handwritten and machine printed texts of Kannada etc. The KHTD is available freely to the researchers by contacting to the authors.

ACKNOWLEDGMENT

The authors would like to thank Mrs. Vijaya Kumari, Mrs. Noor Sara, Mrs. Archana and Mrs. F. Alaei of the Department of Studies in Computer Science, University of Mysore, Mysore, India who has helped us for collecting the documents and preparing the ground truths. The authors would also like to thank all writers who contributed for this dataset.

REFERENCES

[1] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, 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.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffiner, "Gradient based learning applied to document recognition", Proc. of IEEE, 86(11), 1998, pp. 2278–2324.

[3] C.Y. Suen, C. Nadal, R. Legault, T. Mai, L. Lam, "Computer recognition of unconstrained handwritten numerals," Proc. of the IEEE, 80(7), 1992, pp. 1162–1180.

[4] J. Hull, "A database for handwritten text recognition research," IEEE Trans. on PAMI, 16(5), 1994, pp. 550–554.

[5] B. Gatos, N. Stamatopoulos, G. Louloudis, "ICDAR 2009 Handwriting Segmentation Contest," Proc. of 10th ICDAR, 2009, pp. 1393–1397.

[6] U. Marti, H. Bunke. "A full English sentence database for off-line handwriting recognition", Proc. of the 5th ICDAR, 1999, pp. 705–708.

[7] T. M. Rath, R. Manmatha, "Word Spotting for Historical Documents," International Journal on Document Analysis and Recognition (IJDAR), 2005, pp. 139 – 152.

[8] D. Kim, Y. Hwang, S. Park, E. Kim, S. Paek, S. Bang, "Handwritten Korean character image database PE92," Proc. of 2th ICDAR, 1993, pp. 470–473.

[9] [T.-H. Su, T.-W. Zhang, D.J. Guan, "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text," IJDAR, 10(1), 2007, pp. 27–38.

[10] Y. Al-Ohali, M. Cheriet, C.Y. Suen, "Databases for recognition of handwritten Arabic cheques", Pattern Recognition, 36(1), 2003, pp. 111–121.

[11] M. Pechwitz, S.S. Maddouri, V. Maergner, N. Ellouze, H. Amiri "IFN/ENIT - database of handwritten Arabic words", The 7th Colloque International Francophone surl'Ecrit et le Document, 2002, pp.129–136

[12] F. Solimanpour, J. Sadri, Suen, Y. Ching, "Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi Language," Proc. of 10th IWFHR, 2006, pp. 3–7.

[13] M. Ziaratban, K. Faez, F. Bagheri, "FHT: An Unconstraint Farsi Handwritten Text Database," Proc. of 10th ICDAR, 2009, pp. 281–285.

[14] U. Bhattacharya, B.B. Chaudhuri, "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals", IEEE Trans. on PAMI, 31(3), 2009, pp. 444–457.

[15] U. Pal, T. Wakabayashi, N. Sharma, F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", Proc. of 9th ICDAR, 2007, pp. 749–753.

[16] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Off-Line Handwritten Character Recognition of Devnagari Script", Proc. of 9th ICDAR, 2007, pp. 496–500.

[17] B. Nethravathi, C. P. Archana, K. Shashikiran, A. G. Ramakrishnan, V. Vijay Kumar, "Creation of a Huge Annotated Database for Tamil and Kannada OHR", Proc. of 12th ICFHR, 2010, pp. 415–420.

[18] R. Bertolami, H. Bunke, "Hidden Markov model based ensemble methods for offline handwritten text line recognition", Pattern Recognition, 41(11), 2008, pp. 3452–3460.

[19] B. B. Chaudhuri, U. Pal. A complete printed Bangla OCR system. Pattern Recognition, 31(5), 1998, pp. 531–549.

[20] U. Pal and B. B. Chaudhuri, "Indian Script Character Recognition: A Survey", Pattern Recognition, Vol. 37, 2004, pp. 1887–1899.

[21] N. Otsu, A threshold selection method from gray-level histograms, IEEE Transactions on System, Man. and Cybernetics 9(1), 1979, pp.62–69.

[22] A. Alaei, U. Pal, P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation", Pattern Recognition, 44 (4), 2011, pp. 917–928.