Individuality of Bangla Numerals

Chayan Halder
Department of Computer Science,
West Bengal State University,
Barasat, Kolkata 700126, WB,
India
Email: chayan.halderz@gmail.com

Jaya Paul
Department of Information
Technology, Govt. College of
Leather Technology, Kolkata
700098, WB, India
Email: jayapl2005@gmail.com

Kaushik Roy
Department of Computer Science,
West Bengal State University,
Barasat, Kolkata 700126, WB,
India
Email: kaushik.mrg@gmail.com

***Abstract*- This paper presents analysis of individuality of handwritten Bangla numerals. It has a great prospect in Writer Identification, Writer Verification, Forensic Science etc. After collecting and extracting characters from filled in forms, 400 dimensional feature vectors is computed based on gradient of the images. A total of 450 documents were used for this work. In our experiment we have used LIBLINEAR classifier of WEKA environment. We have computed and analyzed the Individuality of each numeral and observed that the numeral 5 has the most individuality property than other numerals and 0 has the least. We have also done the writer identification with all the numerals and obtained 96.5% accuracy with all writers.**

Keywords*- *Individuality of Handwriting; Writer Identification; Bangla Handwriting Analysis; WEKA.

I. INTRODUCTION

It's been a decade or two that writer Identification and verification is an active area of research in document level analysis. It has such potential that it can be used in forensic science, banking, graphology etc. As the handwritten numerals carries additional information about the personality and characteristics of the writer compared to electronic or printed numerals, there exists a high possibility to authenticate and identify the writer. Writer identification rests on the hypothesis that there exists a certain degree of stability in the writing style of an individual which makes it possible to identify the writer from his/her handwritten numerals. Our objective is to verify this hypothesis using our experiment, testing and results. The task of writer Identification focuses on extracting the characteristic attributes like character shape, size, and slat angle etc. from Bangla handwriting which can be done by averaging out the variation in handwriting of different individuals. These attributes are used by the expert document analyzers to quantitatively establish individuality of a writing style. Here we have tried to use the same mechanism for our work.

There exist various pieces of works in literature on writer identification/verification of non-Indic scripts [1-9]. Said et al. [2] have used 1000 test documents from 40 writers to develop a text independent writer identification system. They have used a texture analysis based approach and achieved an accuracy of 96.0%. Marti et al. [3] have used 100 pages of text written by 20 writers as the data for their experiment. They have computed twelve features based on visible characteristics of the writings. K-nearest-neighbour classifier and Feed forward Neural network are being used for their experiment. They have achieved Identification accuracy of 87.8% and 90.7% respectively. Srihari et al. [4] have used 1568 writers for their database and each writer is being asked to copy out the sample document three times. They have extracted Macro and Micro features from total text document, paragraphs, separated words and even from characters. An accuracy of 98% has been achieved by them. Schomaker et al. [5] have used 500 documents from 250 writers for their experiment. The edge-based directional features are being used for identification procedure. Siddiqi et al. [8] have used 50 documents from same nuber of writers for their work. They have used a local approach, based on the extraction of characteristics that are specific to a writer. Bayesian classifier has been used in their work and an identification accuracy of 94% has been achieved.

Some works are also available in Indic script like the works of U. Pal et al. on Oriya script [10]; B. Chanda et al. on Telugu script [11]; Garain et al. [12], U.pal et al [13] and A. K. Das et al. [14] on Bangla script.

U.Pal et al. [10] have used the directional chain-code and curvature feature for their work. SVM classifier has been used for the work and they have achieved an accuracy of 94% on writer identification. In the work of Garain et al. [12] on Bangla characters, they have used 60 documents from 20 writers. Gradient based contour encoding feature and 192 bit feature vector have been used in this respect. The K-means clustering has been used to get an accuracy of 40% for identification. U.pal et al. have worked on the database of 204 documents from 102 writers for text independent writer identification using the Bangla characters [13]. They have used 400 dimensional gradient features with SVM classifier to achieve an accuracy of 95.19%. In the work of A. K. Das et al. [14] they have used their own Database consisting of 55 writers. Each writer has four sample documents on two different topics. Radon transform projection profile has been used as the feature for their work.

978-1-4673-5119-5/12/$31.00 ©2012 IEEE

To the best of our knowledge there exists no work in literature on the Individuality of Bangla numerals. Only the work of Garain et al. [12]; that is on individuality of handwriting based on Bangla characters is comparable with our work, if only the individuality of numerals is considered. Here for our work the 400 dimensional feature based on gradient has been used.

The subsequent part of the paper is organized as follows: section 2 describes the properties of Bangla script. In section 3 data collection is described. Pre-processing strategies and feature extraction is discussed in section 4 and 5 respectively. Section 6 describes about WEKA tool followed by results in section 7. At last we concluded in the section 8.

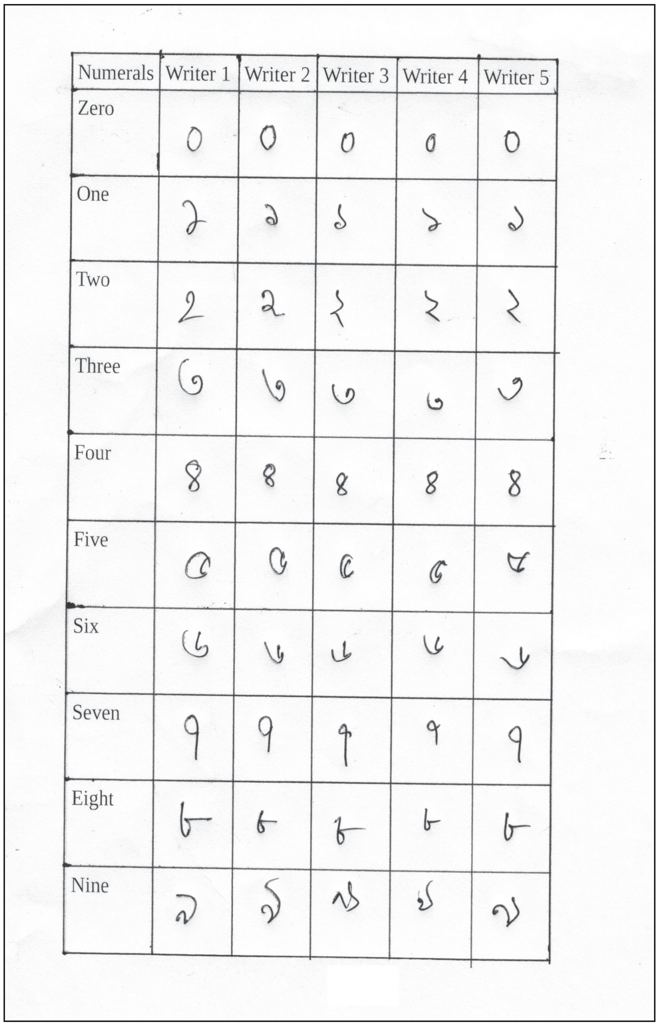
II. PROPERTIES OF BANGLA SCRIPT

Bangla, the second most popular language in India and the sixth most popular language in the world, is an ancient Indo-Aryans language. Bangla script alphabet is used in texts of Bangla, Assamese and Manipuri languages. Bangla is also the national language of Bangladesh. Also Bangla is the official language of the state West Bengal in India [16]. To get an idea of Bangla numerals and their variability in handwriting, five sets of handwritten Bangla numerals of different writers are shown in Fig.1 (a).

Each person writes differently from other and each person write differently from himself/herself but intuitively, the intra-writer variation (the variation within a person's handwriting samples) is quite less compared to the inter-writer variation (the variation between the handwriting samples of two different people). There are two main point of concern while comparing handwritings: the variability of the handwriting of the same individual and the variability of the handwriting from one individual to another. These two variations can be seen when several individuals are asked to write the same numerals many times (in our work it is 5 times). For example if we consider Fig 1(a) which exhibits the numeral 5 and 0 of different writers, where the variation is much more in between same numerals than in Fig 1(b) when a single writer writes the numerals. From the figure it can be noted that the variation in writing for numeral 5 is more than the numeral 0. It also describes that for every writer there exist a certain amount of uniqueness for every character they write. For this reason using this uniqueness of each numeral the identification of writer can be possible. This means each numeral with its uniqueness can able to identify a writer with certain amount of accuracy. For this very reason we worked on the Individuality of Bangla numerals.

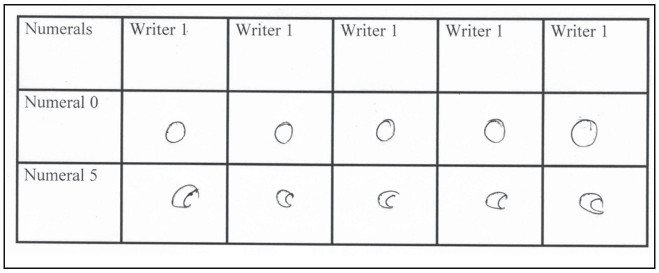
III. DATA COLLECTION

As we are interested in computing individuality of handwriting, we have designed a sample document consisting of all Bangla alphabets (vowel and consonants), numerals and vowel modifiers. A total number of 120 writers, predominantly students, were asked to copy-out the printed characters in the particular box area of the sample document form. Total 5 documents are being given to each writer for data collection. Most of the writers are in between age group 17-25. We also have writers in between age group 30-45 and even 50-60. Most of the writers are right handed. Out of 120 writers until now we have managed to collect full 5 sets of data from 90 writers (used for current work purpose) and for remaining writers till now we have collected fewer numbers of sets. Each set contains 10 Bangla numerals and 51 Bangla alphabets and 10 Bangla vowel modifiers. We have a total of 26367 Bangla alphabets, 5170 numerals and 5170 vowel modifiers. An example of our designed character sample collection document form is shown in Fig 2. There exists no boundary for writers regarding the type of pen and ink they use. We scanned these documents using a flatbed scanner for digitization. The images are in gray tone and digitized at 300/600 dpi and stored in Tagged Information File Format (TIFF). In this work we have used only the numerals for the writer identification and individuality. For our current work we have used the 300 dpi gray toned TIFF format image.



Numerals	Writer 1	Writer 2	Writer 3	Writer 4	Writer 5
Zero	০	০	০	০	০
One	১	১	১	১	১
Two	২	২	২	২	২
Three	৩	৩	৩	৩	৩
Four	৪	৪	৪	৪	৪
Five	৫	৫	৫	৫	৫
Six	৬	৬	৬	৬	৬
Seven	৭	৭	৭	৭	৭
Eight	৮	৮	৮	৮	৮
Nine	৯	৯	৯	৯	৯

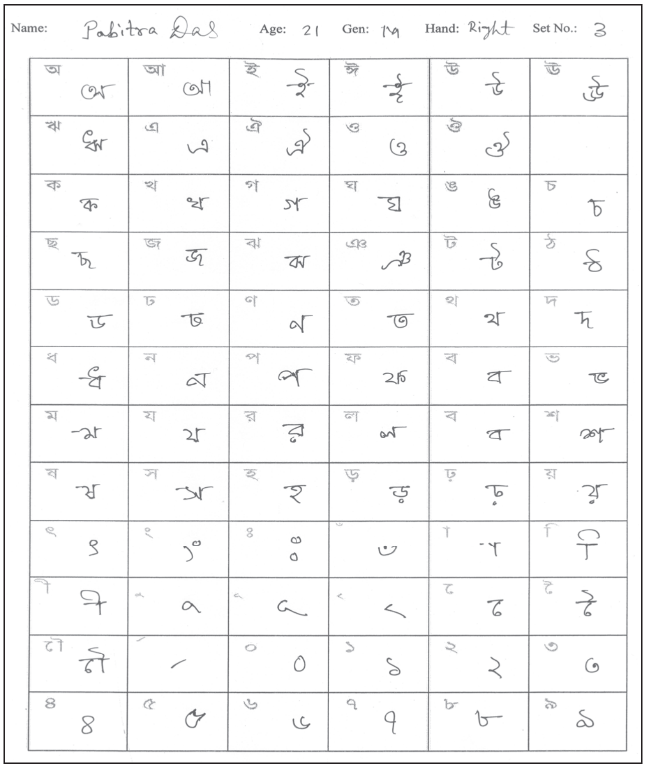
(a)



Numerals	Writer 1	Writer 1	Writer 1	Writer 1	Writer 1
Numeral 0	০	০	০	০	০
Numeral 5	৫	৫	৫	৫	৫

(b)

Fig. 1. (a) Sample of all Bangla handwritten numerals of 5 different writers, (b) Sample of numerals 0 and 5 of the same writer.



Name: Pabitra Das Age: 21 Gen: M Hand: Right Set No.: 3

অ অ	আ আ	ই ই	ঈ ঈ	উ উ	ঊ ঊ
ঋ ঋ	এ এ	ঐ ঐ	ও ও	ঔ ঔ	
ক ক	খ খ	গ গ	ঘ ঘ	ঙ ঙ	চ চ
ছ ছ	জ জ	ঝ ঝ	ঞ ঞ	ট ট	ঠ ঠ
ড ড	ঢ ঢ	ণ ণ	ত ত	থ থ	দ দ
ধ ধ	ন ন	প প	ফ ফ	ব ব	ভ ভ
ম ম	য য	র র	ল ল	ব ব	শ শ
ষ ষ	স স	হ হ	ড় ড়	ঢ় ঢ়	য় য়
ৎ ৎ	ং ং	ঃ ঃ	ঁ ঁ	া া	ি ি
ী ী	ু ু	ূ ূ	ৃ ৃ	ে ে	ৈ ৈ
ো ো	্ ্	০ ০	১ ১	২ ২	৩ ৩
৪ ৪	৫ ৫	৬ ৬	৭ ৭	৮ ৮	৯ ৯

Fig. 2. Sample document form for collection of Bangla Handwritten Alphabet, Numerals, Vowel modifires

IV. PRE-PROCESSING

The digitized images have been binarized using a global binirization method [16]. Then a character extraction technique has been used to extract the characters from the digitized form and they are being stored in gray mode.

A. Character extraction Technique

This technique is used for extraction of each individual character from the document form of handwritten characters. The steps are:

Firstly, the global binirization of the whole document has been carried out. Then maximum run length has been computed on horizontal and vertical histogram of that document form. Using the maximum run length of horizontal and vertical histogram, we have identified the horizontal lines and vertical lines of the document form. After the identification of vertical and horizontal lines we have deleted those lines from the document form image to get an image which contains only the suggestive characters and the original handwritten characters. Then using the horizontal and vertical line information we have calculated the top corner point values of each block and then we have removed the suggestive characters. After this, bounded box for each handwritten character has been calculated and these information have been stored for further processing.

V. FEATURE EXRACTION

In this work 400 dimensional feature extraction technique [16] is used. We have used this feature extraction technique as it has given some encouraging results for different Indic script numeral recognition for the work of U. Pal et al. [17]. To obtain 400 dimensional features we have applied the following steps:

STEP 1 At first the binirization of the input gray image is done.

STEP 2 The normalization of the binary image is done. Here we normalize the image into 73 x 73 pixels.

STEP 3 The binary image is then converted into a gray-scale image applying a 2 × 2 mean filtering 5 times.

STEP 4 The gray-scale image is normalized so that the mean gray scale becomes zero with maximum value 1.

STEP 5 Normalized image is then segmented into 9×9 blocks.

STEP 6 A Roberts filter is then applied on the image to obtain gradient image. The arc tangent of the gradient (strength of gradient) is quantised into 16 directions and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient (f(x, y)) we mean

$$f(x,y)= \sqrt{(\Delta u)^2 + (\Delta v)^2} \text{ and}$$

by direction of gradient (θ(x,y)) we mean

$$\theta(x,y)= \tan^{-1} \Delta v/\Delta u, \text{ where}$$

$$\Delta u = g(x+1, y+1) - g(x, y) \text{ , and}$$

$$\Delta v = g(x+1, y) - g(x, y+1) \text{ , and}$$

g(x,y) is a gray scale point (x,y).

Step 7 Histograms of the values of 16 quantized directions are computed in each of 9 x 9 blocks.

Step 8 9×9 blocks is down sampled into 5×5 by a Gaussian filter. Thus, we get 5×5×16 = 400 dimensional feature.

VI. WEKA

WEKA is one of the popularly used tools for machine learning algorithms for data mining tasks [15]. We have chosen the WEKA tool for our classification purpose as it is one of the most widely used open source classification and data mining tools available. In WEKA we can use different classifiers actively used in pattern recognition like SVM, MLP, KNN etc. simultaneously and easily using the same data set. The algorithms can be called from own Java code or using the weka.jar file of the package. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. For our work purpose we have used the LIBLINEAR classifier. LIBLINEAR is a good linear classifier for data with large number of instances or features. The main features of LIBLINEAR are:

- Same data format as LIBSVM, our general-purpose SVM solver, and also similar usage.
- Multi-class classification: 1) one-vs.-the rest, 2) Crammer & Singer.
- Cross validation for model selection.
- Probability estimates (logistic regression only).
- Weights for unbalanced data.

For some large data without using kernels, one can quickly train a much larger set via a linear classifier. This is one of the reason we have used LIBLINEAR here. Another reason is, it converged faster for our dataset than other classifiers of WEKA and it has given some encouraging results for our work. We have used the L2-Loss Support Vector Machine (dual) as the SVM Type parameter of the LIBLINEAR. Both the Bias and Cost parameters are 1.0. The EPS (the tolerance of the termination criterion) is 0.01.

VII. Results

The present work has been carried out on 4500 numerals from 450 documents written by 90 writers. We have used 5-fold cross validation scheme for computing individuality of numerals and writer identification. We have computed the individuality of each numeral for all the writers. Then writer identification has been computed on all numerals. The results are discussed below:

A. *Result on Individuality of Numerals*

We have computed the results of individuality on numerals that are described in Fig 3. From the figure we can observe that the numeral 5 is most individual followed by numerals 6, 2 and 9 where as numeral 0 has the least individuality. The corresponding values of individuality for 5 and 0 are 35.90% and 16.32%. Analyzing all the numerals that are shown in Fig 1(a), it can be said that numeral 0 has more similarity among different writers. But if we consider numeral 5 then the variation in writing is much higher among the writers. This is because the starting stroke angle, shape, writing pattern of numeral 5 is different among the writers and also the writing complexity of the numeral itself is higher. These are some of the reasons that numeral 5 has the most individuality.

B. *Result of Writer Identification*

We have used all numerals for writer identification and an accuracy of 96.5% has been achieved. We have also experimented by varying the number of writer sets in this purpose. Table 1 shows the details of different identification results for different writer sets. We have achieved an accuracy of 100% upto 20 writers for writer identification. The accuracy has been dropped to 98% for 40 writers. Strangely the drop rate in the accuracy has been quite higher for 50 writers and this rate continues upto 65 writers. From 70 writers the accuracy has been increased afterward. We have tried and tested several times but the results that we have got remained unchanged, which we are not able to justify for the time being.

C. *Comparison of Results*

The work of Garain et al. [12] on 60 documents from 20 writers; that is on individuality of handwriting based on Bangla characters can be compared with our work, if we consider only numerals individuality. In their work the percentage of individuality for numeral 4 is around 5.60% for numeral 5 it is around 11.9% and for 7 it is around 18.5% where for those numerals in our work the corresponding accuracies are 25.17%, 35.90% and for 7 it is 25.41% respectively.

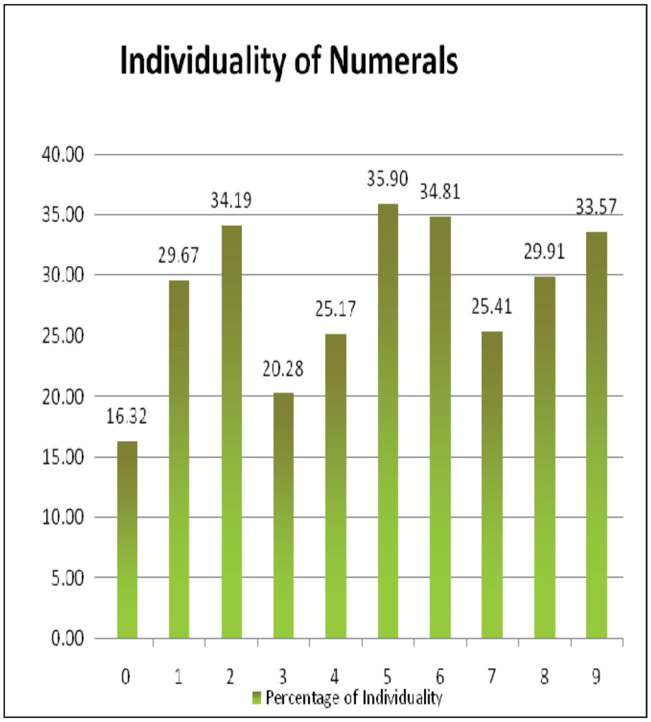


Fig 3 Individuality of Numerals

TABLE I. TABLE FOR WRITER IDENTIFICATION USING NUMERALS INDIVIDUALITY

Number of Writers	Accuracy
20	100%
40	98%
50	96%
55	96%
60	95.33%
65	95.08%
70	96.29%
90	96.5%

VIII. CONCLUSION

In this paper we have presented numeral based Individuality of handwriting and writer identification on Bangla numerals. The main emphasis has been on data collection and evaluation of the individuality of numerals. The research has been done on all numerals from 90 writers with 5 sets each. The LIBLINEAR classifier of the WEKA tool has been used for the identification purpose.

We intend to increase the number of writers for our latter works. Also we are planning to do individuality calculation and writer identification using all the alphabets, numerals and vowel modifiers. Results in this proposed work are encouraging and we will try to increase the data amount and also try different classifiers and features for our future works.

ACKNOWLEDGEMENT

One of the authors would like to thank Department of Science and Technology (DST) for support in the form of INSPIRE fellowship.

REFERENCES

[1] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification – the state of the art", Pattern Recognition, Vol. 22(2), pp. 107-131, 1989.

[2] H. E. S. Said, T. N. Tan and K. D. Baker, "Personal Identification Based on Handwriting", Pattern Recognition, Vol. 33(1), pp. 149-160, 2000.

[3] U. V. Marti, R. Messerli and H. Bunke, "Writer Identification using Text Line Based Features", In Proc. of 6th ICDAR, pp. 101-105, 2001.

[4] S. N. Srihari, S. H. Cha, H. Arora and S. Lee. "Individuality of Handwriting", Journal of Forensic Science, Vol. 47(4), pp. 1-17, 2002.

[5] M. Bulacu, L. Schomaker and L. Vuurpijl, "Writer Identification using Edge-Based Directional Features", In Proc. of 7th ICDAR, pp. 937-941, 2003.

[6] A. Bensefia, T. Paquet and L. Heutte, "Information Retrieval Based Writer Identification", In Proc. of 7th ICDAR, pp. 946-950, 2003.

[7] S. N. Srihari, M. J. Beal, K. Bandi and V. Shah, "A Statistical Model for Writer Verification ", In Proc. of 8th ICDAR, pp. 1105-1109, 2005.

[8] I. Siddiqi and N. Vincent," Writer Identification in Handwritten Documents", In Proc. of 9th ICDAR, pp. 108-112, 2007.

[9] M. Bulacu and L. Schomaker, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features", IEEE Trans. PAMI, Vol. 29(4), pp. 701-717, 2007.

[10] S. Chanda, K. Franke, U. Pal, "Text Independent Writer Identification for Oriya Script", In Proc. of 10th IAPR International Workshop on DAS, pp. 369-373, 2012

[11] P. Purkait, R. Kumar, B. Chanda, "Writer Identification for Handwritten Telugu Documents Using Directional Morphological Features", In Proc. of 12th ICFHR, pp. 658-663, 2010.

[12] A. Sarkar, U. Garain "Individuality of Handwriting: a Study on Handwriting in an Indic Script", In Proc. of ReTIS, pp. 188-191, 2006.

[13] S. Chanda, K.Franke, U.Pal and T.Wakabayashi, "Text Independent Writer Identification for Bengali Script", In Proc. of ICPR, pp. 2005-2008, 2010.

[14] S. Biswas and A. K. Das, "Writer Identification of Bangla handwritings by Radon Transform Projection Profile", In Proc. of 10th IAPR International Workshop on DAS, pp. 215-219, 2012.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Vol. 11(1), pp. 10-18, 2009.

[16] K. Roy., "On the Development of an OCR System for Indian Postal Automation", PhD Thesis, Jadavpur University, 2008.

[17] U. Pal, T. Wakabayashi, N. Sharma1 and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", In Proc. of 9th ICDAR,Vol. 2, pp. 749-753, 2007.