A Numeral Script Identification from a Multi-lingual Printed Document Image

Rajkumar Benne[1(✉)], Shivanand Gornale[2], and Gayatri Patil[2]

[1] Government Autonomous College, Kalaburagi, India
rgbenne@gmail.com
[2] Rani Channamma University, Belagavi, India

Abstract. India is a multi-lingual multi-script country, where a printed document which contains information in the form of texts, images, etc.; the texts part may have composed with characters and numerals of one or more scripts. So, it is necessary Identify the scripts of numerals/characters from multilingual document before feeding them to their individual script OCR systems. In this paper, the system made an attempt to recognize the script of numerals belongs to Kannada, Devanagari, and English based on structural features like water reservoir, aspect ratio, horizontal and vertical strokes. Initially, Bi-script and tri-script numerals script identification experiments are conducted on a dataset of 2100 numerals string(word), by taking 700 samples for each script and noticed average accuracy for tri-script numerals is 93.62%.

Keywords: Script identification · documents · OCR

1 Introduction

India is Multi-lingual Multi-script country, the Indian printed document which contains information in the form of texts of various scripts and images so we need multilingual OCR system. The problem of developing an OCR system can be simplified by sub-categorizing the problem into script identification fallowed by numeral/character recognition. The characters and numerals recognition is out of the scope this paper. We can understand the recognition process of printed and handwritten numerals and characters is not complete without identifying the script of the numerals or characters, before developing a multi-script OCR.

The various authors have attempted to identify script of the text written in handprinted or machine-printed at word level/block level/line level for Indian documents with various techniques [2–5, 16]. All these techniques, identifies the scripts of the text words or lines or blocks of text. But no one has reported about script identification of numerals. Meanwhile, many authors have made an attempt to recognize the numerals of single script and multi-script without script identification of numerals. We repeat and highlight some of the works reported for recognition of numerals with different feature sets including template based approaches and they can be seen in [6, 7]. In addition, structural feature based recognition system [8, 9], statistical feature based recognition

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 178–186, 2023.
https://doi.org/10.2991/978-94-6463-196-8_16

system [1], and hybrid approach based recognition systems [11, 12] are also seen. The task of numerals recognition without script identification can also be done; this kind of work is reported by Dhandra and U.Pal [9, 13, 14]. But, increase in the number of scripts increases the number of classes. In this case, the search space of recognition system increase and hence time complexity also increases. Therefore, it is not an appropriate way of dealing with multi-script numerals identification. In this direction, only few authors have made an attempt to identify the script of the numerals. For instance, G S Lehal and Nivedan Bhatt [15] presented a bilingual recognition system for handwritten numerals of Devanagari (Hindi) and English scripts and also attempted the problem of identification of numeral's script. They have used a set of global and local features to recognize the script of numerals, which are derived from the right and left projection profiles of the numeral image. The task of Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning is presented by Amirreza Fateh [19] and the proposed system was tested with six different languages. Hangairulappan and others [17] reported work on isolated digits of Handwritten Numeral Recognition System. Shrey Malvi [20] claims the work on Variable Length Digit Recognition system for Gujarati Language. Sk Md Obaidullaha and others [18] reported a system of Numeral Script Identification from Handwritten Document Images only.

1.1 Motivation

All the works available in literature are mainly based on script identification from printed and handwritten documents. Some works are reported on numeral recognition from printed and handwritten documents. Till date very few works has been reported on printed Numeral Script Identification, which inspired us to carry out the present work. It has its applicability in different domain of 'smart computing' like automatic sorting of postal documents based on PIN code script, automatic classification of application forms, examination forms etc. written by native languages based on a numeral string.

2 Proposed Method

In this Section, an integrated approach of script identification of numerals for three scripts is presented. The tri-script numerals recognition problem is the ultimate solution to deal with tri-script documents of India. Such kind of recognition systems are helpful in bank transactions, income tax form processing, postal mail processing and various reservation counters. Recognition of numerals from multilingual document images has two approaches: (1) Recognition of numerals without script identification. (2) Identification of the numeral script is first and followed by the recognition of numerals. In the first approach, recognition of multilingual numerals is carried out by adding number of numeral classes to recognize numerals. In second approach, identification of script followed by recognition of numerals. Here, an attempt is made for script identification of numerals from multi-script document. We have used three scripts for experimentation and identification purpose.

For experimentation purpose, we have created our own printed numerals database. Printed numerals are collected from documents of Kannada, Devanagari, and English

scripts of length 10 digits. A dataset of 2100 unconstrained printed numerals of Kannada, Devanagari and English scripts (700 each) are created. The Fig. 1 shows samples of numerals belonging to Kannada, English, and Devanagari scripts.

The India is Multi-lingual Multi-script country, the Indian printed document which contains information in the form of texts and images.

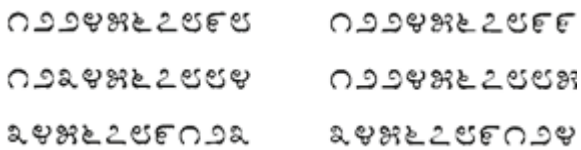
Identification of the script of the numerals consists of the following four stages:

- Acquisition and binarization of document image containing three script numerals.
- Pre-processing and segmentation of numerals from a document using the method proposed by [10].
- Extracted numerals are used for feature extraction.
- Structural features are used to describe each numeral and a single feature vector is formed.
- Using these features, NN classifier is trained. In the same way features are extracted from test numerals and used for identifying the script of the numerals.

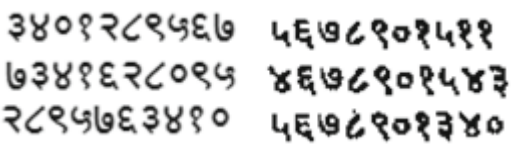
The computation of structural feature like aspect ratio, density (fillhole), vertical and horizontal strokes, and water reservoir are discussed in detail.

Water reservoir based principle

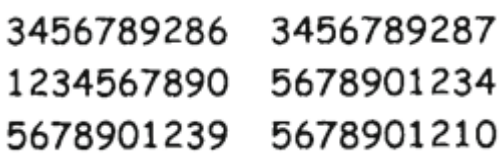
If water is poured from one side of a component, the cavity regions of the component where water will be stored are considered as reservoirs. There are four types of water reservoirs, namely Left, Right, Top and Bottom reservoirs. In this chapter, top and bottom water reservoir that are used to identify the script.



(a)



(b)



(c)

Fig. 1. Sample of numerals: (a) A samples of Kannada numerals. (b) A samples of English numerals. (c) A samples of Devanagari numerals

Top reservoir

The storage region of the water when the water is poured from top of the numeral image. Bottom reservoir: The storage region of the water when water is poured from bottom of the numeral image. Top and bottom reservoir of Kannada, English and Devanagari numerals samples are illustrated in Fig. 2. The observation of the scripts reveals that, the top and bottom reservoirs are present in Kannada script, bottom reservoir is absent in Devanagari script, whereas top and bottom absent in numerals of the English script.

Fill hole density

The looping area of the digits for a numeral is filled with ON pixels, the looping area of the digits varies from script to script. The fill hole density is calculated with respect the image (before fill the hole) and considered as a one of the feature for script identification problem.

Aspect Ratio

Aspect ratio of numeral is calculated by dividing height of the numeral by width of the numeral. The Average aspect ratio is calculated by using the following equation and considered as a one of the feature for script identification problem

$$Average\ Aspect\ Ratio\ (AVR)$$

$$(AVR) = \frac{1}{\boldsymbol{n}} \sum\nolimits_{i=1}^{n} \frac{Height(image\ i)}{Width(image\ i)}$$

Directional stroke estimation:

The directional stroke of numeral image computed on vertical and horizontal direction using morphological transformation with line structuring element. Vertical stroke (VS) and horizontal stroke (HS) are extracted from words of numeral image and finally calculate the AVS and AHS features from below equation with respect to total on pixels of image, and considered for script identification problem. The Fig. 3 shows vertical and horizontal strokes obtained from words of numerals of Kannada, Devanagari and English script.

$$Average\ Vertical\ Stoke\ (AVS)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{On\ pixel\ from\ VD\ (image\ i)}{On\ pixel\ (image\ i)}$$

$$Average\ Horizontal\ Stoke\ (AHS)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{On\ pixel\ from\ HD\ (image\ i)}{On\ pixel\ (image\ i)}$$

Algorithm

The process of identification of script of the numerals is started by extracting a numeral from a document image. On extracted numerals the above proposed features are computed. The computations of features for test and training numerals remain same. The features extracted on trained images inputted to the classifier as a knowledge base. At the end, classifier decides the script of the numerals based on its knowledge base. The complete system of script identification of numerals is briefed out step-wise.

Input: Segmented numeral of three scripts.

Output: Identification of the script of numeral.

Method: Structural feature and Nearest Neighbor classifier.

Step 1. Pre-process the input image [numeral].

Step 2. Fit the minimum rectangle-bounding box to numeral.

Step 3. Extract the Water reservoir based features [Top reservoir and Bottom reservoir] and stored in the library.

Step 4. Calculate Aspect Ratio of numeral and stored in the library

Step 5. Find Fill hole density of numeral and stored in the library

Step 6. Extract the Directional stoke estimation in Vertical direction and Horizontal direction and stored in the library

Step 7. Classify the test image to its appropriate class label using Feature vector stored in the library with NN classifier

Step 8. Stop.

3 Experimental Results

For the purpose of experimentation, 2100 samples of printed numerals of Kannada, Devanagari and English scripts are used. Same data set is used for training and testing purposes, which includes 700-Kannada, 700-English and 700-Devanagari numerals of length ten digits. In the proposed system, simple structural features: Water reservoir, fill hole density, aspect ratio, horizontal and vertical stroke are considered. The classification of numeral's script is carried out with basic Nearest Neighbour (NN) classifier and obtained encouraging results which are shown in the Tables 1, 2 and 3. The result of Kannada-Devanagari script is shown in Table 1, Kannada-English in Table 2, and Devanagari-English in Table 3.

The Tri-script including Kannada-English-Devanagari identification accuracies are presented in Table 4. It can be noticed that, the average bi-script identification accuracies in three cases is high as compared to the results of tri-script identification. It reveals that when combination of more scripts are considered for experimentation, the recognition accuracy falls down. The reason for this is experimentally investigated; it is due to the similarity in shape of the digits of different scripts. For example, the shape of a digit zero of Kannada, Devanagari and English script remains same. Similarly, the shape of a digit four in Kannada resembles to a digit four of Devanagari. These are the reasons for decreased in the script identification accuracies of tri-scripts.

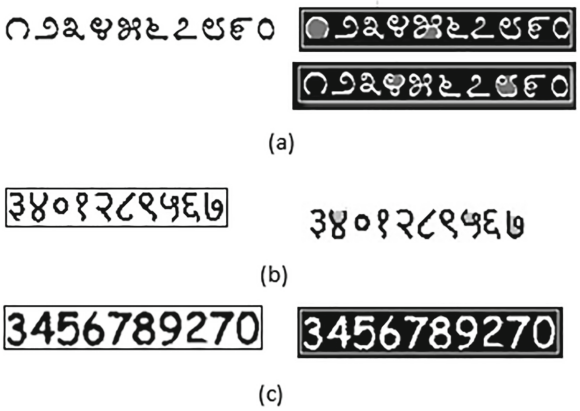


Fig. 2. Water Reservoirs for sample numbers of three different scripts (a) effect of top and bottom Reservoir present in Kannada numeral (b) effect of top and bottom Reservoir present/absent in English numeral (c) effect of top and bottom Reservoir present/absent in Devanagari numeral

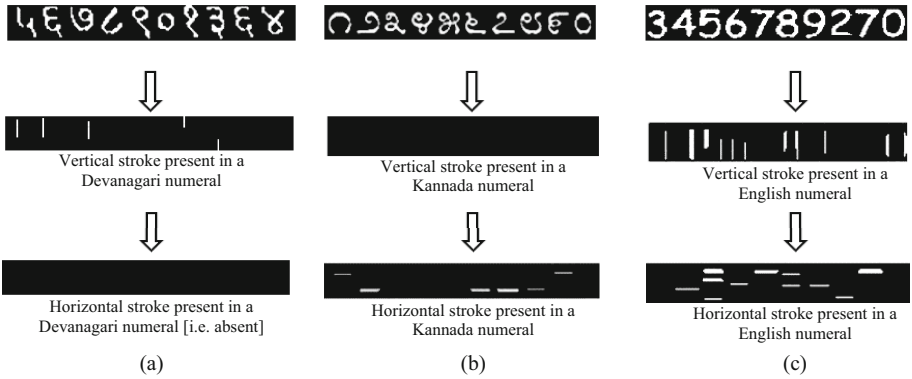


Fig. 3. Effect of Vertical and horizontal stroke present in a Devanagari(a), Kannada(b) and English(c) Numerals

Table 1. Identification accuracy of Kannada and Devanagari numeral script

Numeral script	Recognition accuracy in %
Kannada	93.14
Devanagari	95.29

Table 2. Identification accuracy of Kannada and English numeral script

Numeral script	Recognition accuracy in %
Kannada	93.14
English	94.14

Table 3. Identification accuracy of English and Devanagari numeral script

Numeral script	Recognition accuracy in %
Devangari	95.43
English	94.29

Table 4. Identification accuracy of Kannada, Devnagari, and English numeral script

Numeral script	Recognition accuracy in %
Kannada	92.58
Devanagari	94.43
English	93.86
Average identification rate	**93.62**

4 Conclusion

This paper summarizes a method of structural features for bi-script and tri-script identification script of numerals. The proposed identification system is to identify the script of the printed numeral belonging to Kannada, Devanagari, and English scripts. The average identification accuracy of the Kannada, Devanagari, and English script is 93.62%. The novelty of the proposed method is that recognition accuracy is high with the simple structural feature and basic nearest neighbour classifier. The work proposed in this paper is an attempt towards recognition of the script of numerals for bilingual/multilingual scripts.

References

1. Ivind Trier, Anil Jain, TorfiinnTaxt, "A feature extraction method for character recognition-A survey ", Pattern Recognition, vol. 29, No 4, pp-641–662.
2. Banashree N.P. and R.Vasanta, "OCR for Script identification of Hindi (Devanagari) Numerals using Feature Sub Selection by Means of End-Point with Neuro-Memetric Model", Proceedings of World Academic of Science, Engineering and Technology (PWASET-July 2007) , ISSN 1307–6884, Volume 22, pp. 78–82, 2007.

3. Banashree N.P. and R.Vasanta, "OCR for Script identification of Hindi (Devanagari) Numerals using Error Diffusion Half toning Algorithm with Neural Classifier", Proceedings of World Academic of Science, Engineering and Technology (PWASET-April 2007), ISSN 1307–6884, Volume 20, pp. 46–50, 2007.
4. R.J.Ramteke, P.D.Borkar, S.C.Mehrotra, "Recognition of Isolated Marathi Handwritten Numerals: An Invariant Moments Approach", Proceedings of the International Conference on Cognition and Recognition, pp.482–489.
5. M.Hanmandlu and O.V. Ramana Murthy, "Fuzzy Model Based Recognition of Handwritten Hindi Numerals", Proceedings of the International Conference on Cognition and Recognition, pp.490–496
6. Anil K.Jain, Douglass Zonker, "Representation and Recognition of handwritten Digits using Deformable Templates", IEEE, Pattern analysis and machine intelligence, vol.19, no-12, 1997.
7. J.D.Tubes, "A Note on Binary Template Matching". Pattern Recognition, 22(4):359-365, 1989.
8. B.V.Dhandra, V.S.Mallimath, Mallikargun Hangargi and Ravindra Hegadi, "Multi-font Numeral recognition without Thinning based on Directional Density of pixels", IEEE International conference on Digital Information Management (ICDIM-2006) Bangalore, India, pp.157–160, Dec-2006.
9. R Sanjeev Kunte and Sudhakar Samuel R.D, "Script Independent Handwritten Numeral recognition". VIE -2006, pp. 94–98, September 2006
10. B.V.Dhandra, and Mallikargun Hangargi, "Morphological Reconstruction for Word Level Script Identification" International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 41–51
11. Dinesh Acharaya U, N.V.Subba Reddy and Krishnamoorthi, "Multilevel classifier in Recognition of Handwritten Kannada Numerals", PWASET-2008, ISSN-2070–3740, pp.308–313, 2008
12. SubhangiD.C., P.S. Hiremath, "Handwritten English character and Digit Recognition Using Multiclass SVM classifier and Using structural micro features", International Journal of Recent Trends in Engineering. Vol.2, No.2, pp. 193-195, 2009.
13. U.Pal, N.Sharma, F.Kimura, "Handwritten Numeral recognition of six popular Indian scripts" IEEE-explorer, 2008.
14. B.V.Dhandra, R.G.Benne, and Mallikarjun Hangarge, "Kannada, Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network: A Script Independent Approach", International Journal of Computer Application, IJCA (0975–8887), Volume 26, No-9, july-2011.
15. G S Lehal and Nivedan Bhatt ,"A Recognition System for Devnagri and English Handwritten Numerals"
16. Mallikarjun Hangarge, Kc Santosh and Rajmohan Arjunsingh Pardeshi, "Directional DCT for Handwritten Script Identification", International Conference on Document Analysis and RecognitionAt: Washington DC, USA, August 2013.
17. Hangairulappan Kathirvalavakumar, M. Karthigai Selvi and R. Palaniappan, "Efficient Handwritten Numeral Recognition System Using Leaders of Separated Digit and RBF Network", Second International Conference MIKE 2014, Cork, Ireland, December 10-12, 2014. Proceedings, pp 135–144.
18. Sk Md Obaidullaha, Chayan Halderb, Nibaran Dasc and Kaushik Roy, "Numeral Script Identification from Handwritten Document Images", Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Elsevier-Procedia Computer Science 54 (2015) 585 – 594.

19. Amirreza Fateh, Mansoor Fateh, and Vahid Abolghasemi," Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning", Elsevier-Information Sciences, Volume 581, December 2021, Pages 479-494
20. Shrey Malvi, Nirmal Patel and Pratik Prajapati," Variable Length Digit Recognition for Gujarati Language", Easy Chair Preprint no. 7672, March 29, 2022.