INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Classification of North and South Indian Handwritten Scripts using Gabor Wavelet Features

S K Shreesha[1], **H B Anita**[2]*

1 Department of Computer Science, PES University, Bengalure, Karnataka, India
2 Department of Computer Science, CHRIST(Deemed to be University), Bengaluru, Karnataka, India

OPEN ACCESS

Received: 12-01-2022

Accepted: 28-02-2022

Published: 29.04.2022

Citation: Shreesha SK, Anita HB (2022) Classification of North and South Indian Handwritten Scripts using Gabor Wavelet Features. Indian Journal of Science and Technology 15(16): 712-717. https://doi.org/10.17485/IJST/v15i16.88

* **Corresponding author**.

anita.hb@christuniversity.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Handwritten script identification plays a vital role in processing handwritten data electronically. Most of the methods fail to provide accuracy due to variation in handwriting, hence the classification of the Indic script before providing it to OCR is crucial. The anticipated work helps increase the accuracy by categorizing the handwritten documents as north or South Indic script before further classification. **Methods**: This study has proposed a method, using Gabor filters to extract features from the text image for recognizing the kind of script, and seven widely used Indian scripts were considered for this experiment. The handwritten documents were collected from distinct individuals on request, under supervision. The database was manually created by extracting portions of lines from the scanned document images. **Findings**: A recognition accuracy of 100% was obtained for classifying North and South scripts while an average accuracy of 92% was obtained for bi-script classification using KNN classifier at a portion of the line level. **Novelty**: The proposed method improves the accuracy by acting as a pre-processor to the OCR system by classifying the script according to North Indian script or South Indian Script. Further, it can be processed to find out the script type within the North or South Indian Scripts.

Keywords: Handwritten Script; Gabor Filter; KNN Classifier; OCR; Indic Script

1 Introduction

Interpretation of scripts from digitalized document images is at the heart of every document image understanding system. Multi-script documents are pretty common as India is a multilingual country with a colonial past. Most of the researchers have worked on bi-script, tri-script classification at document, page and line levels. Also there is a great demand for automatic processing of multilingual documents. A lot more work is required on script identification. The success rate of recognition can be increased by classifying north and south Indian scripts before identifying the type of script.

This paper presents a pre-processor for handwritten script recognition system at portion of the line level. The portion of line may contain two or more words in a line, considering requirement of the same script on a single line. Proposed technique segregates handwritten scripts into North (Bengali, Punjabi, Oriya, and Devanagari) and South (Kannada, Malayalam, Telugu, and Tamil) handwritten Indian Scripts. This can be used as a pre-processor to the system which classifies the any Indic script. Once the given script is identified as North or South, it can be further classified. This particular subdivision helps to improve the accuracy. Proposed word identifies the script as North of South Indian Script then classifies the type of specific script.

In the past two to three decades many researchers have worked on Script Identification. Most of their work consisted feature extraction in spatial or frequency based domains and later classified using machine learning techniques like K nearest neighbor, Support Vector Machine, Neural Network etc. Now a days many researches are using Conventional Neural Networks, Deep Learning models etc. As per survey, classifying of North and South Indian scripts was not attempted.

Identification techniques for handwritten documents at word, line and block level is attempted (1). In this texture was used as a factor for assessing the script of handwritten document image with consideration on the observation that, text has a distinct visual texture to classify the scripts, namely English, Devanagari and Urdu. Some of the authors worked on numeric (2) and deep learning (3) technique for script recognition. Recently many researchers have tried Deep Learning for script recognition. In (4) the researchers applied HMB1 and HMB2 convolutional neural networks for Arabic character recognition. In (5) the authors experimented a fully convolution based deep network architecture for cursive handwriting recognition from line level images of English and French. Similarly, Convolutional Neural Network experimented on Arabic letters dataset which is written by the children of age group 7 to 12 to recognize the Arabic handwritten letters in (6). In (7) Meitei Mayek handwritten character recognition is also experimented using CNN and got encouraging results. In (8) the authors have evaluated SVM and PCA used for recognizing digit character of Devanagari script. Local Binary Pattern operator method is experimented and the features are extracted from a block of handwritten text images for classifying handwritten documents written in English, Hindi, Kannada, Malayalam, Telugu, and Urdu scripts (9). Tesseract Tool and CNN experimented (10) and researchers obtained very good result for recognizing the Kannada characters. (11) CNN model is created to recognize Gujarati characters. (12) Essential and relevant features are selected by using Hybrid Swarm and Gravitation-based method. Total numbers of 144 features were selected and DHT algorithm was used with KNN classifier. Similarly, Histogram of Oriented Gradients (HOG) method is used for extracting 149 features, This newest review indicated most of the researchers got average accuracy of 90% for classifying handwritten Indic scripts at line level. Detail case study on handwritten script classification at block, Line and Word level is carried out (13). 12 Indian Scripts, 12 Text line level datasets and six best classifiers are considered for the experiment. Study shows that by using MLP Classifier at Text Line-level Using GLCM Feature Descriptor and Combination of DWT with RT Feature set the researchers achieved TP rate of 0.897 and 0.889 respectively.

In conclusion, certain authors worked only on bi-script, tri-script classification (1,5) and few carried out their research on character or word level dataset for Indic scripts and not line level (2). Certain researchers selected only a particular script and implemented character recognition (4,7,8,10,11), others worked on non-Indic scripts like Arabic, French etc (4,6). (9) Experimented Indic script classification at block level. (12) Attempted Indic script classification but the accuracy obtained was lesser than 90% for line level handwritten scripts. (13) Implemented new hybrid method called Hybrid Swarm and Gravitation-based Feature Selection and attained 97.74% but had a feature vector of size 130.

This was an inspiration to design a system for the classification of North and South Indian Handwritten scripts. The proposed work can act as a pre-processor to the system which identifies the type of Indic script. This method uses 54 features extracted by Gabor filter to classify the given script as North or South Indian Script. Further details of the proposed method are elaborated in the following sections.

2 Methodology

2.1 Data Collection and Pre-processing

The handwritten documents were collected from distinct individuals under supervision. The writers were asked to write few lines of text in A-4 size sheets. Also, there were no restrictions imposed regarding the use of pen and content of the text. The chosen writers were from different professions, different states, and from different geographical locations in India, that included both native and non-native writers. The writing style of native writer could be different from a writer for whom the same language would be his/her second language. The scanning of scripts took place at 300 dpi resolution and were stored as gray-scale images. Creation of database took place by extracting portions of line from images of scanned documents.

Pepper noise could appear in a document image during the conversion process and also caused from the dirt on the document. This noise can be composed of one or more pixels, but by definition, they are assumed to be much smaller than the size of the text objects. Isolated pepper noise was removed by median filtering. Median filtering was applied so as to replace the current point in the image by the median of brightness in its neighborhood. It is necessary to apply thinning operation to the text image since the text written vary in thickness depending on the writer and the type of pen used to write. Thinning operation reduces the thickness of the written text facilitating better feature extraction. The width of the thinned line in the text is of one pixel. It lies along the center of the original line.

It is necessary to convert handwritten text image to binary image for efficient processing since the text may be written using different pens. Thresholding of document images is vital in image binarization. Ostu's global thresholding method was used for binarization. Pepper noise as well as the salt noise around the boundary was removed using morphological opening. Further this operation was also extended to remove discontinuity at the pixel level. We do not perform any processing to homogenize these parameters. It has been ensured that text occupies at least 50% of the handwritten page. Further, we do not eliminate dots and punctuation marks appearing in the document image, since this contributes to the features of respective scripts. A total of 700 handwritten line images containing text were created, with 100 lines per script. The dataset was obtained considering width equal to 512 pixels and height equal to that of the largest character appearing in line. Figure 1provides a sample of line images representing different scripts.

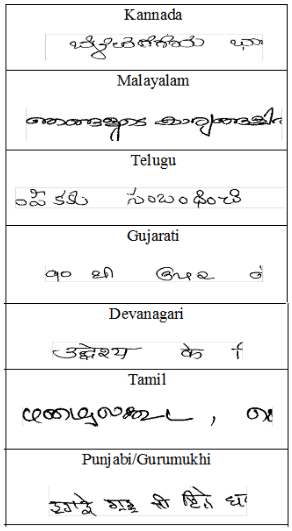

Fig 1. Sample of Line Images representing different Scripts

2.2 Feature Extraction

Feature extraction is most substantial component of script recognition system. The objective of feature extraction is to accurately retrieve the features from the given image object. These extracted features should maximize the distinction between bi-scripts. In frequency domain of the image, we can obtain features which are not noticeable in the time domain. In this proposed method, we employ two-dimensional Gabor filters to extract the features from input text image to identify the script type. Features are extracted using two-dimensional Gabor functions by transforming the image into frequency domain. Description of the extracted or used features is provided below.

Gabor Filter is used recognize and analyze the frequency content of an image in a particular orientation. For this, certain frequency is taken which is oriented and modulated by a two-dimensiona Gaussian function.

The Gaussian function is provided below.

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) exp\left(-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2}+\frac{y'^2}{\sigma_y^2}\right)\right) exp(2\pi jWx')$$

$$x' = xcos\varnothing + ysin\varnothing$$

$$y' = -xsin\varnothing + ycos\varnothing$$

The spatial extension of the filter is controlled by σx2 and σy2, whereas the orientation of the filter is represented by θ and w is the frequency of the sinusoid. Filtering process on the input images is carried about by Gabor filters tuned at several orientations and resolutions. However it is not feasible or computationally convenient for having a large number of filters responding at multiple resolutions and orientations, hence Gabor filters were chosen at certain frequency bands and orientations. There was a restriction of five orientations made on computational savings and the applied feature extraction method is prescribed below.

Algorithm-1

Input: Gray scale image at line level.

Output: A feature extracted vector is obtained.

Method:

1) Remove noise by the application of median filtering (Figure 2a).

2) Apply Otsu's method and invert the image to yield results in binary. Binary 1 representing text and 0 representing the background (Figure 2b).

3)Use morphological opening for removal of unwanted small objects around the boundary (Figure 2d).

4) Apply thinning operation (Figure 2e).

5) Place bounding box over the portion of line to crop the image.

6) Consider six unique orientations at three different frequencies to obtain 18 filters.

7) Convolve the input image with the created Gabor filter Bank (Figures 3 and 4).

8) Perform the following steps considering each of the output images from step 7.

a) Compute the standard deviation on both sine and cosine part separately (36 features).

b) Compute the standard deviation of the entire output image (18 features)

9) This Form the feature vector of length 54 by computing Standard Deviation on the 54 convolved images obtained.

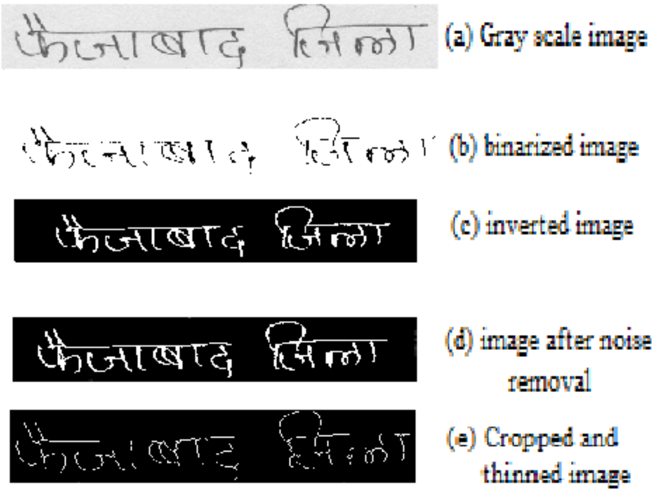

Fig 2. Pipeline process for feature extraction.

2.3 Script Recognition

K-NN classifier was mainly adopted for the purpose of recognition and is a well-known non-parametric classifier, where posterior probability is estimated from the frequency of the nearest neighbors of the unknown pattern. The training phase comprised of extracting the features from the training set by performing the feature extraction algorithm as mentioned above. In order to form a knowledge base, these features were given as input to K-NN classifier that was subsequently used to classify the test images. In the test phase, the test image that is to be recognized is processed in a similar way and features are computed as per the same algorithm. The required Euclidean distances between the test feature vector with that of the stored features is computed by the classifier and it also identifies the k-nearest neighbor. Finally, the K-NN classifier assigns the test image to a class that has the minimum distance with voting majority. The script corresponding to this is declared as a recognized script.

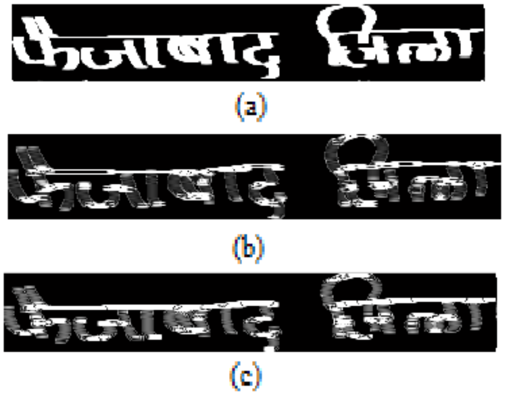

Fig 3. Gabor filtered images for zero degree orientation and frequencies a, b, and c.

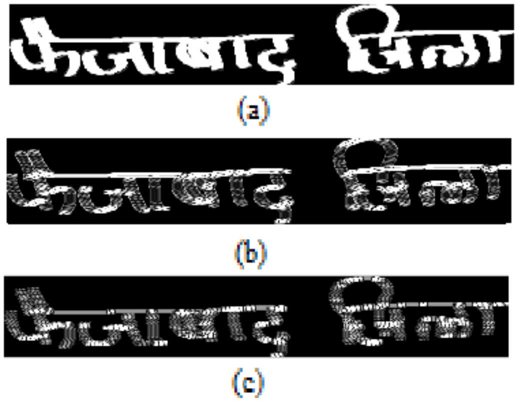

Fig 4. Gabor filtered images for 60 degree orientation and frequencies a, b, and c

3 Results

Conducting script classification for portion of the line is much faster as compared to the analysis of the entire line extracted from handwritten documents. Experiments were performed for identifying the script type. The dataset consisted of 700 pre-processed images (Figure 1). Test script was identified using the k-NN classifier and performance of the proposed system was evaluated over a dataset of 700 pre-processed images. Out of this, 300 images used for north Indian Script and remaining 400 images used for South Indian Script. While training, 60% of the images were used and rest 40% of the images for testing. Optimal results were obtained when the value of k was unity (one) as compared to other values of k. The empirical ratio of 60:40 split was considered for breakdown of the training and testing data. The proposed work achieved a remarkable accuracy when the classification was performed at two stages. First the North and south script classification and in the next stage bi-script classification was carried out. Recognition results are provided in Tables 1 and 2 respectively. The obtained results clearly indicates that features extracted using Gabor function yield good results for the classification of North and South Indian Scripts. As per our knowledge no author worked on north and south indic script classification. Also the proposed method had worked well for classifying Kannada and Hindi scripts with other South and North Indian Scripts respectively.

Table 1. Recognition results in % for North and South Indian scripts documents at portion of the line.

Scripts considered for Classification	Recognition in %
North, South	100

Table 2. Recognition results in % for bi-script documents at portion of the line.

Scripts considered for Classification	Recognition in %
Kannada, Tamil	100
Kannada, Malayalam	100
Kannada, Telugu	100
Hindi, Gujrati	100
Hindi, Punjabi	100
Gujrati, Punjabi	82.5
Tamil, Telugu	83
Tamil, Malayalam	80.5
Malayalam, Telugu	83

4 Conclusion

This study has developed a recognition system to differentiate North- and South-Indian scripts. The obtained accuracy considering all scripts without the proposed North and South script classification was less than 70%. The proposed method when used as a pre-processor increases the accuracy to a minimum of 80% for bi-script classification as indicated in Table 2. Here, Gabor filter was used for feature extraction and KNN classifiers for recognition. A remarkable recognition rate is achieved. The method proposed is highly reliable, robust, and independent of the factors such as the style of handwriting and thickness of text. This method can be used as a pre-processor to the script identification system. In the future, we intend to extend this proposed method for identification of North and South Indian scripts at the word level.

References

1) Rajput DGG, B AH. Handwritten Script Recognition using DCT, Gabor Filter and Wavelet Features at Line Level. *International Journal of Electronics Signals and Systems*. 2011;1(2):85–90. doi:10.47893/IJESS.2011.1017.
2) Bhunia AK, Mukherjee S, Sain A, Bhunia AK, Roy PP, Pal U. Indic handwritten script identification using offline-online multi-modal deep network. *Information Fusion*. 2020;57:1–14. Available from: https://dx.doi.org/10.1016/j.inffus.2019.10.010.
3) Obaidullah SM, Santosh KC, Das N, Halder C, Roy K. Handwritten Indic Script Identification in Multi-Script Document Images: A Survey. *International Journal of Pattern Recognition and Artificial Intelligence*. 2018;32(10):1856012–1856012. Available from: https://dx.doi.org/10.1142/s0218001418560128.
4) Balaha HM, Ali HA, Saraya M, Badawy M. A new Arabic handwritten character recognition deep learning system (AHCR-DLS). *Neural Computing and Applications*. 2021;33(11):6325–6367. Available from: https://dx.doi.org/10.1007/s00521-020-05397-2.
5) Sharma A, Jayagopi DB. Towards efficient unconstrained handwriting recognition using Dilated Temporal Convolution Network. *Expert Systems with Applications*. 2021;164:114004–114004. Available from: https://dx.doi.org/10.1016/j.eswa.2020.114004.
6) Altwaijry N, Al-Turaiki I. Arabic handwriting recognition system using convolutional neural network. *Neural Computing and Applications*. 2021;33(7):2249–2261. Available from: https://dx.doi.org/10.1007/s00521-020-05070-8.
7) Inunganbi S, Choudhary P, Manglem K. Meitei Mayek handwritten dataset: compilation, segmentation, and character recognition. *The Visual Computer*. 2021;37(2):291–305. Available from: https://dx.doi.org/10.1007/s00371-020-01799-4.
8) Khamparia A, Singh SK, Luhach AK. SVM-PCA Based Handwritten Devanagari Digit Character Recognition. *Recent Advances in Computer Science and Communications*. 2021;14(1):48–53. Available from: https://dx.doi.org/10.2174/2213275912666181219092905.
9) Rajput GG, Ummapure SB. Script Identification from Handwritten document Images Using LBP Technique at Block level. *2019 International Conference on Data Science and Communication (IconDSC)*. 2019;p. 1–6. doi:10.1109/IconDSC.2019.8816944.
10) Fernandes R, Rodrigues AP. Kannada Handwritten Script Recognition using Machine Learning Techniques. *2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. 2019;p. 1–6. doi:10.1109/DISCOVER47552.2019.9008097.
11) Shirke A, Gaonkar N, Pandit P, Parab K. Handwritten Gujarati Script Recognition. *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2021;p. 1174–1179. doi:10.1109/ICACCS51430.2021.9441811.
12) Singh PK, Sarkar R, Abraham A, Nasipuri M. A Case Study on Handwritten Indic Script Classification: Benchmarking of the Results at Page, Block, Text-line, and Word Levels. *ACM Trans Asian Low-Resour Lang Inf Process*. 2021;21(2):36–36. Available from: https://doi.org/10.1145/3476102.
13) Guha R, Ghosh M, Singh PK, Sarkar R, Nasipuri M. A Hybrid Swarm and Gravitation-based feature selection algorithm for handwritten Indic script classification problem. *Complex & Intelligent Systems*. 2021;7(2):823–839. Available from: https://dx.doi.org/10.1007/s40747-020-00237-1.