Script Identification using Discrete Curvelet Transforms

B.V.Dhandra
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Gulbarga, India

Vijayalaxmi.M.B and Gururaj Mukarambi
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Gulbarga, India

Mallikarjun Hangarge
Department of P.G. Studies and Research in Computer Science
Karnatak Arts Science and Commerce College, Bidar

ABSTRACT

This paper presents discrete curvelet transform (DCvT) based block level handwritten script identification. The conventional two-dimensional (2-D) discrete wavelet transforms (DWTs), de-emphasizes directional discriminating properties such as curves, lines and edges of the texture under study and whereas discrete curvelet transform (DCvT) efficiently extracts directional selective features. Typically it can be observed that the patterns of any handwritten text blocks encompass directionally dominant texture primitives. Therefore, the primary aim of this paper is to show the efficiency of discrete curvelet transform (DCvT) in describing the handwritten text blocks of six Indian scripts. Exhaustive experimentations were conducted on a large dataset with various combinations of scripts. For instance, average script classification accuracy achieved in case of bi-scripts and tri-scripts combinations are 94.19% and 95.24% respectively.

General Terms

Document Image Processing, Pattern Recognition.

Keywords

Bilingual, Trilingual, Multilingual, Script Identification, Curvelet Transform, Nearest Neighbor classifier, texture features.

1. INTRODUCTION

Script identification is one of the important pre-processing steps of automatic processing of multilingual document images. The problem of script identification may be addressed by considering bi-scripts, tri-scripts and multi-scripts documents. Script identification in a multi-lingual environment has various applications such as indexing and retrieval of text as an initial step towards optical character recognition. Automatic script and language identification facilitates to read and process the multi-script documents for various applications.

All the existing works on automatic handwritten script identification can be classified into two categories: i) Local Approach ii) Global Approach. The local approaches [2, 4] employ morphological, water reservoir principle, cavities, corner points, end point connectivity, top and bottom profiles based features. Basically local approaches works on connected component analysis and their performance is poor in case of broken characters and improper segmented components, slower in computation, sensitive to noise. On the other hand, global approaches involve analysis of large images or the regions (blocks) consisting of two or more text lines, hence segmentation at line, word and character level is not necessary. So script classification task is simple and faster using global approach compared to local approach. Moreover, most of the global methods have proposed are not efficient in capturing the directional edge and curve information which of course plays a significant role in shape analysis. These observations motivated us to present a generalized global method to overcome all the limitations of the aforementioned methods.

2. PREVIOUS WORK

A tool was developed for the identification of Tamil, English, Hindi, Malayalam, Kannada and Telugu printed scripts irrespective of their font styles and sizes at word level in [2]. The shape, density and transition features were used to perform the nine zone segmentation over the characters. Then script was determined by using rule based classifiers containing set of classification rules which are raised from the zones and obtained accuracy of 97.8%, 89.8%, 92.1%, 86.1%, 89.3% and 86.2% for Tamil, English, Hindi, Malayalam, Telugu and Kannada words respectively.

Patil et.al. [3] have used black pixel distribution in each script as a potential feature for English, Hindi and Kannada script identification at block level. The recognition accuracy of 96% is obtained using single feed forward neural network and of 99% by using modular NN. Pal et. al reviewed the OCR system on Indian scripts in [4]. Dhandra et. al. have classified three handwritten Indian scripts namely English, Devanagari and Urdu based on 13 spatial spread features extracted from morphological filters at block level and line level in [5]. The experiments were performed on Urdu, English and Devanagari scripts by considering the block size of 128 x 128 pixels. Using KNN classifier with five fold cross validation they have achieved an average recognition accuracy of 99.2% for bi-script and 88.6% for tri-script at text line and block level respectively. Pati et.al. [6] have used Gabor and discrete cosine transform (DCT) based features for word level multi-script identification of printed document that have been independently evaluated using nearest neighbor, linear discriminant and support vector machine (SVM) classifiers and showed recognition accuracy of 98% for bi-scripts and tri-scripts and above 89% for the eleven-scripts scenario. Joshi et al [7] have extracted features consistent with human perception and used hierarchical classification scheme for script identification from Indian documents. They have extracted local energy based features using log-Gabor filter for printed script classification at block level and with KNN and Parzen window classifier achieved the recognition accuracy of 97%.

Dhanya et al [8] have reported word level script identification in a bilingual document image containing Roman and Tamil scripts. The SVM classifier gave 88.39% using spatial features and 96.03% using Gabor filter responses. Rajput et. al. [9] used DCT and Wavelets of Daubechies Family based features and achieved the recognition accuracy of 96.4% using nearest neighbor classifier.

Hangarge et.al. [10] have considered automatic handwritten script identification as a texture analysis problem. The Gabor filters are used to extract oriented energy features of size 24. The KNN classifier with two fold cross validation gave average tri-script classification accuracy of 91.99 %.

Lindsay et al [13] developed an automated imaging system for classification of tissues in medical images obtained from Computed Tomography (CT) scans. The approach consisted of two steps: automatic extraction of the most discriminative texture features of regions of interest and creation of a classifier that automatically identifies the various tissues. The discriminating power of several curvelet-based texture descriptors were investigated. Tests indicated that Energy, Entropy, Mean and Standard Deviation signatures were the most effective descriptors for curvelets, yielding accuracy rates in the 97- 98% range.

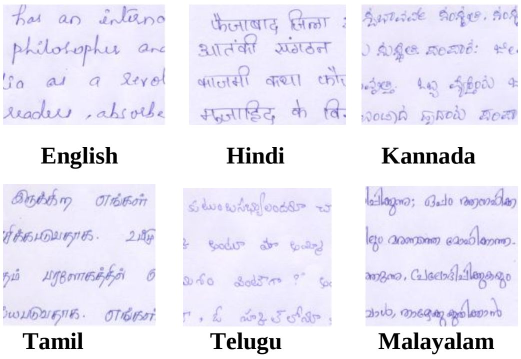
Hangarge et al [14] used two different methods to capture directional edge information. One method by performing 1D-DCT along left and right diagonals of an image and another by decomposing 2D-DCT coefficients in left and right diagonals. The mean and standard deviations of left and right diagonals of DCT coefficients were computed and using linear discriminant analysis (LDA) and K-nearest neighbor (K-NN) classification of the words is performed at biscripts, triscripts and multiscripts cases and the identification accuracies of 96.95%, 96.42% and 85.77% were achieved where 9000 words belonging to six different scripts were considered for validation.

In this paper a generic global method is developed using discrete curvelet transform (DCvT) to extract curves, edges and small line segments to discriminate the text patterns of the scripts.The paper is organized as follows. The database collected for testing the proposed algorithm is presented in Section 3. Section 4 describes feature extraction. The experimental results obtained are presented in Section 5 followed by conclusion in Section 6.

3. DATA COLLECTION AND PREPROCESSING

3.1 Data Collection

The standard database for Indian scripts is not available. So the handwritten documents are collected from different writers of different age groups and professions. The collected documents of English, Hindi, Kannada, Tamil, Telugu and Malayalam are scanned through scanner HP Scanjet G2410 to obtain digitized images. The scanning is performed at 300 dpi resolution. The 100 blocks of each script are segmented from the scanned document images. The size of the text block considered for experimentation is 512x512 pixels. Few sample text block images of six scripts are shown in Fig. 1.



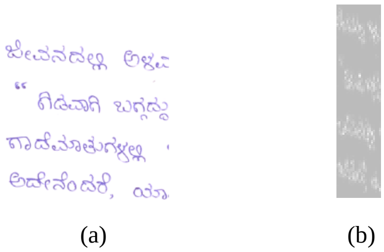
English **Hindi** **Kannada**

Tamil **Telugu** **Malayalam**

Figure 1: Sample blocks in 6 different scripts

3.2 Preprocessing

The scanned documents are binarized using Otsu's global threshold approach. The small connected components such as commas, hyphens and isolated characters are removed using morphological openings. The sample document image and its curvelet image are presented in Fig 2.



(a) (b)

Figure 2: Sample blocks of a) Original image b) its Curvelet image at scale 2 and angle 8

4. FEATURE EXTRACTION

As discussed above the handwritten text patterns have curves as well as straight lines, so curvelet transform is designed to extract these. It allows representing edges and other singularities along lines in a more efficient way than other transforms.

The texture features used in this algorithm are derived from the Discrete Curvelet Transform (DCvT), introduced by Candes and Donoho in [11]. This is a discretization of their continuous curvelet transform [12], which uses a "wrapping" algorithm. Curvelet coefficients have different scales and angles. Energy of these coefficients is different for different coefficients based on angle and scale.

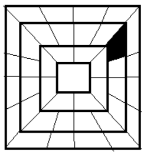


Figure 3: Discrete curvelet frequency tiling domain, wedge samples are shaded.

Two parameters are involved in the discrete implementation of the curvelet transform: number of scales and number of angles at the coarsest level. In the proposed method, the 512x512 image block is decomposed into four scales using real-valued curvelets. The number of second coarsest level angles used is 8. We have taken only the standard deviation of curvelet coefficients obtained for each of the wedge created by four levels of resolution and 8 angular orientations.

Algorithm: Curvelet Feature Extraction for Script identification

Input: Text block of size 512X512

Output: Recognition of the script of the text block

Begin

1. Convert gray scale image into binary image using Otsu's thresholding method.
2. Preprocess the image by applying morphological operations
3. Apply Discrete Curvelet Transform with "wrapping" algorithm. The transform consists of four steps: application of a 2-dimensional Fast Fourier transform of the image, formation of a product of scale and angle windows, wrapping this product around the origin, and application of a 2-dimensional inverse fast Fourier transform. The scale of 4 and angular orientations 8 are used to

produce ‘wedges’. For each wedge curvelet coefficients are obtained.

4. Compute standard deviation of curvelet coefficients (except for scale=1), obtained in step 3 to get feature set size of 20 dimensions.

5. Store feature vector of each script of each text block in the database.

6. The features of unknown script is computed as explained in step 3 and 4 and then given to nearest neighbor classifier to identify the script of the text block.

End.

From experiment it is found that for scale=1 the curvelet coefficient obtained is not dominating feature, hence we have considered other three scale’s that is scale=2, 3, 4). Applying Discrete Curvelet Transform on the preprocessed image gave 8, 16 and 16 numbers of wedges for scales 2, 3 and 4 respectively. Among these 40 wedges the standard deviation is obtained for curvelet coefficients of 20 wedges only i.e, 4, 8 and 8 wedges (wedges of upper half part with scale 2, 3 and 4 respectively) after observing negligible contributions of the wedges of upper half part with scale 2, 3 and 4 respectively.

5. EXPERIMENTAL RESULTS

The experiments are carried out on 100 text blocks of each script that are segmented from the scanned document images. The size of the text block considered is 512x512 pixels. The proposed method gave outperforming results with nearest neighbor classifier with two fold cross validation. The average recognition accuracy for bilingual scripts is 94.19% as shown in Table 1 and the maximum recognition accuracy is 100% for Telugu and English and is due to dissimilar shapes of the scripts. The minimum accuracy is 86% for Tamil and Malayalam and is due to similarity of their character shapes.

The average recognition rate of Kannada, Telugu is 87.5%, and Kannada, Malayalam is 87.5% and that of Tamil and Malayalam is 86%. Due to the shape similarity of Kannada and Telugu characters, Kannada and Malayalam characters the classifier has more confusion. On the other hand average recognition rate of Kannada, Tamil is 97.50%, it has shown high accuracy because there is less similarity between Kannada and Tamil characters.

The average recognition accuracy for trilingual scripts with four combinations is shown in Table 2.

From the Table 3, the overall average recognition accuracy of bilingual scripts is 90.07%. The maximum recognition accuracy is 96.67% for English, Hindi, and Telugu and is due to dissimilar shapes of the scripts. The minimum accuracy is 82% for Telugu, Tamil, and Malayalam and is due to similarity between character shapes. In Hindi, Kannada, Malayalam combination some of the Kannada blocks are misclassified as Malayalam, most of the Malayalam blocks are misclassified as Kannada due to similarity of Kannada and Malayalam scripts. And most of the Hindi blocks are misclassified as Kannada. This is due to the effect of writing style of native Kannada writer used to write Hindi.

Table 1. Average recognition accuracy for bilingual script identification using 2 fold cross validation with Nearest Neighbor classifier

Bilingual script group	**Bilingual scripts**	**Recognition accuracy in (%)**
1	English, Kannada	97
2	English, Telugu	100
3	English, Tamil	93.50
4	English, Malayalam	96.50
5	Hindi, Kannada	93
6	Hindi, Telugu	97.50
7	Hindi, Tamil	98
8	Hindi, Malayalam	95
9	English, Hindi	97
10	Kannada, Telugu	87.50
11	Kannada, Tamil	97.50
12	Kannada, Malayalam	87.50
13	Tamil, Malayalam	86
14	Tamil, Telugu	97
15	Malayalam, Telugu	89.90
Average Recognition Accuracy		94.19

Table 2. Average recognition accuracy for four combinations of south Indian trilingual scripts using 2 fold cross validation with Nearest Neighbor classifier

Trilingual script group	**Trilingual scripts**	**Recognition accuracy in (%)**
1	English, Hindi, Kannada	93.33
2	English, Hindi, Telugu	96.67
3	English, Hindi, Tamil	95.67
4	English, Hindi, Malayalam	95.33
Average Recognition Accuracy		95.24

Table 3. Average recognition accuracy for all combinations of South Indian trilingual scripts using 2 fold cross validation with Nearest Neighbor classifier

Trilingual script group	Trilingual scripts	Recognition accuracy in (%)
1	English, Hindi, Kannada	93.33
2	English, Hindi, Telugu	96.67
3	English, Hindi, Tamil	95.67
4	English, Hindi, Malayalam	95.33
5	Hindi, Kannada, Telugu	86.00
6	Hindi, Kannada, Tamil	91.33
7	Hindi, Kannada, Malayalam	87.33
8	Hindi, Telugu, Tamil	94.33
9	Hindi, Telugu, Malayalam	87.67
10	Hindi, Tamil, Malayalam	86.33
11	English, Kannada, Telugu	87.33
12	English, Kannada, Tamil	94
13	English, Kannada, Malayalam	92
14	English, Telugu, Tamil	96
15	English, Telugu, Malayalam	89.67
16	English, Tamil, Malayalam	90.67
17	Kannada, Telugu, Tamil	90
18	Kannada, Telugu, Malayalam	79.67
19	Kannada, Tamil, Malayalam	86
20	Telugu, Tamil, Malayalam	82
Average Recognition Accuracy		90.07

.(K-Kannada, H-Hindi, E-English, T-Telugu, Tm-Tamil, M-Malayalam)

The proposed method gave 95.24% of recognition rate with EHK, EHT, EHTm, and EHM combinations of scripts as shown in Table 4, where as Hangarge et. al.'s method gave 91.99% of recognition rate. This enhancing recognition rate of scripts is due to role of Curvelet features. The proposed method gives 90.07% of recognition rate with all combinations ie., EHK, EHT, EHTm, EHM, HKT, HKTm, HKM, HTTm, HTM, HTmM, EKT, EKTm, EKM, ETTm, ETM, ETmM, KTTm, KTM, KTM and TTmM

Table 4. Comparative Analysis of Block Level trilingual script identification

Method Used	Feature Set	No. of Features	Classifier Used	Accuracy (%)
Hangarge et.al[10]	Gabor Features	24	KNN for K=1	91.99
Proposed Method	Curvelet Features	20	Nearest Neighbor	95.24

6. CONCLUSION

In this paper, we have presented a technique based on discrete curvelet transform to identify the script of the handwritten text blocks. Typically it is true that all Indian script character shapes are curvilinear in nature. Therefore we have exploited these properties using discrete curvelet transform. Exhaustive experimentations are carried out on various combinations of scripts and noticed comparable performance in all the cases. Further, we are extending this method to various Asian handwritten and printed scripts in future.

7. REFERENCES

[1] Judith Hochberg, Patrick Kelly, Timothy Thomas, Lila Kerns, "Automatic Script Identification From Document Images Using Cluster-Based Templates", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 19, No. 2, February 1997.

[2] Sharmila S., Abirami S., Murukappan S., Bhaskaran R. "Design and Development of a Script Recognition Tool for Indian Document Images" , IJIDCS, Vol. 2 No. 1, 2012.

[3] Basavaraj Patil, N. V. Subbareddy, "Neural network based system for script identification in Indian documents", Sadhana Vol. 27, Part 1, February 2002, pp. 83–97.

[4] U. Pal, B. B. Chaudhuri, "Indian script character recognition: A Survey", Pattern Recognition, Vol. 37, Issue 9, September 2004, pp. 1887-1899.

[5] B.V.Dhandra., Mallikarjun Hangarge, "Offline Handwritten Script Identification in Document Images", IJCA (0975-8887), Vol. 4, No. 6, July 2010.

[6] Peeta Basa Pati, A.G. Ramakrishnan, "Word level multi-script identification", Pattern Recognition Letters, Vol. 29, Issue 9, 1 July 2008, pp. 1218-1229.

[7] Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy, "Script Identification from Indian Documents", DAS 2006, LNCS 3872, pp.255-267.

[8] D Dhanya, A G Ramakrishna, Peeta Basa Pati, "Script identification in printed bilingual documents", Sadhana Vol. 27, Part I, February 2002, pp. 73-82.

[9] G. G. Rajput, Anita H. B, "Handwritten Script Recognition using DCT an Wavelet Features at Block Level", IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, RTIPPR, 2010.

[10] Mallikarjun Hangarge, Gururaj Mukarambi, B. V. Dhandra, "South Indian Ha.ndwritten Script Identification at Block Level from Trilingual Script Document Based on Gabor Features", Multimedia Processing, Communicating and Computing Applications, Lecture Notes in Electrical Engineering 213.

[11] E. J. Candès, L. Demanet, D. L. Donoho, L. Ying. "Fast discrete curvelet transforms". *Multiscale Model. Simul.*, **5** 861-899.

[12] M.J. Fadili , J.L. Starck, "Curvelets and Ridgelets", October 24, Encyclopedia of Complexity and Systems, 2007.

[13] Lindsay Semler, Lucia Dettori, "Curvelet-Based Texture Classification Of Tissues In Computed Tomography", IEEE International Conference on Image Processing, 2006, pp. 2165 – 2168.

[14] Hangarge, M., Santosh ,K.C., P., Rajmohan: Directional Discrete Cosine Transform for Handwritten Script Identification. In: Proc. of ICDAR pp. 344{348 (2013)