A Multiple Feature based Novel Approach for Identification of Printed Indian Scripts at Word Level

Gopal Prasad
ECE Department
K. P. Engineering College
Agra, UP, India

Atul Kumar Singh
ECE Department
Kamla Nehru Inst. of Technology
Sultanpur, UP, India

Pawan Kumar
ECE Department
G. L. Bajaj Inst. of Technology & Mgt.
Greater Noida, UP, India

ABSTRACT
In a country like India where different scripts are in use, automatic identification of printed script facilitates many important applications such as automatic transcription of multilingual documents and for the selection of script specific OCR in a multilingual environment. In this paper a novel method to identify the script type of the collection of documents printed in seven Indian languages at word level is proposed. These languages are Bangla, Hindi, English, Malayalam, Oriya, Tamil and Kannada. The recognition is based upon multiple features extracted using Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Script classification performance is analyzed using the K-nearest neighbor classifier by comparing the majority of voting's between the outputs of DCT and DWT based methods. The proposed scheme utilizes the strength of both the DCT and DWT based features. The results of experimentation found the overall accuracy to be 98.11 % which show the superiority of the proposed multiple features based scheme over several existing schemes of script identification.

General Terms
Script Identification, Image Preprocessing, OCR.

Keywords
Multilingual Document Images, Multi-scripts Images, Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Standard Deviation, K-NN classifier.

1. INTRODUCTION
India is a multi-script multi-lingual country which has more than 18 regional languages derived from 12 different scripts. Most of the states in India have more than one language of communication. Therefore many official documents are multi-script in nature. Identification of the script from a multi-scripts document is one of the challenging tasks for a developer of an OCR system. Script identification makes the task of analysis and recognition of the text easier by suitably selecting the modalities of OCR. A few results have already been reported in the literature, identifying the scripts in a multi-lingual and multi-script document dealing with Roman and other oriental scripts such as Chinese, Korean and Japanese. A few attempts have already been made to isolate and identify the scripts of the texts in the case of multi-script Indian documents also. Existing script identification techniques mainly depend on various features extracted from document images at character, word, text line or block level.

Many commercial OCR systems are now available in the market that work for Roman, Chinese, Japanese and Arabic characters. Multilingual document recognition technology and its applications in China, which is useful for building multilingual digital library, are reported in [1]. A survey of offline cursive script word recognition is presented in [2]. The survey is classified into three categories such as segmentation free methods, segmentation based methods and the perception oriented approach. Most of this survey focuses on the algorithms that were proposed in order to realize the recognition phase. Chain code based representation and manipulation of hand written images is reported in [3].

Although there are twelve major scripts in India and the multi-script multilingual documents are quite common in Indian environment. A few schemes of script identification for Indian scripts are also reported in literature. A review of the OCR work done on Indian language scripts is reported in [4]. A method as presented in [5] presents an automatic technique for the identification of printed Roman, Chinese, Arabic, Devnagari and Bangla text lines from a single document with an overall accuracy of 97.33%. Shape based features, statistical features and some features obtained from the concept of water reservoir are used for script identification. Experimental results show an average script-line identification accuracy of 97.52%. Two different approaches have been proposed in [6] for script identification at the word level, from a bilingual document containing Roman and Tamil scripts. In the first approach, words are divided into three distinct spatial zones. The spatial spread of a word in upper and lower zones, together with the character density, is used to identify the script. The second approach analyses the directional energy distribution of a word using Gabor filters with suitable frequencies and orientations. The work presented in [7] describes another method for identification and separation of text words of Kannada, Devanagri, and Roman scripts using discriminating features with an accuracy of 96.7%. A hierarchical classification based scheme is reported in [8] which use features consistent with human perception for script identification from Indian document. In [9], effectiveness of Gabor and discrete cosine transform (DCT) features for word level multi-script identification has been independently evaluated using nearest neighbor, linear discriminant and support vector machine (SVM) classifiers. A Gabor function based multichannel directional filtering approach for both text area separation and script identification at the word level is reported in [10] and an overall classification accuracy of 97.11% was achieved. Some background information about the past researches on both global based approach as well as local based approach for script identification in document images is reported in [11]. Both the systems can perform script/language identification in document images at document, line and word level. Thus, all the reported studies accomplish script recognition either at the line level or at the word level.

In this proposed work a multiple feature based approach that combines discrete cosine transform (DCT) and discrete wavelet transform (DWT) based frequency information for seven Indian scripts including Roman script is presented. The classification is done using k-nearest neighbor (K-NN) classifier by using majority voting between DCT and DWT. The experiments are carried out on the sample document

images at word level. The rest of the paper is divided into various sections. Section 2 presents the proposed approach. Section 3 gives experimental results and analysis. Section 4 summarizes the conclusions followed by section 5 of references.

2. PROPOSED ALGORITHM

2.1 Data Collection and Preprocessing

At present, in India, standard databases of printed Indian scripts are not available. Hence, data for training and testing the classification scheme was collected from different sources like newspaper, websites, Google images, Google translate, from different books etc. A block of image of size 512 x 512 pixels is then extracted manually from different areas of the document image. It should be noted that the text block may contain two or more lines. Numerals that may appear in the text are not considered. It is ensured that at least 50% of the text block region contains text. These blocks representing a portion of the printed document are then binarized using Otsu's method [12] so that text represents value 1 and background represents value 0. A total of 700 printed image blocks are created, 100 blocks for each of the scripts. A sample of blocks representing different scripts is shown in figure 1.

(a) Bangla (b) English (c) Kannada

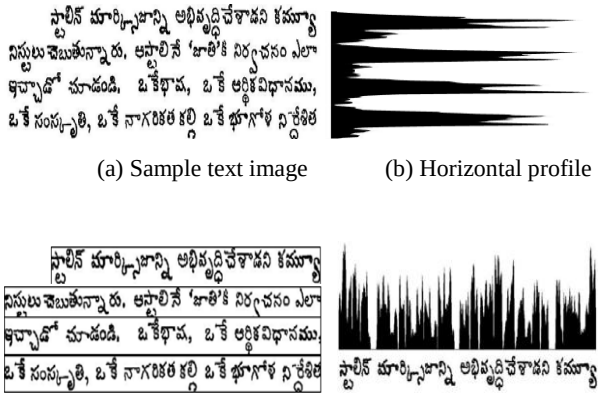
(d) Orriya (e) Malayalam (f) Hindi

(g) Tamil

Figure 1: Blocks of Indian scripts

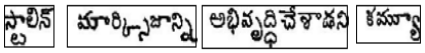
2.2 Line and Word Segmentation

To segment the document image into several text lines, the valleys of the horizontal projection computed by a row-wise sum of black pixels is utilized. The position between two consecutive horizontal projections where the histogram height is least denotes a boundary line. Using these boundary lines, document image is segmented into several text lines. Similarly, to segment each text line into several text words, the valleys of the vertical projection of each text line obtained by computing the column-wise sum of black pixels is utilized. The position between two consecutive vertical projections where the histogram height is least denotes a boundary line. Using these boundary lines, every text line is segmented into several text words. Figure 2 shows line and word segmentation. Each word is then normalized to keep constant aspect ratio.



(a) Sample text image (b) Horizontal profile

(c) Extracted text lines (d) Vertical profile



(e) Extracted words

Figure 2: Line and Word Segmentation

2.3 DCT based Feature Extraction

The discrete cosine transform (DCT) concentrates energy into lower order coefficients. The DCT is purely real. The DCT expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies that are necessary to preserve the most important features. With an input image A_{mn} the DCT coefficients for the transformed output image B_{pq} are computed according to equation 2. In the equation A is the input image having M×N pixels, A_{mn} is the intensity of the pixel in row m and column n of the image and, B_{pq} is the DCT coefficient in row p and column q of the DCT matrix.

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le p \le M-1 \end{cases}$$

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le q \le N-1 \end{cases} \quad (1)$$

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)q}{2N}$$

$$0 \le p \le M-1$$

$$0 \le q \le N-1 \quad (2)$$

The steps involved in DCT feature extraction as shown in figure 3 are,

a) Calculate discrete cosine transform (DCT) of each word I_c.

b) Perform zigzag operation on the DCT coefficients I_{dct}. The zigzag matrix I_z is a row vector matrix containing high frequency coefficients in its first 100 values that contain 95% word information. This forms features vectors f_{dct} for each words of different scripts.

c) These features vectors f_{dct} are stored in a feature library and used later in the testing stage.

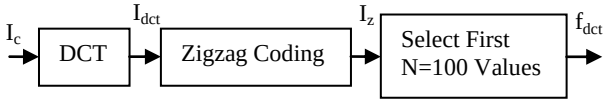


Figure 3: DCT based Feature Extraction

2.4 DWT based Feature Extraction

The discrete wavelet transform (DWT), which is based on sub-band coding is found to yield fast computation of wavelet transform [13, 14]. It is easy to implement and reduces the computation time and resources required. The wavelet transforms are used to analyze the signal (image) at different frequencies with different resolutions. It represents the same signal, but corresponding to different frequency bands. Wavelets are used for multi resolution analysis, to analyze the signal at different frequencies with different resolutions, to split up the signal into a bunch of signals, representing the same signal, but all corresponding to different frequency bands, and provides what frequency bands exist at what time intervals. Many wavelet families have been developed with different properties [14]. For 2-D images, applying DWT corresponds to processing the image by 2-D filters in each dimension. The filters divide the input image into four non-overlapping multi-resolution sub-bands LL, LH, HL and HH. The sub-band LL represents the coarse-scale DWT coefficients while the sub-bands LH, HL and HH represent the fine-scale of DWT coefficients. Daubechies 10 (Db10) wavelet is employed since it yielded better results. Decomposition is carried out for one level. The Daubechies wavelet is a wavelet used to convolve image data. The wavelets can be orthogonal, when the scaling functions have the same number of coefficients.

The steps involved in DWT feature extraction as shown in figure 4 are,

a) Calculate Discrete Wavelet Transform (DWT) with Db10 wavelet for each word I_c.

b) Compute the Standard Deviation for each sub-band (cA, cH, cV, cD) separately. This forms 4 features σ1, σ2, σ3, σ4.

c) Store all the computed features in a vector to form feature vector f_{dwt}. For each word of different scripts, features vector is formed.

d) These features vectors are stored in a feature library and used later at the testing stage.

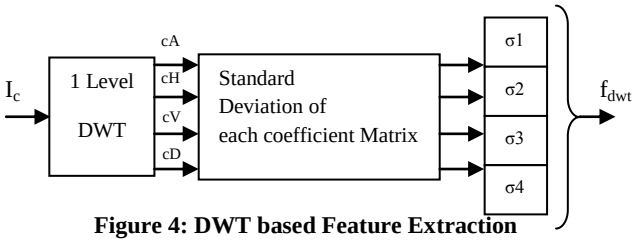


Figure 4: DWT based Feature Extraction

2.5 Script Recognition with Voting Scheme

In the proposed model, K -nearest neighbor classifier is used to classify the test samples. The features are extracted from the test image using the proposed feature extraction algorithm and then compared with corresponding feature values stored in the feature library using the Euclidean distance formula given in equation 3 as,

$$D(M) = \sqrt{\sum_{j=1}^{N}[f_j(x) - f_j(M)]^2} \quad (3)$$

Where N is the number of features in the feature vector f, $f_j(x)$ represents the j_{th} feature of the test sample and $f_j(M)$ represents the j_{th} feature of M^{th} class in the feature library. Then, the test sample is classified using the k-nearest neighbor (K-NN) classifier. In the K -NN classifier, a test sample is classified by a majority vote of its k neighbors, where K is a positive integer, typically small. If K =1, then the sample is just assigned the class of its nearest neighbor.

It is better to choose K to be an odd number to avoid tied votes. So, in this method, the K=5 nearest neighbors are determined and the test image is classified as the script type of the majority of the votes between DCT based decision and Wavelet based decision.

3. RESULTS AND ANALYSIS

3.1 Database Creation

The performance of the described multi-script identification system on a dataset of 10500 pre-processed word images is evaluated. The complete dataset is manually processed to generate the ground truth for testing and evaluation of the algorithm. Test sample block is the input to our system and performance is noted in terms of recognition accuracy. 50 blocks (approx 1500 words) of each script shown in figure1 are used for training purpose and 20 blocks (approx 500 words) of each script are used as a test dataset.

3.2 Training

The steps involved in training phase as shown in figure 5 are,

a) Preprocessing of the document image I_d.

b) Perform line and word segmentation of the preprocessed image.

c) Calculate Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) of each extracted character I_c. This forms DCT and DWT based features vectors f_{dct} and f_{dwt}.

d) These features vectors f_v are stored in a feature library and used later in the testing stage.

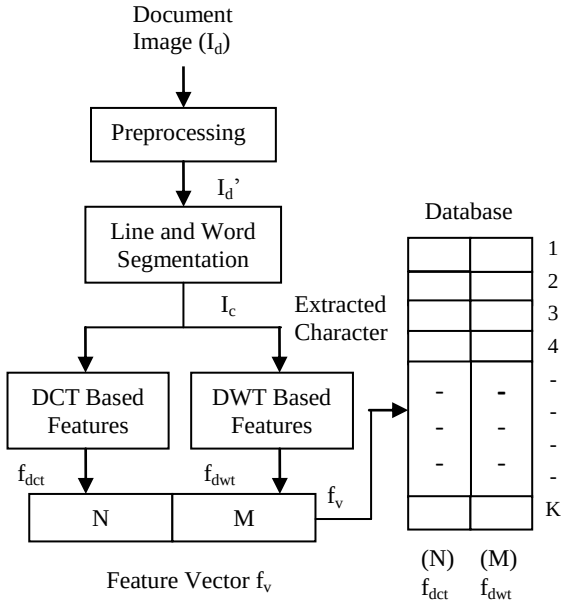


Figure 5: Training phase

3.3 Testing

The steps involved in training phase as shown in figure 6 are,

a) Preprocessing of the document query image I_q.

b) Perform line and word segmentation of the preprocessed image.

c) Calculate Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) of each extracted character I_c. This forms DCT and DWT based features vectors f_{dct} and f_{dwt}.

d) Now in both cases (DCT and DWT) separately, compare the feature values of the test image with the feature values stored in the database using K-nearest neighbor classifier.

e) Classify the script type I_{cd} of the document query image by calculating the majority of the votes between DCT based decision and Wavelet based decision.

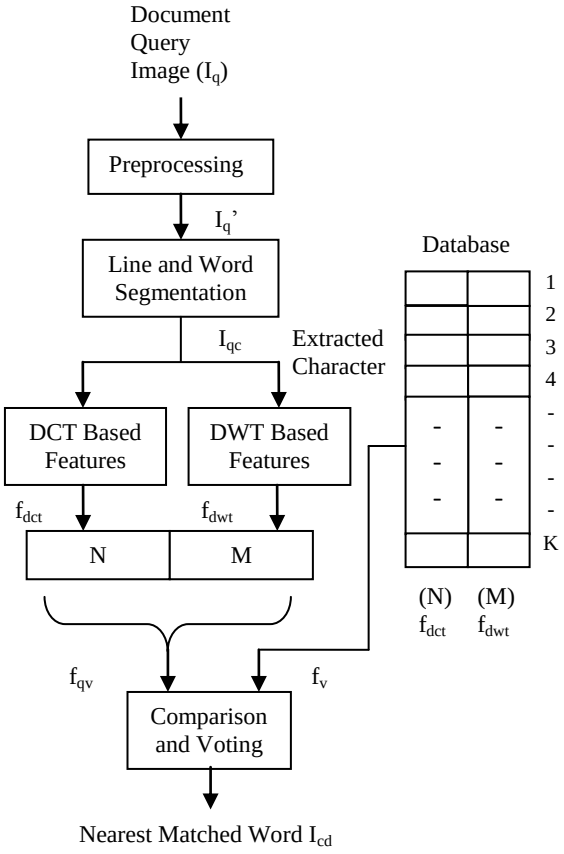
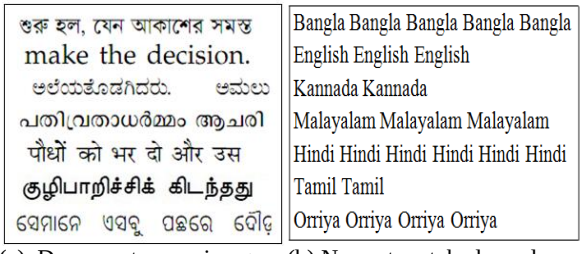


Figure 6: Testing phase

The method was implemented using Matlab 7.8 software. Figure 7 shows the document query image and the corresponding script outputs. The results of testing documents images in the form of confusion metric are tabulated in Table 1. The results clearly show that the combined features that constitute DCT and wavelets yield better results. The recognition accuracy of multi-scripts is presented in figure 8.



(a) Document query image (b) Nearest matched word

Figure 7: Multi-scripts query image with its recognition

Table 1: Percentage of Recognition of 7 Indian scripts (Ban=Bangla, Ori=Oriya, Eng=English, Mal=Malayalam, Hin=Hindi, Kan=Kannada and Tam=Tamil)

Input Script	Total Words	Recognized Script							% Accuracy	% Error
		Ban	Ori	Mal	Hin	Eng	Kan	Tam		
Ban	519	505	1	-	7	4	2	-	97.3	2.7
Ori	516	2	513	-	-	1	-	-	99.4	0.6
Mal	506	-	4	493	1	6	1	1	97.4	2.6
Hin	508	7	1	-	497	2	1	-	97.8	2.2
Eng	513	-	4	7	-	502	-	-	97.8	2.2
Kan	506	-	-	1	-	-	505	-	99.8	0.2
Tam	518	1	3	1	2	6	1	504	97.3	2.7
Average %									98.11	1.89

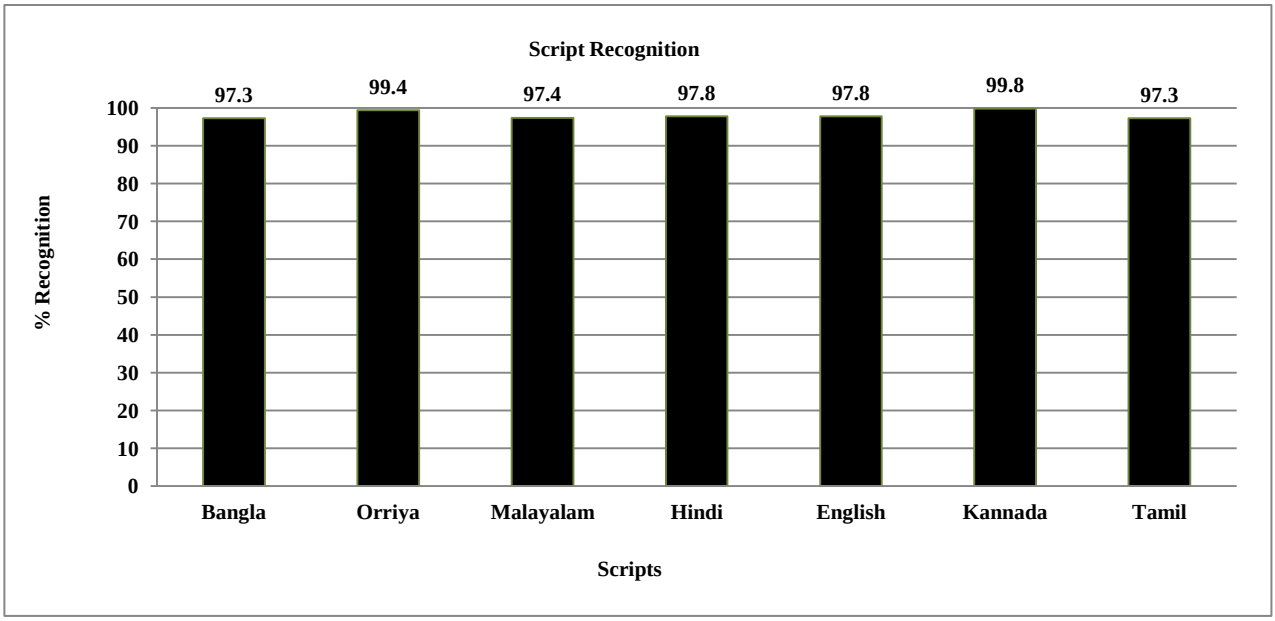


Figure 8: Recognition accuracy of multi-scripts test images

4. CONCLUSIONS

Script Identification plays an impartment role in OCR application, application forms, examination papers, government reports etc. Before further processing in OCR, it is necessary to identify the type of the scripts followed by text recognition. In this paper a multiple feature based approach is presented which combines the best results of DCT based approach and DWT based approach. Identification is carried out at the word level. The results of section 3 show that Kannada and Orriya scripts are identified with an accuracy of 99.80 % and 99.40 % respectively. The overall accuracy of seven major Indian scripts is found to be 98.11%. The experimental results demonstrate that the multiple feature based approach performs better in classifying the different Indian scripts. This multiple feature based approach can also used for the classification of different Indian scripts.

5. REFERENCES

[1] J.S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition," *Neurocomputing—Algorithms, Architectures and Applications*, F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989.

[2] W.-K. Chen, *Linear Networks and Systems.* Belmont, Calif.: Wadsworth, pp. 123-135, 1993.

[3] H. Poor, "A Hypertext History of Multiuser Dimensions," *MUD History*, http://www.ccs.neu.edu/home/pb/mud-history.html. 1986.

[4] K. Elissa, "An Overview of Decision Theory," unpublished.

[5] R. Nicole, "The Last Word on Decision Theory," *J. Computer Vision*, submitted for publication.

[6] C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992.

[7] D.S. Coming and O.G. Staadt, "Velocity-Aligned Discrete Oriented Polytopes for Dynamic Collision Detection," *IEEE Trans. Visualization and Computer Graphics*, vol. 14, no. 1, pp. 1-12, Jan/Feb 2008, doi:10.1109/TVCG.2007.70405.

[8] S.P. Bingulac, "On the Compatibility of Adaptive Controllers," *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory*, pp. 8-16, 1994.

[9] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representation," *Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS '07)*, pp. 57-64, Apr. 2007, doi:10.1109/SCIS.2007.367670.

[10] J. Williams, "Narrow-Band Analyzer," PhD dissertation, Dept. of Electrical Eng., Harvard Univ., Cambridge, Mass., 1993.

[11] E.E. Reber, R.L. Michell, and C.J. Carter, "Oxygen Absorption in the Earth's Atmosphere," Technical Report TR-0200 (420-46)-3, Aerospace Corp., Los Angeles, Calif., Nov. 1988.

[12] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification*, vol. 2, no. 4, pp. 193-218, Apr. 1985.

[13] R.J. Vidmar, "On the Use of Atmospheric Plasmas as Electromagnetic Reflectors," *IEEE Trans. Plasma Science*, vol. 21, no. 3, pp. 876-880, available at http://www.halcyon.com/pub/journals/21ps03-vidmar, Aug. 1992.

[14] J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746.