Handwritten Script Recognition using DCT and Wavelet Features at Block Level

G. G. Rajput
Department of Computer Science,
Gulbarga University, Gulbarga-585106
Karnataka, India

Anita H. B.
Department of Computer Science,
Gulbarga University, Gulbarga-585106
Karnataka, India

ABSTRACT
In a country like India where different scripts are in use, automatic identification of handwritten script facilitates many important applications such as automatic transcription of multilingual documents and for the selection of script specific OCR in a multilingual environment. Existing script identification techniques depend on various features extracted from document images at character, word, text line or block level. In this paper, we propose a novel method towards multi-script identification at block level. The recognition is based upon features extracted using Discrete Cosine Transform (DCT) and Wavelets of Daubechies family. The proposed method is experimented on handwritten documents of eight Indian scripts that include English script and yielded encouraging results.

General Terms
Script identification, Wavelets, multi-script identification

Keywords
Multi-script documents, handwritten script, Discrete Cosine Transform, Wavelets, K-NN classifier.

1. INTRODUCTION
A document containing text information in more than one script is called a multi-script document. An automatic script identification technique is useful to sort document images, select appropriate script-specific OCRs and search online archives of document images for those containing a particular script. Handwritten script identification is a complex task due to following reasons; complexity in pre-processing, complexity in feature extraction and classification, sensitivity of the scheme to the variation in handwritten text in document (font style, font size and document skew), and performance of the scheme. Existing script identification techniques mainly depend on various features extracted from document images at character, word, text line or block level. A brief review of different methods available in the literature is given below.

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 application in China, which is useful for building multilingual digital library, is reported in [1]. A survey of offline cursive script word recognition is presented in [2]. The survey is classified into three categories: 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, there are no sufficient numbers of papers on Indian printed/handwritten script /language recognition. A review of the OCR work done on Indian language scripts is reported in [4]. Pal and Choudhary [5] have presented an automatic technique for the identification of printed Roman, Chinese, Arabic, Devnagari and Bangla text lines from a single document. Shape based features, statistical features and some features obtained from the concept of water reservoir are used for script identification. Two different approaches have been proposed by Dhanya et al.[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. Padma and Nagbhushan [7] have described a method for identification and separation of text words of Kannada, Devanagri, and Roman scripts using discriminating features. Gopal Datt Joshi et. al. [8] have proposed hierarchical classification scheme which uses 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]. 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. However, these components are applicable only after line and word (LWC) segmentation of the underlying document image [12]. In contrast, the method proposed in this paper employs analysis of regions (block) comprising at least two lines and hence does not require

fine segmentation. Consequently, the script classification task is simplified and performed faster.

Block level script identification identifies the script of the given document in a mixture of various script documents. Input document to the block level identification is a mono script document. Very few publications are found in the literature for differentiating the Indian handwritten/printed scripts at block level. Dhandra et. al. [13] have proposed script identification method at block level by extracting the features in two stages. In the first stage, the morphological erosion and opening by reconstruction is carried out on a document image in horizontal, vertical, left and right diagonal directions. In the second stage, average pixel distribution is found in these directions. The classification is done using nearest neighbor classifier. The experiments are performed on Kannada, Urdu, English, and Devanagari scripts by considering the block size of 128 x 128 pixels.

Motivation for our work is as follows. First, many of the Indian documents, handwritten or machine printed, contain three scripts, namely, the state's official language (local script), Hindi and English. In [14], tri-script identification from an Indian language document has been considered. However, the literature on this problem of tri-script identification is very less. Second, many of Indian handwritten scripts have descenders and ascenders, and due to variability of handwriting in the documents the methods proposed may fail to recognise the script accurately. In many cases the most distinguished information is hidden in the frequency content of the signal rather than in the time domain. Hence, in this paper, we present a multiple feature based approach that combines DCT and Wavelet based frequency contents for eight Indian scripts including English. The classification is done using k-nearest neighbour (K-NN) classifier. The experiments are carried out on the sample document images at block level. The method proposed in this paper is successfully applied to the documents containing bi-scripts and is reported in [15]. The rest of the paper is described as follows. Data collection and pre-processing is described in section 2. A brief description of the properties of the scripts considered is given in section 2. Description of the proposed method is presented in section 3. Section 4 describes experimental results and some conclusions are given in section 5.

2. PROPERTIES OF SEVEN SCRIPTS

A brief description of the properties of the scripts considered in our study is given below. All these scripts are written from left to right.

1) **English**: The modern English alphabet is a Latin-based alphabet consisting of 26 letters each of upper and lower case characters. In addition, there are some special symbols and numerals. The letters A, E, I, O, U are considered vowel letters, since (except when silent) they represent vowels; the remaining letters are considered consonant letters, since when not silent they generally represent consonants. However, Y commonly represents vowels as well as a consonant, as very rarely does W The structure of the English alphabet contains more vertical and slant strokes.
2) **Hindi**: Hindi is derived from Devanagari script. Devanāgarī alphabet descended from the Brahmi script sometime around the 11th century AD. It was originally developed to write Sanskrit but was later adapted to write many other languages. Type of writing system is alphasyllabary / abugida. The script has 12 vowels and 34 consonants. Consonant letters carry an inherent vowel which can be altered or muted by means of diacritics or *matra*. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. This feature is common to most of the alphabets of South and South East Asia. When consonants occur together in clusters, special conjunct letters are used. Devanagari script is used to write the languages Bhojpuri, Hindi, Marathi, Mundari, Nepali, Pali, Sanskrit, Sindhi and many more including Hindi. Devanagari is recognizable by a distinctive horizontal line running along the tops of the letters that links them together.
3) **Gujarati**: The Gujarati script is one of the modern scripts of India, and is derived from the Devanagari script during the 16th century CE. The major difference between Gujarati and Devanagari is the lack of the top horizontal bar in Gujarati. Otherwise the two scripts are fairly similar. Gujarātī is a syllabic alphabet in which all consonants have an inherent vowel. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. Gujarati character set provides 14 vowels and 34 (+2 compound -ksha, gna) consonants.
4) **Punjabi**: The Gurmukhi (Punjabi) alphabet was devised during the 16th century and is modeled on the Landa alphabet. This is a syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel. Modern Gurmukhi has forty-one consonants, nine vowel symbols, two symbols for nasal sounds, and one symbol which duplicates the sound of any consonant. In addition, four conjuncts are used.
5) **Telagu**: The origins of the Telugu alphabet can be traced by to the Brahmi alphabet of ancient India, which developed into an alphabet used for both Telugu and Kannada, which in turn split into two separate alphabets between the 12th and 15th centuries AD. The writing system is syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel and consists of sequences of simple and/or complex characters. The overall pattern consists of 60 symbols, of which 16 are vowels, 3 vowel modifiers, and 41 consonants.
6) **Kannada**: The earliest inscriptional records in Kannada are from the 6th century. Kannada script is closely akin to Telugu script in origin. Under the influence of Christian missionary organizations, Kannada and Telugu scripts were standardized at the beginning of the 19th century. Writing system is alphasyllabary in which all consonants have an inherent vowel. Other vowels are indicated with diacritics, which can appear above, below, before or after the

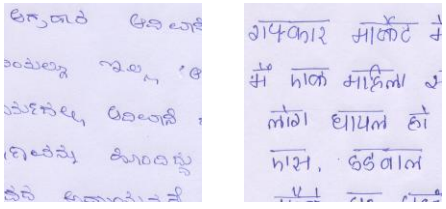
consonants. Kannada has 16 vowels and 34 consonants. There are about 250 basic, modified and compound character shapes in Kannada.

7) **Tamil**: The earlier Tamil inscriptions were written in braahmi, grantha and vaTTezuttu scripts. The Tamil script is partially "alphabetic" and partially syllable-based. Writing system of Tamil is syllabic alphabet. There are twelve vowels and eighteen consonants. Consonants are made up of six surds and their corresponding six sonants and six medials. Combinations of consonants with vowels give rise to new symbols or result in modified symbols.
8) **Malayalam**: Malayalam belongs to the southern group of Dravidian languages along with Tamil, Kota, Kodagu and Kannada. It has high affinity towards Tamil. In the early thirteenth century the Malayalam script developed from a script known as vattezhuthu (round writing), a descendant of the Brahmi script. This is a syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel. The modern Malayalam alphabet has 13 vowel letters, 36 consonant letters, and a few other symbols.

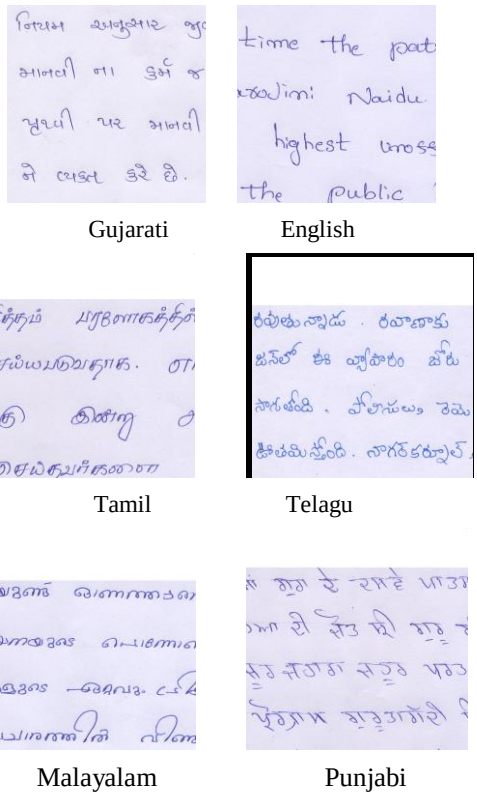
3. METHOD DESCRIPTION

3.1 Data collection and Preprocessing

At present, in India, standard databases of handwritten Indian scripts are not available. Hence, data for training and testing the classification scheme was collected from different sources. Handwritten documents belonging to English, Devnagari, Kannada, Tamil, Bangla, Telagu, Punjabi, and Malayalam scripts are collected from different persons belonging to different professions. The documents are scanned at 300 dpi and stored as gray scale images. 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 handwritten text block may contain two or more lines with different font sizes and variable spaces between lines, words and characters. Numerals that may appear in the text are not considered. We do not perform any processing to homogenise these parameters. It is ensured that at least 50% of the text block region contains text. These blocks representing a portion of the handwritten document are then binarized using Ostu's method [14, 15] so that text represents value 1 and background represents value 0. The salt and pepper noise around the boundary is removed using morphological opening. This operation also removes discontinuity at pixel level. A total of 800 handwritten image blocks are created, 100 blocks for each of the scrips. A sample of blocks representing different scripts is shown in Figure 1.



Kannada Hindi



Gujarati English

Tamil Telagu

Malayalam Punjabi

Figure 1. Sample images of handwritten document images in different scripts.

3.2 Feature Extraction

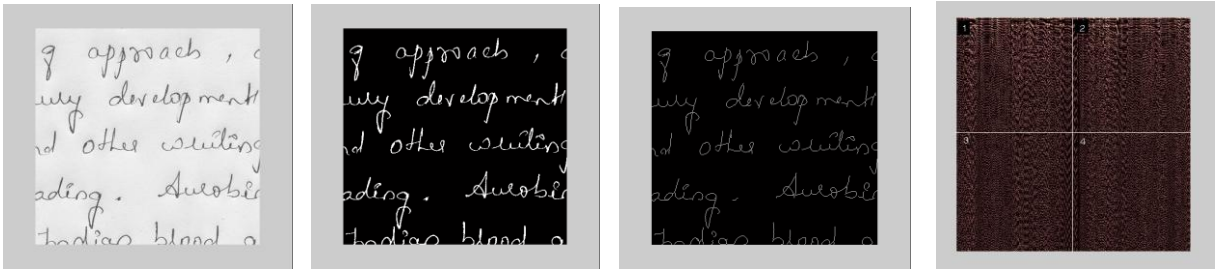
Features are the representative measures of a signal, which distinguish it from other signals. The selected features should maximize the distinction between English, Devnagari and local official language scripts. Features are extracted by transforming the image in time domain to the image in frequency domain. The term frequency refers to variation in brightness or color across the image, i.e. it is a function of spatial coordinates, rather than time. The frequency information of image is needed to see information that is not obvious in time-domain. A brief description of the features is given below.

Cosine transforms: 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 [15]. With an input image, Amn, the DCT coefficients for the transformed output image, B_{pq}, are computed according to equation shown below. In the equation, A, is the input image having M-by-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.

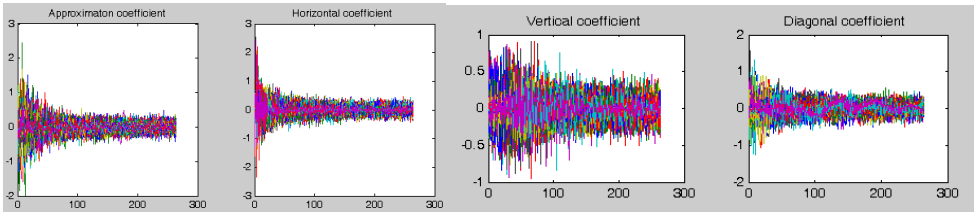
$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)q}{2N}, \quad \begin{matrix} 0 \le p \le M-1 \\ 0 \le q \le N-1 \end{matrix}$$

$$\alpha_p = \begin{cases} 1/\sqrt{M}, & p = 0 \\ \sqrt{2/M}, & 1 \le p \le M-1 \end{cases} \quad \alpha_q = \begin{cases} 1/\sqrt{N}, & q = 0 \\ \sqrt{2/N}, & 1 \le q \le N-1 \end{cases}$$

Wavelet Transforms: The discrete wavelet transform (DWT), which is based on sub-band coding is found to yield fast computation of wavelet transform [15,16]. 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



(a) original image (b) binarized image (c) thinned image (d) DCT image (shown in pink color)



(e) DWT coefficients

Figure 2. Pipeline for feature extraction

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 [16]. 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. We have used Daubechies 9 since it yielded better results. We have carried out the decomposition 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 feature extraction algorithm is described in following steps.

Input: Image block of size 512 x 512 pixels in grey scale.

Output: Feature vector computed by performing DCT and DWT.

Method:

1. Binarize the image using Otsu method to yield text representing binary 1 and background binary 0.
2. Remove small objects around the boundary using morphological opening.
3. Apply thinning operation.
4. Apply DCT and divide the magnitude (image) of DCT into 4 equal non-overlapping block and extract the local features by computing the Standard Deviation for the first and second block. This forms 2 features.
5. Perform Wavelet (Daubechies 9) decomposition for the magnitude (image) of DCT to obtain approximation coefficients (cA), vertical coefficients(cV), horizontal coefficients(cH), and diagonal coefficients(cD)
6. Compute the Standard Deviation for each frequency (cA, cV, and cH) bands separately. This forms 3 features.
7. Store all the computed features in a vector.

Figure 1 shows the pipeline for feature extraction

3.3 Script Recognition

KNN classifier is adopted for recognition purpose. This method is well-known non-parametric classifier, where posterior probability is estimated from the frequency of nearest neighbours of the unknown pattern. The key idea behind k-nearest neighbor classification is that similar observations belong to similar classes. The test image is classified to a class, to which its k-nearest neighbor belongs to. Feature vectors stored priori are used to decide the nearest neighbor of the given test image feature vector. The recognition process is described below.
During the training phase, features are extracted from the training set by performing algorithm given in section 3.2. These features are input to K-NN classifier to form a knowledge base that is subsequently used to classify the test images. During test phase, the test image which is to be recognized is processed in a similar way as described in section 3.2 and features are computed performing the algorithm described in section 3.2. The classifier computes the Euclidean distances between the test feature vector with that of the stored features and identifies the k-nearest neighbor. Finally, the classifier assigns the test image to a class that has the minimum distance with voting majority. The corresponding script is declared as recognized script.

4. EXPERIMENTAL RESULTS

We evaluate the performance of the described multi-script identification system on a dataset of 800 pre-processed images. The complete dataset is manually processed, as described in section 3.2, to generate the ground truth for testing and evaluation of the algorithm. For evaluation, we consider English script, Hindi script, and a local language script. Samples of one script are input to our system and performance is noted in terms of recognition accuracy. 60 blocks are used for training purpose and remaining 40 blocks are used as a test dataset. Identification of the test script is done using K-NN classifier. The results were found to be optimal for k=1 using feature size of 5. The method was implemented using Matlab 6.1 software. The results of all the tri-script considered for testing are tabulated in Table 1. The results clearly show that the combined features that constitute DCT and wavelets yield better results. The results are promising for groups 1, 2, 4, and 6. The performance of the system is on lower side for groups 3 and 5. Table 2 presents the details of recognition for each group in the form of confusion matrices. The recognition accuracy of each script, in each of the groups, is presented in Figure3.

Table 1: Recognition results for tri-scripts

Tri-script Group	Tr-scripts	Recognition %
1	Kannada, English and Hindi	98%
2	Malayalam, English and Hindi	99.2%
3	Punjabi, English and Hindi	93%
4	Tamil, English and Hindi	99.2%
5	Gujarati, English and Hindi	90%
6	Telagu, English and Hindi	99%

Table 2: Recognition results of tri-scripts in the form of confusion matrices

Script	Kannada	English	Hindi
Kannada	37	3	0
English	2	38	0
Hindi	0	0	40

Script	Punjabi	English	Hindi
Punjabi	38	0	2
English	0	40	0
Hindi	5	0	35

Script	Malayalam	English	Hindi
Malayalam	39	1	0
English	0	40	0
Hindi	0	0	40

Script	Tamil	English	Hindi
Tamil	40	0	0
English	1	39	0
Hindi	0	0	40

Script	Gujarati	English	Hindi
Gujarati	35	5	0
English	7	33	0
Hindi	0	0	40

Script	Telagu	English	Hindi
Telagu	40	0	0
English	2	38	0
Hindi	0	0	40

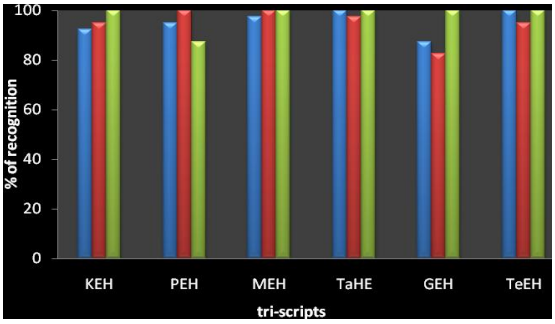


Figure 3. Recognition accuracy of scripts in each tri-script group

4. CONCLUSION

The discrete cosine transform (DCT) and discrete wavelet transform (DWT) have been applied successfully in script recognition. In this paper, we have presented a combined DCT-DWT approach for tri-script identification at block level for the handwritten documents. KNN classifier is used in recognition phase that yielded better results for k=1. The proposed method is robust and independent of style of hand writing. Overall, it can be successfully used for identifying the scripts. The proposed method can be extended to other scripts.

5. ACKNOWLEDGMENTS

The authors are thankful to the referees for their critical comments. We also thank Dr. P. S. Hiremath, Professor, Department of Computer Science, Gulbarga University, Gulbarga, for his helpful suggestions.

6. REFERENCES

[1] Liangrui Peng, Changsong Liu, Xiaoqing Ding, Hua Wang, "Multilingual document recognition research and its application in China," dial, pp.126-132, Second International Conference on Document Image Analysis for Libraries (DIAL'06), 2006.

[2] Offline cursive script word recognition: a survey, Ta Steinherz, Ehud Rivlin, Nathan Intrator, International journal on Document Analysis and Recognition, IJDAR (1999) 2: 90-110.

[3] S Madhvanath, G Kim, V Govindaraju, .Chaincode Contour Processing for Handwritten Word Recognition. IEEE transactions on pattern analysis and machine intelligence, 1996, Vol 21 , No 9, pg 928 . 932.

[4] Indian script character recognition: a survey Pattern Recognition, Volume 37, Issue 9, September 2004, Pages 1887-1899, U. Pal and B. B. Chaudhuri.

[5] U. Pal and B. Chaudhuri. Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line. In International Conference on Document Analysis and Recognition, pages 790-794, 2001

[6] Script identification in printed bilingual documents, D Dhanya, A G Ramakrishnan_ And Peeta Basa Pati, Sadhana Vol. 27, Part 1, February 2002, pp. 73–82.

[7] M. C. Padma and P. Nagbhushan, Identification and separation of text words of Kannada, Hindi, and English languages through discriminating features, Proceedings of NCDAR 2003, 252-260

[8] Script Identification from Indian Documents, Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy, DAS 2006, LNCS 3872, pp. 255–267, 20069.

[9] Peeta Basa Pati and A.G. Ramakrishnan Word level multi-script identification, Pattern Recognition Letters, Volume 29, Issue 9, 1 July 2008, Pages 1218-1229.

[10] Gabor filters for Document analysis in Indian Bilingual Documents Peeta Basa Pati S Sabari Raju Nishikanta Pati A G Ramakrishnan, ICISIP 2008, pp 123-126.

[11] A Survey of Script Identification techniques for Multi-Script Document Images, S. Abirami, Dr. D. Manjula, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[12] U. Pal, S. Sinha, and B. B. Chaudhuri, Multi-script line identification from Indian document. Seventh International Conference on Document Analysis and Recognition 2 (2003) 880–884.

[13] B. V. Dhandra. P. Nagabhushan, Mallikarjun Hangarge, Ravindra Hegdi, V. S. Malemath, Script Identification Based on Morphological Reconstruction in Document Images, Proceedings of 18th International Conference on Pattern Recognition, Honkong, 2006

[14] Peeta Basa Patil and A. G. Ramkrishnan, HVS Inspired System for Script Identification in Indian Multi-script Documents, LNCS, Volume 3872, pp 380-389, 2006

[15] Kannada, English, and Hindi Handwritten Script Recognition using multiple features, Proc. of National Seminar on Recent Trends in Image Processing and Pattern Recognition (RTIPPR-10), Editors: Dr. P. S. Hiremath et. al., Excel India Pub., New Delhi, ISBN: 93-80043-74-0, pp 149-152.

[16] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram", IEEE Tansaction Systems, Man and Cybernetics, vol 9, no.1, pp.62-66,1979.

[17] Digital Image processing, 3/e,Gonzalez and Woods, Pearson Education, 2008.

[18] I. Daubechies 1992, "Ten Lectures on Wavelets", SIAM CBMS-NSF, Series on Applied Mathematics, SIAM.

[19] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, New York.

.