

On Separation of English Numerals from Multilingual Document Images

B.V.Dhendra

P.G.Department of Studies and Research in Computer Science, Gulbarga University, Gulbarga, India
dhendra_b_v@yahoo.co.in

Mallikarjun Hangarge

P.G.Department of Studies and Research in Computer Science, Gulbarga University, Gulbarga, India
mhangarge@yahoo.co.in

Abstract— For Optical Character Recognition (OCR) of bilingual or multilingual document containing text words in regional language and numerals in English, it is necessary to identify different script forms before running an individual OCR of the scripts. In this paper, an attempt is made for separation of English numerals at word level from bilingual and trilingual documents representing Kannada, Devnagari, Tamil, Odiya and Malayalam scripts by using discriminating features such as aspect ratio, strokes densities, eccentricity, etc. as a tool. The k-nearest neighbour algorithm is used to classify the new word images and the algorithm is tested on 6000 sample words with a five fold cross validation test. The algorithm is robust with respect to font styles, sizes and noise. The results obtained are quite encouraging.

Index Terms—Script identification, OCR, morphological reconstruction, eccentricity, and cross validation

I. INTRODUCTION

A very important area in the field of document analysis is that of optical character recognition (OCR), which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form. To date, many algorithms have been presented in the literature to perform this task for a specific language, and such OCRs will not work for a document containing more than one script. Most of the work reported in the literature relates to Roman, Arabic, Chinese, and Korean and Japanese scripts. Though, some work has already been reported involving Indian scripts, the work is still in its emerging stage.

In a multi-lingual multi-script country like India, English has proven to be the binding language due to the diversity of languages and scripts. Therefore, a bilingual or trilingual document page may contain text words in regional language and numerals in English (see Fig.1).

So, multilingual OCR is needed to read these documents. To make a multilingual OCR successful, it is necessary to separate portions of different script regions of the multilingual document at word level and then identify the different script forms before running an individual OCR system. The script/language identification research work reported in the literature, can be classified into three different categories, (1) word-wise (2) text line and (3) text block script identification.

A. Word-wise Script Identification

Peeta Basa Pati *et al.* [19], used global approach based on Gabor filter bank having three different radial frequencies and six different angles of orientation with a radial frequency bandwidth of 1 octave and an angular bandwidth of 30°. They obtained a combination of 18 odd and 18 even filters with three radial frequencies and six degrees. The size of each filter mask used for experimentation is 13 x 13. Thus, a 36-dimensional feature vector of the total energy in each of the filtered

രണ്ടാഴ്ച വൈകിട്ട് ബാഗ്ദാദിലെ സന്ദേശ മാർക്കറ്റിലുണ്ടായ സ്ഫോടനത്തിൽ 132 പേർ കൊല്ലപ്പെട്ടതിനു പുറമെ 305 പേർക്ക് പരിക്കുണ്ടായി. രണ്ടു വർഷം മുമ്പ് ഹിജ്റയിൽ 125 പേർ കൊന്ന സ്ഫോടനത്തിലെ മരണസംഖ്യ മറികടന്നു സന്ദേശത്തെ വശത്തിനുശേഷം ഇറാഖിലുണ്ടായ ഏറ്റവും മാരകമായ ചാരഫോടനപരമ്പരകളിൽ ഒറ്റ ദിവസം ഇരുന്നൂറിൽപ്പരമാളുകൾ കൊന്നു.

(a)

1986 ஆம் ஆண்டு நாடாளுமன்றத்தில் நிறைவேற்றப்பட்ட தேசியக் கல்வி கொள்கையில் பொதுப்பள்ளி முறையை அடைவதற்கு அண்மைப் பள்ளி முறையை தடைமுறைப் படுத்துதல், அவசியமில்லாத, இது குறித்துக் கல்விக் கொள்கையில் உறுதியான நடவடிக்கைகள் எடுக்கப்படவேண்டும். 1988 ஆம் ஆண்டு தடுவன் கல்விக் குழு பொதுப்பள்ளி மீதான வழிகாட்டுதல் தாடு முழுவதும் பொதுப்பள்ளி முறையை அமல்படுத்தத் தேசியசெயல் திட்டத்தை வடிவமைத்துக் கொடுத்தது. 1990 இல் ஆச்சாரிய ராஜ்ரத்தி குழுவும் இதனளபே முன் வைத்தது.

(b)

ಶವೇರಿ ಸ್ವಯಮಂಶಯ: 1991 ರಂದು ಮಧ್ಯೆರ ತೀರ್ಪು ನೀಡಿದ ಸಂದರ್ಭದಲ್ಲಿ ಉಪವಿಧಾನ ಸಭೆ ಕನ್ನಡದಲ್ಲಿ ಮುಂಜಾಗತ ಕ್ರಮವಾಗಿ ಈ ಎಲ್ಲ ವ್ಯವಸ್ಥೆಗಳನ್ನು ಮಂಜೂರಿಸಿದ ಎಂದು ಅವರು ತಿಳಿಸಿದರು.

(c)

Figure 1. (a) A sample of Malayalam document containing English numerals (b) A sample of Tamil document containing English numerals and (c) A sample of Kannada document containing English numerals

This paper is an extended and modified vision of "Word Level Script Identification in Bilingual Documents through Discriminating Features" by B.V.Dhendra, Mallikarjun Hangarge, Ravindra Hegadi, and V.S.Malemath which has appeared in the proceedings of ICSCN, Feb-2007, Chennai, India and © IEEE-2007.

images is used. The Linear discriminant (LD) and nearest neighbour (NN) classifiers are used to classify the word images of five different scripts namely, Roman, Devnagari, Kannada, Tamil and Oriya in bi-script, tri-script and five-script scenarios. They used prototypes to reduce the training set to smaller size and in turn saved 87% of memory and computation time. However, this method assumes that a word should contain at least two characters. Thus, it is word size dependent and still it involves time complexity as it depends on 36-dimension feature vector for classification.

The other algorithms proposed for word level script identification are by Dhanya *et al.* [13] based on Gabor filters and spatial spread features, Pal *et al.* [15,26] based on water reservoir, conventional, topological and structural features and Padma *et al.* [16] based on discriminating features, have recognition rate of more than 95%. The recognition accuracy of these algorithms falls drastically for the words of size less than three characters. Hence, the algorithms are word size dependent. Peeta Basa Pati *et al.* [17] have proposed word level script identification for Tamil, Devnagari and Oriya scripts based on 32 features using Gabor filters. They have not reported about the performance of their algorithm for various font sizes and styles. Further, these algorithms deal with only text words separation from bi-script, tri-script and five-script scenarios. However, English numeral separation from multilingual documents is ignored but, it is necessary for successful design of multi-script/ multi-lingual OCR. That is what this paper is about.

B Script Identification at Text block and Text line Level

The number of other approaches for automatic script identification at text blocks as well as at text lines has been proposed in the literature and is briefly presented here. Spitz [1] proposed a method for distinguishing between Asian and European languages by examining the upward concavities of connected components. Tan *et al.* [6] proposed a method based on texture analysis for automatic script and language identification from document images using multiple channel (Gabor) filters and Gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Koreans, Malayalam, Persian and Russian. Hochberg, *et al.* [2,3] described a method of automatic script identification from document images using cluster-based templates and also proposed an algorithm for handwritten script identification of six scripts using statistical features extracted based on connected components. Tan [5] developed rotation invariant features extraction method for automatic script identification for six languages. Wood *et al.* [4] described projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters. A. Busch *et al.* [9] proposed a texture based script identification system using wavelet features. Pal *et al.* [11] proposed an automatic technique of separating the text lines from 12 Indian scripts. Gaurav *et al.* [12] proposed a method for identification of Indian languages by combining Gabor filter based techniques and direction

distance histogram classifier for Hindi, English, Malayalam, Bengali, Telugu and Urdu. Basavaraj *et al.* [14] proposed a neural network based system for script identification of Kannada, Hindi and English. Nagabhushan *et al.* [18] discussed an intelligent pin code script identification methodology based on texture analysis using modified invariant moments. In this paper an attempt is made to demonstrate the potentiality of visual discriminating features for script identification at word level. This work is in continuation of [27, 28, and 29] and modification of [30].

In Section 2, the brief overview of data collection, pre-processing is presented. In Section 3, the feature extraction, features Computation and K nearest neighbour classifier are discussed. The experimental details and results obtained are presented in Section 4. Error analysis and comparative study is discussed in Section 5. Conclusion is given in Section 6. The overall method and process proposed is pictorially represented in Fig. 2.

II. DATAT COLLECTION AND PREPROCESSING

A. Data Collection

The data set of 6000 word images are obtained by segmenting 350 document images. Out of 350, one hundred and fifty documents are collected from various books Magazines and Newspapers. Another 100 documents are downloaded from digital library of Indian institute of science. The remaining 100 documents are downloaded from samachar.com, then printed and finally scanned with 300 dpi. Most of these documents are bilingual and trilingual in nature.

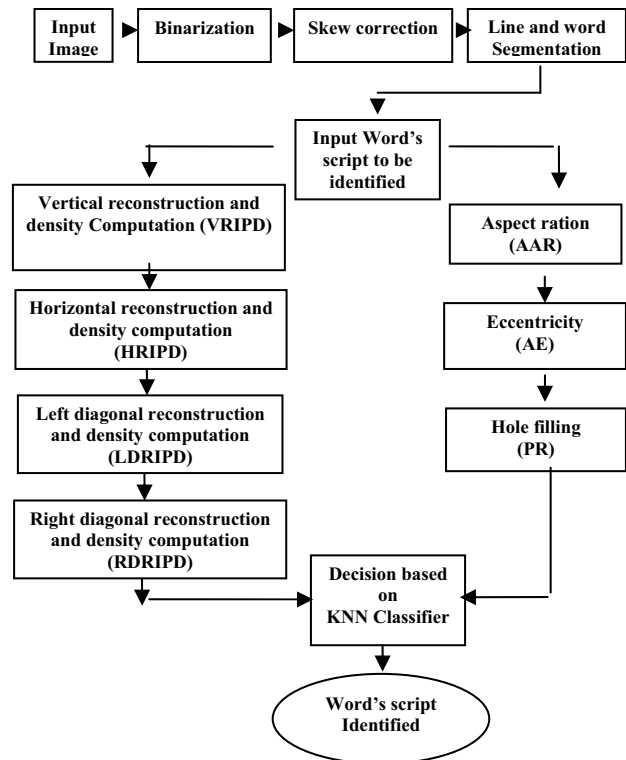


Figure 2. Overall steps involved in the proposed method

These, documents contain lot of variability in terms of font size, styles and scanning resolutions varying from 300 to 600 dpi, as well as the age and nature of the document.

B. Preprocessing

In general, the scanned document images are not good candidates for segmentation and feature extraction. The varying degrees of contrast in gray scale images and the presence of skew and noise will affect such features, leading to high classification error rates. In order to reduce the impact of these factors, the document images from which features are to be extracted must undergo a significant amount of preprocessing. The individual steps, which are performed in this stage, are binarization, deskewing, and segmentation. However, we assume that, the document image contains only text matter.

Binarization can be described as the process of converting a gray scale image into one, which contains only two distinct tones, that is black and white. This is an essential stage in many of the algorithms used in document analysis; especially those that identify connected components, that is, group of pixels, which are connected to form a single entity. For the purpose of this evaluation, a global thresholding approach provides an adequate means of binarization, and the method proposed by Otsu in [20] is used. The morphological area opening operation is performed to remove the noise like, periods, commas and quotation marks of area less than are equal to 20 pixels.

C. Skew Detection and Correction

Knowing the skew of a document is necessary for many document image analysis tasks, for example calculating projection profiles requires knowledge of the skew angle of the image to a high precision in order to obtain accurate results. In practice, the exact skew angle of a document is rarely known, as scanning errors, different page layouts, or even deliberate skewing of text can result in misalignment. In order to correct this, it is necessary to accurately determine the skew angle of a document image or of a specified region of the image, and for this purpose, a number of techniques have been presented in the literature [21, 22, 23 and 24].

Dhandra et al [31] discussed an image dilation and region labeling method for skew angle detection. In order to estimate the skew angle, they dilated the input image horizontally using a line-structuring element with a length of 32 pixels to fill up the gaps between the characters and words. Further, regions are labeled and then orientation of each label has been calculated and their mean is taken as an estimate of skew angle. However, this algorithm works well for English documents and when the script of the document changes, especially for Indian scripts having isolated descenders and ascenders; it fails to estimate the skew angle accurately. Most of Indian scripts have descenders and ascenders; therefore to make this algorithm script independent, we extended it by dilating the input image in vertical and horizontal directions with a line-

structuring element to fill up the gaps between characters, descenders and ascenders to produce the text lines or words as the regions. The length of the structuring element is computed using equation (1) with $k=0.5$ and $k=0.7$ for vertical and horizontal directions respectively. Then, regions are labeled and their average orientation is taken as an estimate of skew angle (θ). Rotating an input image in opposite direction by θ performs the skew correction.

$$Strel=K. \text{Mean (connected components height)} \quad (1)$$

where, strel means the length of the structuring element and K varies from 0.25 to 0.8.

D. Segmentation

To segment the document image into several text lines, we use the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines. Similarly, to segment each text line into several text words, we use the valleys of the vertical projection of each text line obtained by computing the column-wise sum of black pixels. The position between two consecutive vertical projections where the histogram height is least denotes one boundary line. Using these boundary lines, every text line is segmented into several text words.

III. FEATURE EXTRACTION

The feature extraction is the integral part of any recognition system. The aim of feature extraction is to identify patterns by means of minimum number of features that are effective in discriminating pattern classes. The proposed algorithm is inspired by a simple observation that every script/language defines a finite set of text patterns, each having a distinct visual appearance [5], and hence every language could be identified based on its discriminating features.

A. Properties of Scripts

Devnagari: Devnagari is the most popular script in India. It has 12 vowels and 33 consonants. They are called *basic characters*. 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. When vowels are written in this way they are known as *modifiers*. Sometimes two or more consonants can combine and take new shapes these new shape clusters are known as *compound characters*. These types of basic characters, compound characters and modifiers are present not only in Devnagari but in Kannada, Tamil, Malayalam and Odiya.

The Hindi (Devnagari) language characters (alphabets) have a horizontal line at the upper part. This line is called *sirekha*. However, we shall call them *headlines*. When two or more Devnagari characters sit side by side to form a word, the *sirekha* or *headline* touch one another and

generates a big headline [11], which is used as the major feature to distinguish from Kannada, Tamil, Malayalam, Odiya text words and English numerals.

Kannada: The Kannada script has 13 vowels and 37 consonants and they are flat in shape. They have less aspect ratio as compared to English numerals. Kannada script can be discriminated easily from Devnagari using horizontal stroke feature. Kannada characters have dominant left and right diagonal strokes as compared to other five scripts.

Tamil: The Tamil script has 12 vowels and 18 consonants and six special characters. Combination of consonants with vowels gives rise to new symbol or modified symbol. Hence a set of 262 symbols exists in the Tamil script. Tamil script has dominant vertical strokes as compared to Kannada script. Its aspect ratio is less than the English numerals.

Malayalam: Malayalam script has 13 vowels and 37 consonants. Most of the characters in this particular script have a convex curve type shape at their left or right end or both.

Odiya: The alphabet of the modern Oriya script consists of 11 vowels and 41 consonants. Oriya script contains dominant vertical and diagonal strokes.

English numeral: The distinct property of the English numerals is the existence of the vertical strokes like structure. From the experiment, we noticed that the vertical strokes in digits like 1, 3, 4, 6, 8, 9, and 0 are more dominant than that of horizontal strokes. The sample character set (vowels) of proposed scripts is shown in Fig. 3 to exhibit their shapes. The seven features and their method of computation are discussed below.

Aspect Ratio (AR): The word images are used to compute the eight-connected components of white pixels on the image and produce the bounding box for each of the connected components. Aspect ratio can be defined as the ratio of component height to component width [3]. The average aspect ratio (AAR) is defined as

$$AAR = \frac{1}{N} \sum_{i=1}^N \frac{height(component_i)}{width(component_i)} \quad (2)$$

where, N is the number of connected components in an image and the value of the aspect ratio is a real number. It is important feature for word wise script identification [7].

Eccentricity (AE): The eccentricity is a contour based global shape feature [8]. It is defined as the length of the major axis divided by the length of the minor axis [8] of a connected component. The average eccentricity (AE) is defined as

$$AE = \frac{1}{N} \sum_{i=1}^N \frac{len_maj_axis(Component_i)}{len_min_axis(component_i)} \quad (3)$$

where, N is the number of connected components in an image. The value of eccentricity is a real number.

To extract the characters or components containing strokes in vertical, horizontal, right and left diagonal directions, we have performed the erosion operation on

the input binary image (taken as texture) with the line-structuring element. The length of the structuring element is computed using equation (1) with k=0.7. The resulting image is used for opening by reconstruction in the vertical, horizontal, right and left diagonal directions using a fast hybrid reconstruction algorithm [25]. Reconstruction is a morphological transformation involving two images and a structuring element. One image, the marker, is the starting point for the transformation. The other mask image constraints the transformation. In this paper, a fast hybrid reconstruction algorithm [25] is used for reconstruction and erode image is used as the marker image throughout the experiment.



Figure. 3, First row: Devnagari Vowels, second row: Kannada Vowels, third row: Tamil Vowels, fourth row: Malayalam Vowels and fifth row: Odiya Vowels

Horizontal Reconstructed Image Pixels Density (HRIPD): It is defined as the ratio of the total number of on pixels left after horizontal reconstruction with fill holes to total number of pixels in an input image. The value of horizontal pixel density is a real number. It is computed using equation (5).

Vertical Reconstructed Image Pixels Density (VRIPD): It is defined as the ratio of the total number of on pixels left after vertical reconstruction with fill holes to total number of pixels in an input image. The value of vertical pixel density is a real number. It is computed using equation (6). Similarly, the right diagonal reconstructed image pixels density (RDRIPD) and left diagonal reconstructed image pixels density (LDRIPD) are defined.

For fill holes, we choose the marker image (erode image), f_m , to be 0 everywhere except on the image border, where it is set to 1-f. Here f is the original image.

$$f_m(x, y) = \begin{cases} 1-f(x, y), & \text{if } (x, y) \text{ is on the border of } f \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Then $g = [R_f^c(f_m)]^c$ has the effect of filling the holes in f as shown in Fig. 5, where, R_f^c is the reconstructed image.

Pixels Ratio (PR): It is defined as the ratio of the total number of on pixels left after performing hole fill operation on input image to its total number of pixels and is a real number. Equation (7) is used to compute PR. Thus, a vector of seven features i.e. $X = [AAR, AE, HRIPD, VRIPD, RDRIPD, LDRIPD, PR]$ is obtained.

The horizontal reconstruction, vertical reconstruction and pixel ratio computation is pictorially represented in Fig. 4, Fig.5 and Fig. 6. The vertical and horizontal reconstruction process of Kannada, Tamil, Hindi and English numerals is presented in Fig. 7. The sample text word images of Kannada, Devnagari, Tamil, Malayalam and English numerals are shown in Fig. 8.

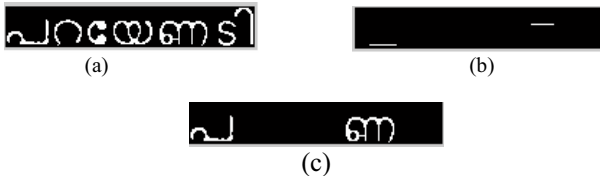


Figure 4. (a)Malayalam word (b) horizontal strokes of (a) and (c) horizontal reconstruction using (b) with fill holes

$$HRIPD = \frac{\text{sumof_onpixels}(c)}{\text{size}(c)} \quad (5)$$

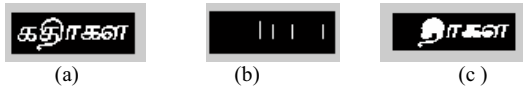


Figure 5. (a)Tamil word (b) vertical strokes of (a) and (c) vertical reconstruction using (b) with fill holes

$$VRIPD = \frac{\text{sumof_onpixels}(c)}{\text{size}(c)} \quad (6)$$



Figure 6. (a) English numeral (b) after hole fill of (a)

$$\text{Pixel ratio (PR)} = \frac{\text{sumof_onpixels}(b)}{\text{size}(b)} \quad (7)$$

B. K-Nearest Neighbour Classifier

In order to identify the most appropriate classifier for the problem of script identification, the KNN classifier is chosen based on its best performance reported in [10] as compared to well known classifiers, namely, Parzen density, quadratic Byes, feed-forward neural net and support vector machine. Our underlying problem is similar to one reported in [10], hence the choice of KNN classifier is justified. The K-nearest neighbour is a supervised learning algorithm. It is based on minimum distance (Euclidian distance calculated using Eq. 8) from the query instance to the training samples to determine the k- nearest neighbours. After determining the k nearest neighbours, we take simple majority of these k-nearest neighbours to be the prediction of the query instance. The experiment is carried out by varying the number of neighbours ($K= 3, 5, 7$) and the performance of the algorithm is optimal when $K = 3$. To asses the performance of the classifier the feature set of 6000 word images are randomly divided (approximately equal) into five groups and a 5-fold cross validation is performed to get optimum result as reported in Tables-2 to 7.

$$D_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - y_{jk})^2} \quad (8)$$

IV. RESULTS AND DISCUSSIONS

Out of 6000 word images, Kannada, Devnagari-Tamil, Malayalam, Odiya are 1000 each and 850 are English numerals. The average script identification results of KNN classifier with five fold cross validation test for bi-script and tri-script problem is presented in Table 2, 3, 4, 5, 6 and 7 and are quite comparable. Though, the aim of this paper is achieved, that is the word-wise script identification in bilingual documents; a second experiment is conducted on 150 word images to test the sensitivity of the algorithm with respect to font sizes and styles. These words are first created in different fonts using DTP packages, and then printed from a laser printer. The printed documents are scanned as mentioned earlier. The most commonly used five fonts of Kannada, Hindi and English are considered for experiment. For each font 10 word images of 10 to 36 points are considered. These 150 word images are distributed equally among Kannada, Devnagari and English numerals. The Kannada font styles used are KN-TTKamanna, TTUma, TTNandini, TTPadmini and TT-Pampa. The Devnagari font styles considered are DV-TTAakash, TTBhima, TTNatraj, TTRadhika, and TTSurekha. Times New Roman, Arial, Times New Roman italic, Arial Black and Bookman Old Style of English numerals are used for font and size sensitivity testing. It is noticed that, script identification accuracy achieved for second data set is consistent. In the reported work of [13, 15, 16], it is mentioned that, the error rate is more when the word size is less than 3 characters.

V. ERROR AND COMPARATIVE ANALYSIS

Proposed algorithm perform better even for single character words, but it fails when words like से, हो, marks like “|” and broken sireorkha’s are encountered in Devnagari. The touched and broken components of Kannada word images are not recognized correctly because of loss in aspect ratio. The boldface word images of Tamil are also misclassified as Devnagari, due to the increase in stroke densities. As word length reduces to less than two characters the variation of vertical strokes densities between Devnagari, Malayalam, Odiya and English numeral are negligible. Hence, an average of 5% misclassification occurred between these scripts.

The proposed algorithm is implemented in MATLAB 6.1. The average time taken to recognize the script of a given word is 0.3024 seconds on a Pentium-IV with 128 MB RAM based machine running at 1.80 GHz. Since, there is no work reported for script identification of numerals at word level, to the best of our knowledge, the results of this work could not be compared. However, we compared the performance of the proposed algorithm with the algorithms proposed for text words separation. The comparative study is presented in Table 1.

TABLE I
COMPARATIVE ANALYSIS

Method proposed by	Scripts considered for experimentation	Accuracy in %	Remark
D.Dhanya	Tamil and Roman	96.03	These algorithms have word size constraint
U.Pal	Devnagari, Telugu and Roman	96.72	
U.Pal	Roman, Devnagari, Bangala, Telugu, Malayalam and Gujarati	97.00	
M.C.Padma	Kannada, Roman and Devnagari	95.66	
Peeta Basa Pati	Devnagari, Tamil and Oriya	97.33	
Proposed Method	Kannada and Roman	98.11	Proposed algorithm has no word size constraint
	Hindi and Roman	97.78	
	Tamil and Roman	99.89	
	Malayalam and Roman	99.45	
	Odiya and Roman	97.28	
	Kannada, Hindi and Roman	95.77	
	Tamil, Hindi and Roman	98.07	
	Malayalam, Hindi and Roman	94.37	
	Odiya, Hindi and Roman	96.04	

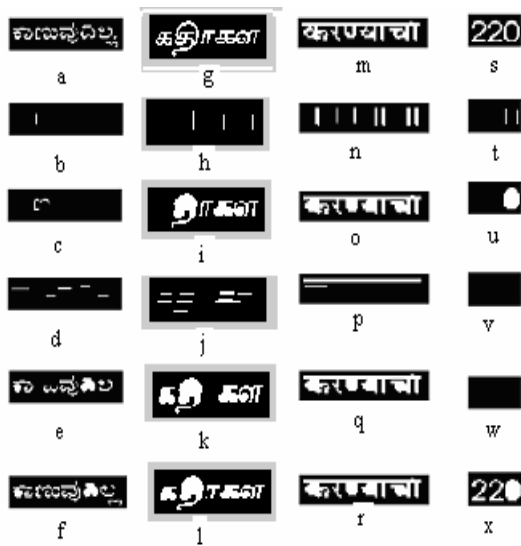


Figure 7. (a), (g) and (m) are input images of Kannada, Tamil and Devnagari text words; (s) input image of English numeral. (b), (h), (n) and (t) are the images of vertical strokes, (c), (i), (o) and (u) are the reconstructed images with fill holes of (b), (h), (n) and (t) that contain vertical strokes, (d), (j), (p) and (v) are the images of horizontal strokes, (e), (k), (q) and (w) are the reconstructed images with fill holes of (d), (j), (p) and (v) that contain horizontal strokes, (f), (l), (r) and (x) are original text word images with fill holes of Kannada, Tamil, Devnagari and English numeral

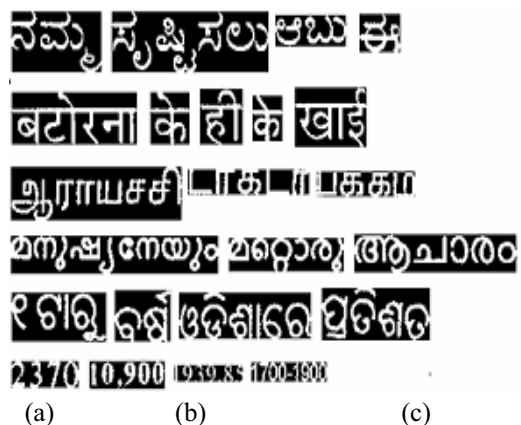


Figure 8. First, second, third, fourth, fifth and sixth rows are sample word images of Kannada, Devnagari, Tamil, Malayalam, Odiya and English numerals respectively.

The performance of the proposed features are experimentally observed for bi-class problem (Kannada and English numeral) and found that the aspect ratio is the dominant feature for recognition whereas horizontal and right diagonal features are weak (see Fig. 9). The maximum average recognition accuracy of the proposed algorithm for bi-script (Tamil and English script) problem is 99.89% and for tri-script (Tamil, Hindi and English) problem is 98.07%, whereas the maximum average accuracy of bi-script problem reported by U.Pal et.al.[26] as 97% and the maximum average recognition accuracy of the tri-script problem reported by Pati et.al.[19] as 99% with 32 features. With this it is clear that the performance of the proposed algorithm in terms of accuracy, time complexity and dimensionality is better than the other algorithms proposed in the literature.

VI. CONCLUSION

In this paper, we have investigated a tool of structural stroke primitives present in different directions and global shape features for script identification at word level. The morphological opening by reconstruction approach is efficiently used for stroke based reconstruction and its contribution for discriminating the proposed scripts is evident from this algorithm. The simplicity of the algorithm is that, it exploits efficiently the distinct visual factors (strokes, holes etc.), which are guiding to our human script identification. Furthermore, our method overcomes the word length constraint of [13, 15, 16] and works well even for single component words. During the extraction of features, the connected components of size less than are equal to 20 pixels are removed from the image prior to features computation. Thus, the approach is robust with respect to noise. It is also clear that the algorithm is insensitive to font styles, sizes, word length and scanning artifacts like resolution. Experimental results have exhibited that relatively simple technique can reach a high level accuracy for discriminating the proposed scripts. It is our future endeavor to modify this algorithm to perform script identification from multilingual document images containing more number of Indian languages.

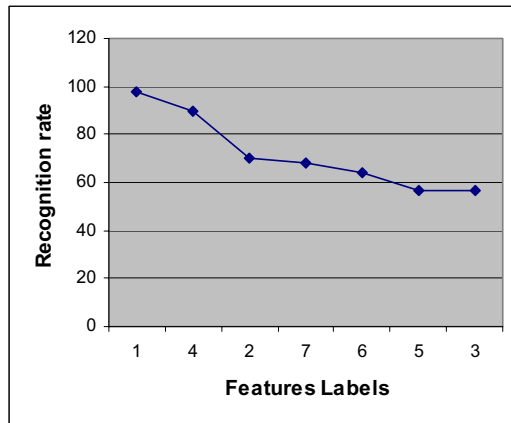


Figure 9. Features performance chart

TABLE 2
RECOGNITION RESULTS OF KANNADA WORDS AND ENGLISH NUMERALS AS

Script	K NN	
	Kannada	English
Kannada	96.89%	3.11%
English	0.67%	99.33%

TABLE 3
RECOGNITION RESULTS OF DEVNAGARI WORDS AND ENGLISH NUMERALS AS

Script	K NN	
	Hindi	English
Hindi	96.78%	3.22%
English	1.22%	98.78%

TABLE 4
RECOGNITION RESULTS OF TAMIL WORDS AND ENGLISH NUMERALS AS

Script	K NN	
	Tamil	English
Tamil	99.78%	0.22%
English	0%	100%

TABLE 5
RECOGNITION RESULTS OF MALAYALAM WORDS AND ENGLISH NUMERALS AS

Script	K NN	
	Malayalam	English
Malayalam	99.22%	0.78%
English	0.33%	99.67%

TABLE 6
RECOGNITION RESULTS OF ODIYA WORDS AND ENGLISH NUMERALS AS

Script	K NN	
	Odiya	English
Odiya	96.67%	3.33%
English	2.11%	97.89%

TABLE 7
SCRIPT IDENTIFICATION AVERAGE RESULTS OF TRI-SCRIPT, IN PERCENTAGE, USING KNN WITH K=3

Scripts	Kannada	Tamil	Malayalam	Odiya
English Numerals	98.44	99.00	98.56	96.78
Hindi	95.67	96.67	92.44	96.22
Local	93.22	98.56	92.11	95.11
Average	95.77	98.07	94.37	96.03

The average accuracy of English numerals separation from Kannada, Hindi, Tamil, Malayalam and Odiya

documents are 98.11%, 97.78%, 99.89%, 99.44% and 97.28% (see Table 2, 3, 4, 5 and 6) respectively. Furthermore, the average accuracy of English numerals separation from the documents containing Kannada/Hindi, Tamil/Hindi, Malayalam/Hindi and Odiya/Hindi are 95.77%, 98.07%, 94.37% and 96.03% (see Table 7) respectively. In Table 7, local means the local scripts Kannada, Tamil, Malayalam and Odiya.

ACKNOWLEDGMENT

Authors are grateful to the referees for their critical comments and suggestions. Further, we extend our gratitude to Dr. P. Nagabushan for his valuable suggestions, and discussions. We are also thankful to Dr. G. Hemantha Kumar and Dr. D. S. Guru, Dept. of Computer Science, University of Mysore, for their helpful discussion and encouragement during this work.

REFERENCES

- [1] A.L.Spitz, "Determination of the script and language content of document images," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp.234-245, 1997.
- [2] J. Hochberg, P. Kelly, T Thomas and L Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.19, pp.176-181, 1997
- [3] Judith Hochberg, Kevin Bowers, Michael Cannon and Patrick Keely, "Script and language identification for hand-written document images," *IJDAR*, vol.2, pp45-52. 1999.
- [4] S. Wood. X. Yao. K.Krishnamurthi and L.Dang "Language identification from for printed text independent of segmentation," *Proc. of Int'l. Conf. on Image Processing*, pp. 428-431, 1995.
- [5] T.N.Tan, "Rotation invariant texture features and their use in automatic script identification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp.751-756, 1998.
- [6] G.S.Peake and Tan, "Script and language identification from document images," *Proc. of Eighth British Mach. Vision Conf.*, vol.2, pp. 230-233, Sept-1997
- [7] Anoop M. Namboodri, Anil K Jain, " Online handwritten script identification", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 26,no.1,pp. 124-130, 2004
- [8] Dengsheng Zhang, Guojun Lu, " Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, pp. 1-19, 2004
- [9] A.Busch ,W.W.Boles and S.Sridharan, " Texture for script identification" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(11) 1720-173,2005
- [10] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy," Script Identification for Indian Documents", *In. Pro. of 7th IAPR workshop on Document Image Systems, (DAS)*, New Zealand,pp.255-267, 2006
- [11] U.Pal and B.B.choudhuri "Script Line separation from Indian Multi- script Documents", *In Proc. of 5th ICDAR*, pp.406-409, 1999
- [12] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, R.B.Shet," Identification of scripts of Indian languages by Combining trainable classifiers," *Proc. of ICVGIP 2000*, Dec-20-22, Bangalore, India
- [13] D Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," *Sadhana*, vol. 27, part-1, pp. 73-82, 2002

- [14] S.Basavaraj, Patil and N.V.Subbareddy. "Neural network based system for script identification in Indian documents," *Sadhana*, vol. 27, part-1, pp. 83-97, 2002.
- [15] U.Pal. S.Sinha and B.B Chaudhuri, "Word-wise Script identification from a document containing English, Devnagari and Telgu Text," *Proc. of NCDAR*, pp 213-220,2003
- [16] M.C.Padma and P. Nagabhushan," Identification and separation of text words of Kannada Hindi and English languages through discriminating features," *Proc. of NCDAR*,pp. 252-260. 2003.
- [17] Peeta Basa pati, S. Sabari Raju, Nishikanta Pati and A.G. Ramakrishnan, "Gabor filters for document analysis in Indian Bilingual Documents," *Proc. of ICISIP*, pp. 123-126,2004
- [18] P. Nagabhushan, S.A. Angadi and B.S. Anami," An Intelligent Pin code Script Identification Methodology Based on Texture Analysis using Modified Invariant Moments," *Proc. of ICCR*,, pp. 615-623,2005
- [19] Peeta Basa Pati and A.G.Ramakrishnan," HVS inspired system for Script Identification in Indian Multi-Script Documents", *In Proc. of 7th International Workshop on Document Analysis System, Nelson Newland*,pp-380-389, Feb-13-15,2006
- [20] N. Otsu, " A Threshold Selection Method from Gray-Level Histogram , *IEEE Trans. Systems,Man, and Cybernetics*, vol.9,no.1,pp.62-66,1979
- [21] H.S. Baird, "The Skew Angle of Printed Documents", *In proc. of Document Image Analysis, L. O' Gorman and R. Kasturi, eds., IEEE CS Press*, pp. 204-208,1995
- [22] B.B.Chaudhuri and U.Pal, "Skew Angle Detection of Digitized Indian Script Documents," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no.2, pp. 703-712, Feb.1997
- [23] W.Postl, " Detection of Linear Oblique Structures and Skew Scan in Digitized Documents", *In Proc. of Int'l Conf. Pattern Recognition*, pp. 687-689,1986
- [24] G.Peak and T Tan, "A General Algorithm for Document skew Angle Estimation," *Proc. Int'l. Conf. Image Processing*, vol.2, pp. 230-233, 1997
- [25] Vincent, L., " Morphological gray scale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. on Image processing*, vol.2, no. 2, pp. 176-201, 1993
- [26] Suranjit Sinha, Umapada Pal and B.B. Chaudhuri"Word-Wise Script Identification from Indian Documents", *Proc. of 6th International Workshop, DAS, Florence*, pp. 310-321,2004
- [27] B.V.Dhandra,P.Nagabhushan, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath, "Script Identification based on Morphological Reconstruction in Document images", *in proc.18th International conference on Pattern Recognition- ICPR, Hong Kong, V. No. II-3*, pp 950-953,2006
- [28] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath, "Word-wise Script Identification based on Morphological Reconstruction in Printed Bilingual Documents," *Proc. of IET International Conference on Vision Information Engineering VIE, Bangalore* pp 389-393, 2006
- [29] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath,"Word-wise Script Identification from Bilingual Documents based on Morphological Reconstruction," *Proc. of First IEEE International Conference on Digital Information Management*, pp. 389-394, 2006
- [30] B.V.Dhandra, H.Mallikarjun, Ravindra Hegadi, V.S.Malemath, "Word Level Script Identification through Discriminating Features from Document Images" *In Proc. of International Conference on Signal Processing, Communications and Networking*, Feb. 22-24 at MIT, Chennai. pp. 630-635,2007
- [31] B.V.Dhandra, V.S.Malemath, Mallikarjun Hangarge, Ravindra Hegadi, "Skew detection in Binary image documents based on Image Dilation and Region labeling Approach", *In Proceedings of ICPR 2006, V. No. II-3*, pp. 954-957

B.V.Dhandra. Professor and Chairman, P.G.Department of Studies and Research in Computer Science, Gulbarga University, Gulbarga, Karnataka, INDIA. He has born on January 1st 1955 in Gulbarga, India. He has received his MA degree in Statistics in 1979 and M.Phil in Statistics in 1986 from Karnataka University Dharwad, Karnataka, India. He obtained his Ph.D. degree from Shivaji University, Kolhapur, India in the year 1993. He has served as lecturer during 1979 to 1993, as a Reader during 1993 to 2001 and since 2001 he has been serving as a professor in Computer Science and presently heading the PG Department of Studies and Research in Computer Science, Gulbarga University, Gulbarga, India He is member of Board of Studies of various Universities. He has published more than 30 research articles in peer reviewed national and international conferences and journals. His research interests are Pattern Recognition, Image Processing and Operations Research.

Mallikarjun Hangarge is a senior faculty member and Head of the Department of Computer Science, Karnataka Arts, Science and Commerce Degree College, Bidar, Karnataka, India. He has born on January 1st 1966 in Gulbarga, India. He has received his MSc degree in Statistics from Gulbarga University, in 1989, India. He has received Post Graduation Diploma in Computer Applications in 1992 from Gulbarga University, India. He has been awarded MPhil in Computer Science in 2003 from M.S.University, Trinnelvelli, Tamil Nadu, India. He has 15 years of experience in teaching for under graduate and post graduate students. Presently, he is carrying out his PhD under the faculty improvement program of University Grant Commission, New Delhi, India. He has published more than 15 research articles in peer reviewed national and international conferences and journals and as part of his PhD work he has published 8 research articles. His research interests are Pattern Recognition and Image Processing.