Script Identification from Bilingual Gujarati-English Documents

Shailesh A. Chaudhari
M.Sc. (I.T.) Programme
Veer Narmad South Gujarat University
U.M. Road, Surat

Ravi M. Gulati, PhD
Dept. of Computer Science
Veer Narmad South Gujarat University
U.M. Road, Surat

ABSTRACT

In a multi-lingual country like India, in most of the official papers, school text books, magazines, it is observed that English words intersperse within the Indian regional languages. So a bilingual Optical Character Recognition (OCR) system is needed which can recognize these bilingual documents and store it for future use. In this paper authors present an OCR system developed for the script identification of Indian language i.e. Gujarati and Roman scripts for printed documents. Here authors propose the line-wise script identification. The spatial spread of pixels on Upper and Lower parts associated with Gujarati and English are used to identify the script. Authors have used horizontal projection for line distinction belonging to different script. Further, K-nearest neighbour algorithm is used to classify 2000 text lines into two scripts: English and Gujarati, based on 4 spatial spread features extracted using connected component and horizontal projection. The proposed algorithm achieves average classification accuracy as high as 99.70% for bi-script separation.

Keywords

Pre-processing, Segmentation, Vector, kNN Classifier, etc.

1. INTRODUCTION

Researchers have been working for pattern recognition since decades. Optical Character Recognition (OCR) is the oldest sub field of the pattern recognition field and has almost achieved a lot of success in the case of Monolingual Scripts. There are 24 official (Indian constitution accepted) languages in India. Two or more of these languages may be written using one script. Twelve different scripts are used for writing these languages. All official documents, magazines, text books and reports can be converted to electronic form using a high performance Optical Character Recognizer (OCR). In a multi-lingual country like India, documents are often bilingual or multi-lingual in nature. English is interspersed in most of the important official documents, reports, magazines and technical papers in addition to an Indian regional language. Monolingual OCRs fail to recognize such documents and there is a need to extend the monolingual systems to bilingual ones. This paper describes one such system, which handles both Gujarati and Roman script. Optical Character Recognition (OCR) system of such a document page can be made through the Development of a script separation technique to identify different scripts present in the document pages and then run individual OCR developed for each script alphabets.

This paper is organized in the following sections; Section 2 describes the early attempts made in Indian language OCR. Section 3 enlightens properties of Gujarati language. Section 4 explains preprocessing steps. Section 5 is devoted to feature extraction techniques and section 6 explains script identification. Lastly in section 7 conclusions is explained.

2. RELATED WORK

Development of a generalized OCR system for Indian languages is more difficult than a single script OCR development. This is because of the large number of characters in each Indian script alphabet. There are many pieces of work on script identification from a single document. Spitz [1] developed a method to separate Han based and Latin based script separation. He used optical density distribution of characters and frequently occurring word shape characteristics for the purpose. Ding et al. [2] proposed a method for separating two classes of scripts: viz. European (comprising Roman and Cyrillic scripts) and Oriental (comprising Chinese, Japanese and Korean scripts). Dhanya et al. [3] proposed a Gabor filter based technique for word-wise segmentation from bilingual documents containing English and Tamil scripts. Using cluster based templates; an automatic script identification technique has been described by Hochberg et al. [4]. Recently, using fractal-based texture features, Tan [5] described an automatic method for identification of Chinese, English, Greek, Russian, Malayalam and Persian texts. Wood et al. [6] described an approach using filtered pixel projection profiles for script separation. Pal and Chaudhuri [7] proposed a line-wise script identification scheme from tri-language (triplet) documents. Later, Pal et al. [8] proposed a generalized scheme for line-wise script identification from a single document containing all the twelve Indian scripts. Pal et al. [9] also proposed some work on word-wise identification from Indian script documents.

A zone-based feature extraction algorithm proposed by Rajeshekararadhya and Ranjan [10] for the recognition of off-line handwritten digits of four popular Indian scripts. The nearest neighbour feed forward back propagation neural network and support vector machine classifiers are used for subsequent classification and recognition purposes. They obtained a recognition rate of 98.65 % for Kannada digits, 96.1 % for Tamil digits, 98.6 % for Telugu digits and 96.5 % for Malayalam digits using the support vector machine.

An algorithm for language identification of Kannada, Hindi and English text lines from printed documents is proposed by Padma, Vijaya and Nagabhushan [11]. The approach is based on the analysis of the top and bottom profiles of individual text lines and hence does not require any character or word segmentation. Experimental results demonstrate that relatively simple technique can reach a high accuracy level for identifying the text lines of Kannada, Hindi and English languages. The performance has turned out to be 95.4% rate.

Bindu Philip and R. D. Sudhaker Samuel [12] addressed the problem of bilingual character recognition using Gabor features for script identification and Dominant Singular Values as features for classification. The proposed algorithm has been tested successfully and an overall recognition rate of 96.5% is achieved. In our earlier work we have separated printed Roman and Gujarati digits using template matching algorithm and kNN classifier[13][14].

All the above pieces of work are done for script separation from printed documents. In the proposed scheme, at first, the documents are pre-processed by binarization and noise removing. Using horizontal projection profile the document is segmented into lines. The first uniqueness property in between the Roman and Gujarati script is that each line consists of more number of Roman characters as compared to that of Gujarati. Basing on these features we have taken a derived feature value by dividing the number of characters in a line and number of hole (loop) in a line with the line width. And after obtaining a unique value we sent these lines to their respective classifiers. The Figure 1 shows the entire process carried out for the recognition of our bilingual document.

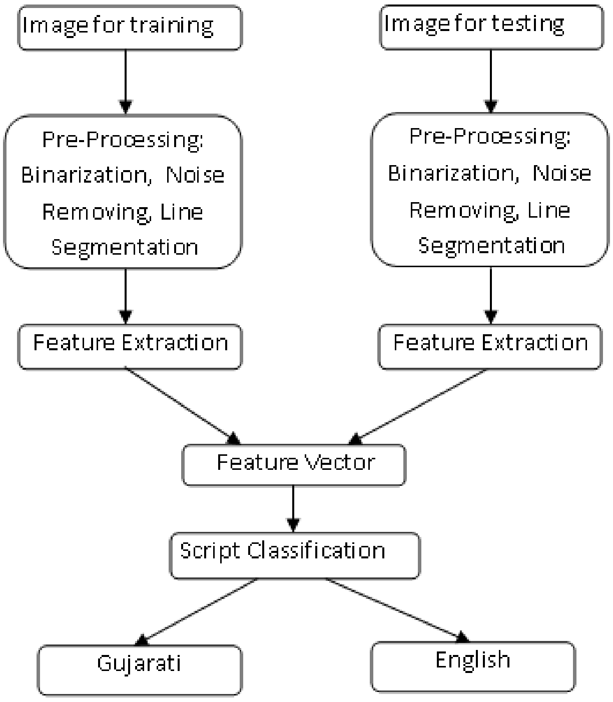


Image for training
Image for testing
Pre-Processing: Binarization, Noise Removing, Line Segmentation
Pre-Processing: Binarization, Noise Removing, Line Segmentation
Feature Extraction
Feature Extraction
Feature Vector
Script Classification
Gujarati
English

Figure 1: Block diagram of proposed system

3. PROPERTIES OF GUJARATI SCRIPT

Gujarati language stands at the 26th position among the most spoken native language in the world and nearly 50 million people throughout the world speak Gujarati. The basic direction of writing Gujarati is from left to right and top to bottom, the same as English. Altogether Gujarati alphabets utilize 94 symbols, which can be categorized into different groupings. Gujarati character set provides 34 (+2 compound *ksha*, *gna*) consonants, 14 vowels which are represented by a single symbol, and 10 numerals. The components of the characters can be classified into:

(a) **Main component**: Either a vowel or a symbol may be consonant.

(b) **Vowel Modifier**: A character can also have a vowel modifier, which modifies the consonants. When the vowel modifier does not touch the main component, it forms separate component, which lies to left or right or top of the main component.

(c) **Consonant modifier**: A symbol can be composed of two or more consonants, the main component and consonant modifier(s) or half consonant. Spatially, the consonant modifier could be at the bottom or the top of the main component, and hence lie above or below the line. More than two up to four consonant vowel combinations are found. These are called conjuncts. The basic characters of Gujarati script are shown in Fig. 2 (a, b, c, d).

ક	ખ	ગ	ઘ	ઙ
ચ	છ	જ	ઝ	ઞ
ટ	ઠ	ડ	ઢ	ણ
ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ
ય	ર	લ	વ	
શ	ષ	સ		
હ	ળ	ક્ષ	જ્ઞ	

Figure 2.a: Gujarati Alphabets

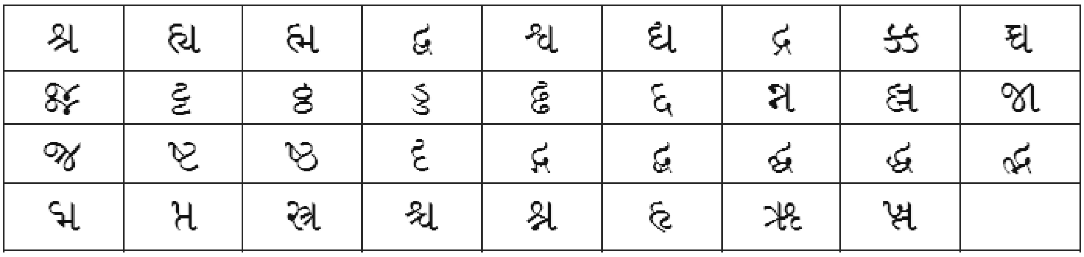
અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ	એ	ઐ	ઓ	ઔ	અં	અઃ
	◌ા	િ◌	◌ી	◌ુ	◌ૂ	◌ૃ	◌ે	◌ૈ	◌ો	◌ૌ	◌ં	◌ઃ

ક કા કિ કી કુ કૂ કૃ કે કૈ કો કૌ કં કઃ

Figure 2b: Gujarati Vowels and its modifiers with character

૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯ ૦

Figure 2.c: Gujarati Digits



શ્ર	હ્ય	હ્મ	ઙ્ગ	શ્વ	દ્ધ	ઙ્ક	ક્ક	દ્વ
ઙ્ક	ટ્ટ	ઠ્ઠ	ડ્ડ	ઢ્ઢ	દ્ગ	ત્ર	હ્ન	જ્ઞ
જ્જ	ટ્ઠ	ડ્ઢ	દ્દ	ઙ્ગ	ડ્ગ	ઙ્ઘ	ઙ્ઘ	ઙ્મ
ષ્મ	મ	સ્ર	શ્ચ	શ્ર	હૃ	ૠ	ક્ષ	

Figure 2.d: Some Gujarati conjunct characters

From the above Figure it can be noted that basic characters have a sharp curve shape. The writing style in the script is from left to right. The concept of upper/lower case is absent in Gujarati script. A consonant or vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character or conjuncts. Compound characters can be combinations of consonant and consonant, as well as consonant and vowel.

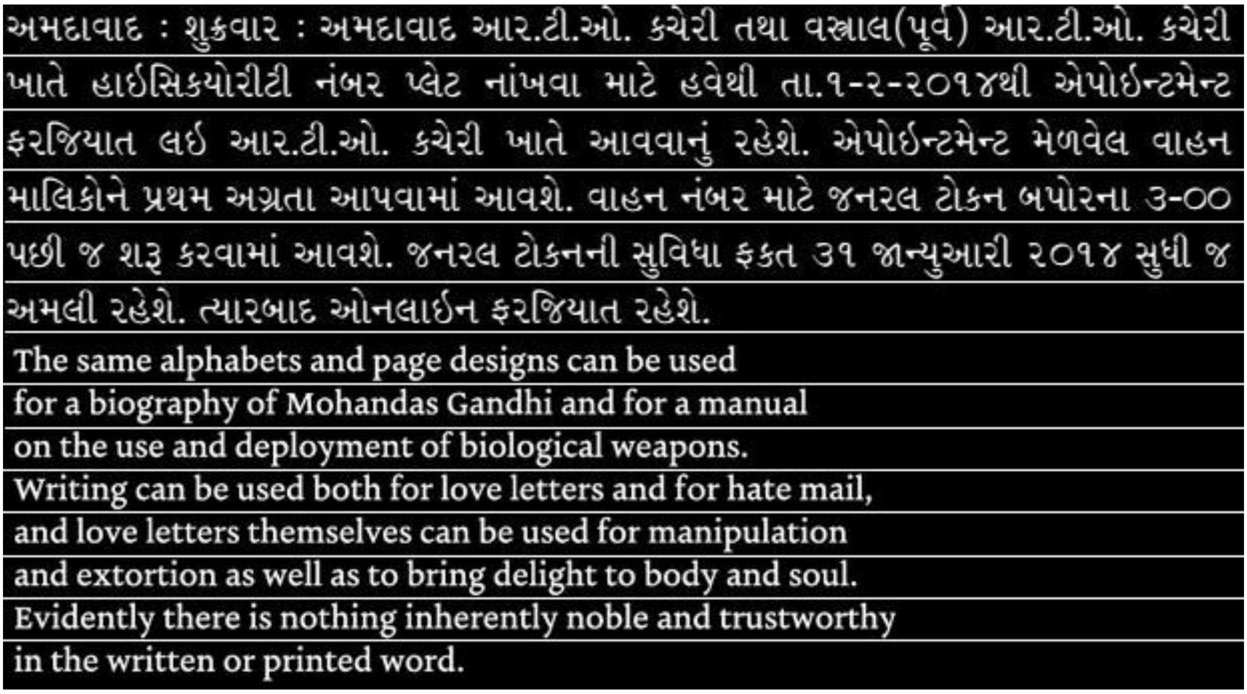
4. PREPROCESSING

4.1 Binarization and Noise removing

The input of an OCR is given in from the scanner. After this we need to binarize the image. In binarization grey-scale image file or colour image file is converted into Binary file (Black & White) using UTSU's global thresholding technique [15]. During noise removal objects containing less than 20 pixels considered as noise and is removed.

4.2 Segmentation

In the present work horizontal projection profile is used for line segmentation. A text line is located between scan lines whose horizontal projection profile values are greater than zero. The major challenge in our work is the separation of lines for script identification. One major factor which we have considered for line identification of different script is the horizontal projection profile. Horizontal projection profile is the sum of black pixels along every row of the image. The purpose of analyzing the text line detection of an image is to identify the physical region in the image and their characteristics. A maximal region in an image is the maximal homogenous area of the image. The property of homogeneity in the case of text image refers to the type of region, such as text block, graphic, text line, word, etc. so we define the segmentation as follows A segmentation of a text line image is a set of mutually exclusive and collectively exhaustive sub regions of the text line image.



અમદાવાદ : શુક્રવાર : અમદાવાદ આર.ટી.ઓ. કચેરી તથા વસ્ત્રાલ(પૂર્વ) આર.ટી.ઓ. કચેરી
ખાતે હાઈસિક્યોરીટી નંબર પ્લેટ નાંખવા માટે હવેથી તા.૧-૨-૨૦૧૪થી એપોઈન્ટમેન્ટ
ફરજિયાત લઈ આર.ટી.ઓ. કચેરી ખાતે આવવાનું રહેશે. એપોઈન્ટમેન્ટ મેળવેલ વાહન
માલિકોને પ્રથમ અગ્રતા આપવામાં આવશે. વાહન નંબર માટે જનરલ ટોકન બપોરના ૩-૦૦
પછી જ શરૂ કરવામાં આવશે. જનરલ ટોકનની સુવિધા ફકત ૩૧ જાન્યુઆરી ૨૦૧૪ સુધી જ
અમલી રહેશે. ત્યારબાદ ઓનલાઈન ફરજિયાત રહેશે.
The same alphabets and page designs can be used
for a biography of Mohandas Gandhi and for a manual
on the use and deployment of biological weapons.
Writing can be used both for love letters and for hate mail,
and love letters themselves can be used for manipulation
and extortion as well as to bring delight to body and soul.
Evidently there is nothing inherently noble and trustworthy
in the written or printed word.

Figure 3: Line Segmentation

Typical top-down approaches proceed by dividing a text image into smaller regions using the horizontal and vertical projection profiles. The proposed line segmentation algorithm, starts dividing a text image into sections based on their projection profiles as shown in figure 3. The algorithm repeatedly partitions the image by alternately projecting the regions of the current segmentation on the horizontal and vertical axes. An image is recursively split horizontally until a final criterion where a split is impossible is met. Projection profile based techniques are extremely sensitive to the skew of the image. Hence extreme care has to be taken while scanning of images.

5. FEATURE EXTRACTION

For script identification, features are identified from segmented line image based on the following observations. From the above projection profile authors observe that:

1. Total numbers of Gujarati characters present in a line are comparatively less than that of the Roman characters.
2. Total number of loops present in a Gujarati script line is comparatively less than that of the Roman script line.
3. The Roman script has very few characters with curve shape as compared to Gujarati script characters. So the horizontal pixel distribution in some portion of Roman script line is more as compared to Gujarati script line (top profile and bottom profile).
4. Due to curvature nature of Gujarati character they have less horizontal pixel distribution in Gujarati text line compared to English text line.

In consideration to the above features for distinction authors have tried to separate the scripts on the basis of the line width. Figure 3 shows different lines extracted for the individual scripts. Here authors have considered the upper and lower *matras* for the Gujarati characters. For script line identification following features is considered.

Connected Component Density: The first feature is the ratio between number of connected components, to line width. To identify connected component authors have used 8-nearest neighbourhood.

$$CCD = \frac{\sum_{i=0}^{N} Connected\_Components}{Line\_Width}$$

***Hole Density*:** From careful analysis it is observed that English line having more holes (loops) as compared to Gujarati line. The ratio between the numbers of number of holes, to line width is stored as second feature.

$$HD = \frac{\sum_{i=0}^{N} Holes}{Line\_Width}$$

Top Projection Max Row Density: The third feature is based on top projection profile. It is calculated by drawing vertical lines from each black pixel on top of the word. If the line reaches the middle of word without encountering any black pixel then the count is incremented by one. Then horizontal projection count is calculated for each row of top projection profile and a max row count is evaluated by dividing line image width and stored as feature. The horizontal projection and top max row count is shown in figure 4.

$$TPMD = \frac{Max\left(Row_{Array}\left(\sum_{i=0}^{N} Object\_Pixels\right)\right)}{Line\_Width}$$

Bottom Projection Max Row Density: The fourth feature is based on bottom projection profile. It is calculated by drawing vertical lines from each black pixel on bottom of the word. If the line reaches the middle of word without encountering any black pixel then the count is incremented by one. Then horizontal projection count is calculated for each row of bottom projection profile and a max row count is evaluated by dividing line image width and stored as feature. The horizontal projection and bottom max row count is shown in figure 4.

$$BPMRD = \frac{Max\left(Row_{Array}\left(\sum_{i=0}^{N} Object\_Pixels\right)\right)}{Line\_Width}$$

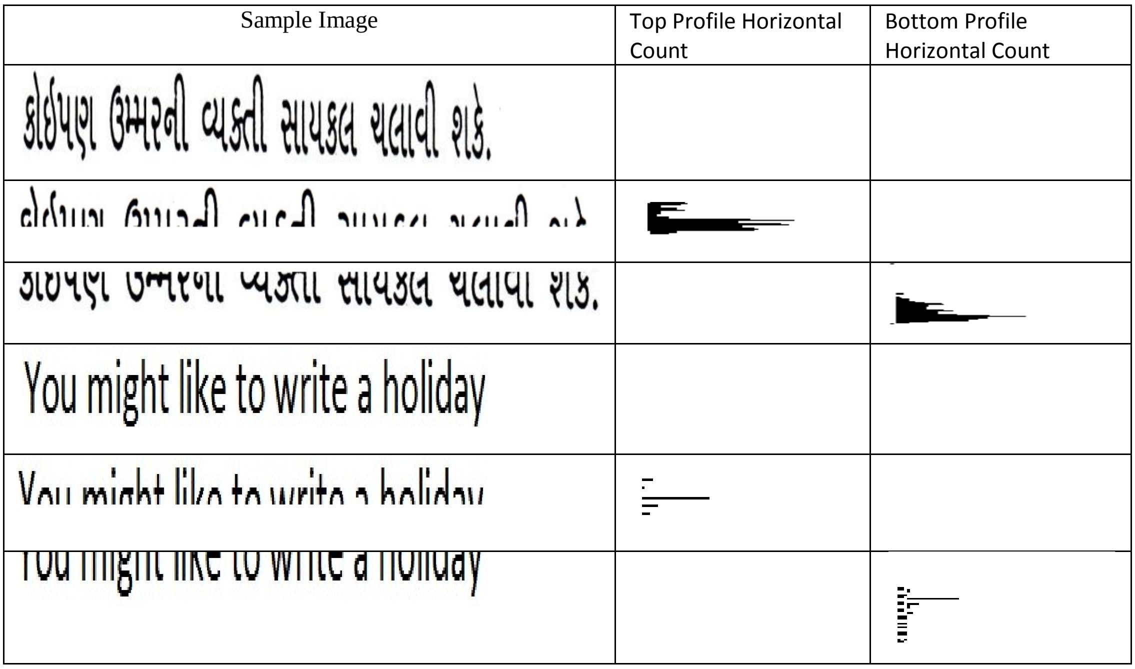

Sample Image
Top Profile Horizontal Count
Bottom Profile Horizontal Count
કોઈપણ ઉમ્મરની વ્યક્તી સાયકલ ચલાવી શકે.
You might like to write a holiday

Figure 4: Top and Bottom Profile Gujarati and English line

6. SCRIPT CLASSIFICATION

A supervised learning algorithm K-nearest neighbour is used for script classification. The four features are extracted from the test image *X* and these feature values are compared with feature values stored in the knowledge base. The Euclidean distance is used to measure the distance between the test sample and the *k* neighbors. After determining the k nearest neighbours, we take simple majority of these k-nearest neighbours to be the prediction of the query image line.

For experimentation, 200 printed document pages obtained from various resources like textbook, news paper, magazine, official documents, etc. are used with an assumption that the document pages contain only text lines. These document pages are scanned using a flatbed scanner at a resolution of 300 dpi. The accuracy of the classification achieved for script identification at text lines is presented in Tables 1.

Table 1. Script Identification Accuracy

Script / Language	kNN Classifier	
	English	Gujarati
English	99.57%	0.43%
Gujarati	0.17%	99.83%

The system is trained to thoroughly understand the nature of the top and bottom profiles using a training data set of 500 text lines from both the scripts. The proposed system is tested thoroughly using a manually created data set of 2000 text lines obtained from different document images. Totally, 2000 text lines from both the scripts are considered for testing. Each test image contained approximately 15 to 25 text lines printed in different font type and font sizes. However, the font size and font type is assumed to be same within the text line. Test document images containing text lines in mixture of Gujarati and English languages were considered. Few test images containing text lines in only one script and mixture of two scripts were also considered. The proposed system is very sensitive with skewed lines; therefore authors have used only skew less line for experimentation.

7. CONCLUSION

In this paper, a new method to identify the script type of the bilingual document containing Gujarati and English text lines is presented. The proposed model is developed based on the distinct features extracted from the top and bottom profiles of the individual text lines. The method looks simple, as it does not require any character or word segmentation. Experimental results demonstrate that relatively simple technique can reach recognition rate of 99.70% for data set constructed from scanned document images. Future work is to develop script identification model at word level for the text lines with the words printed in different scripts. The proposed algorithm suggested in this paper could be modified to apply on the bilingual documents of other Indian states.

8. REFERENCES

[1] L. Spitz. "Determination of the Script and Language Content of Document Images". IEEE Trans. on PAMI, 235-245, 1997

[2] J. Ding, L. Lam, and C. Y. Suen. "Classification of Oriental and European Scripts by using Characteristic Features". In Proceedings of 4th ICDAR, pp. 1023-1027, 1997

[3] D. Dhanya, A. G. Ramakrishna, and P. B. Pati. " Script Identification in Printed Bilingual Documents". Sadhana, 27(1): 73-82, 2002

[4] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns. "Automatic script Identification from Document Images using Cluster-Based Templates" IEEE Trans. on PAMI, 176-181, 1997

[5] T. N. Tan. "Rotation Invariant Texture Features and their use in Automatic Script Identification". IEEE Trans. On PAMI, 751-756, 1998

[6] S. Wood, X. Yao, and K. Krishnamurthi, , L. Dang. "Language Identification for Printed Text Independent of Segmentation". In Proc. Int'l Conf. on Image Processing. 428-431, 1995

[7] U. Pal, and B. B Chaudhuri,. "Script Line Separation from Indian Multi-Script Documents". IETE Journal of Research, 49, 3-11, 2003

[8] U. Pal, S. Sinha, and B. B. Chaudhuri. "Multi-Script Line identification from Indian Documents". In Proceedings 7th ICDAR, 880--884, 2003

[9] S. Chanda, U. Pal, "English, Devnagari and Urdu Text Identification". Proc. International Conference on Cognition and Recognition, 538-545, 2005

[10] S. V. Rajashekararadhya, Dr P. Vanaja Ranjan, "Handwritten Numeral/Mixed Numerals Recognition Of South-Indian Scripts: The Zonebased Feature Extraction Method" Journal of Theoretical and Applied Information Technology, 2009, Vol 7. No 1.

[11] M.C. Padma, P. A. Vijaya, P. Nagabhushan, "Language Identification from an Indian Multilingual Document Using Profile Features", International Conference on Computer and Automation Engineering, IEEE, 2009, 978-0-7695-3569-2.

[12] Bindu Philip and R. D. Sudhaker Samuel, "A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values", International Conference on Digital Image Processing, IEEE, 2009, 978-0-7695-3565-4/09.

[13] S. Chaudhari, R. Gulati, "Character Level Separation and Identification of English and Gujarati Digits from Bilingual (English-Gujarati) Printed Documents", International Journal of computer applications(IJCA), NewYork, USA, 2012.

[14] S. Chaudhari, R. Gulati, "An OCR for Separation and Identification of Mixed English - Gujarati Digits using kNN Classifier", Proc. International Conference on Intelligent Systems and Signal Processig, 2013.

[15] N. Otsu, " A threshold selection method from gray level histogram ", IEEE Trans. Syst. Man Cyb, Vol.9, no.1, pp.62-66, 1979.