

Script Identification from Handwritten Document Images Using LBP features

¹G. G. Rajput, ²Suryakanth Baburao Ummapure

¹Department of Computer Science, Akkamahadevi Women's University, Vijapura 586106, Karnataka, India

²Department of Computer Science Gulbarga University, Kalaburagi-585106, Karnataka, India

Corresponding Author: ²Suryakanth Baburao Ummapure

ABSTRACT:In the context of our country, script recognition is a complex task because of more number of prevailing scripts and that many of the official documents are single script and multi-script in nature, English being commonly used along with regional script. Hence, it is pre-requisite to identify the script in the document and then feed the document to respective OCR for further processing. Texture of a script is a unique feature that can be used to identify the script type. Recently, local binary pattern (LBP) operator is extensively used in extracting statistical based texture features because of its simplicity in implementation and robustness to changes in neighborhood intensity values. In this paper, we explore the significance of LBP operator for script identification. Performing connected component analysis on handwritten document images, lines are extracted and features are computed using LBP operator. Identification of script type is done using K-NN and SVM classifiers. Experiments are performed on handwritten document images written in Kannada, Hindi, English, Malayalam, Punjabi, Tamil, Telugu, Oriya, and Urdu scripts. K-NN yielded overall accuracy of 92.63%, and SVM yielded over all accuracy of 94.91%.

KEYWORDS: script identification, bi-script, LBP, K-NN, SVM.

Date of Submission: 06-09-2018

Date of acceptance: 22-09-2018

I. INTRODUCTION

Research in script identification is not a new one [1]. Past many years, methods have been proposed for script identification from scanned document images. More research is done in identification of script from printed documents compared to research in script identification from handwritten documents. Handwritten documents are common in developing countries like India. Moreover, many scripts are used in India compared to any other country in the world [1-2]. Handwritten and printed documents occur in single script as well as in multi-script, bi-script in common [figure 1]. Formally, script identification facilitates Optical Character Recognition (OCR) system i.e. the script of the document is identified and then fed to OCR for further processing like language identification, digitizing the document, identifying the content of the document and writer identification.

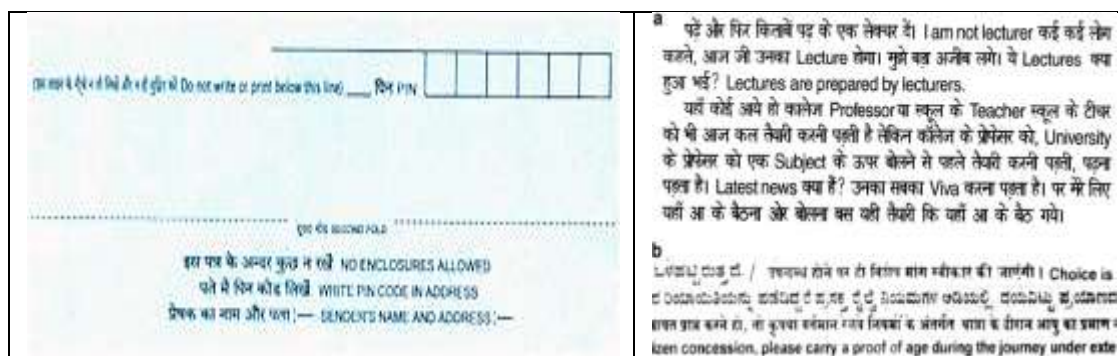


Figure 1: Sample bi-script document images

Script identification from a document image is performed in following ways: by extracting block or by segmenting line and/or segmenting word from the document. That is, generally, script identification is performed at block level, line level or word level. In case of multi-script document, line-level or word level

identification is preferred. If document appears in single script, a block of the document is sufficient to identify the script type [3].

In the context of India, script identification is challenging and complex task because of many scripts and languages being used officially and that many of the documents are multi-script in nature (Figure 1). Each and every state in India has some unique culture and language. There are 22 official languages in India including English. There are officially identified 13 scripts used to write different languages. Sample images of scripts are shown in figure 2. Moreover, each state has its official (regional) language and also uses English language for communication. Languages in the neighboring states also influence the document type to be of two-script, tri-script or in general multi-script document. Hence, automatic identification of script type from Indian documents is a challenging task attracting lot of researchers to work in this field. Different types of feature extraction techniques are proposed in the literature. Different classification models are used in classifying the script based upon the features extracted.

A brief review of literature for script identification from handwritten document images describing various feature extraction and classification techniques is presented below. Md. Obaidullah et.al [4] proposed a method for handwritten script identification based on structural, directional and texture based features. The features were extracted from document image in binary and input to MLP for text identification.


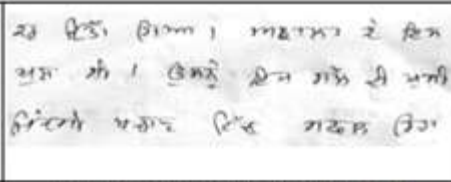

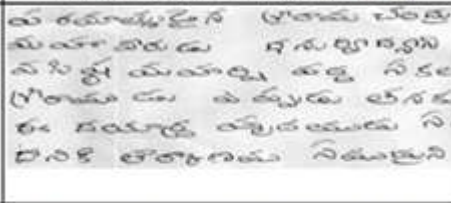
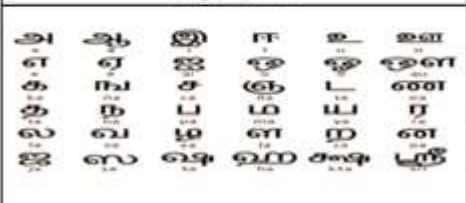
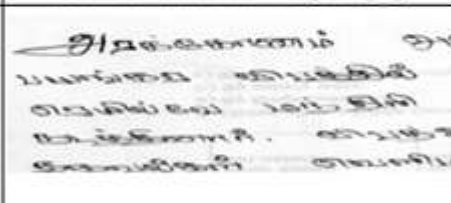
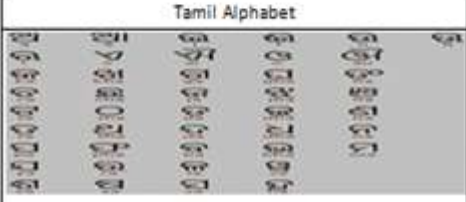
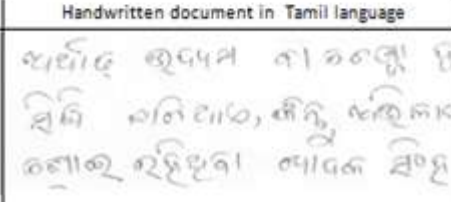
	
Punjabi Alphabet	Handwritten document in Punjabi language
	
Telugu Alphabet	Handwritten document in Telugu language
	
Tamil Alphabet	Handwritten document in Tamil language
	
Oriya Alphabet	Handwritten document in Oriya language

Figure 2: Script images (alphabet and handwritten text)

Rajput et.al [5] proposed SIFT based script identification at line-level. Text lines were extracted using horizontal profile and connected component analysis, SIFT features vector were computed. Euclidean distance measure is used for distance computation and to classify a document as belonging to specific script. Experiments were carried out for bi-script and tri-script classification. Script dependent and script independent feature based classification of script at block level, line level and word level is proposed by S.K. Obaidullah [8]. Classification of script is carried out using MLP and random forest classifiers. B. Shi et.al. [9] Incorporated deep learning based script identification from natural images. They have presented a two stage features extraction methodology i.e., feature extraction and discriminative clustering at mid-level representation and global fine tuning at second stage by modeling feature extraction and classification into one neural network by

transferring learned parameters at first stage. Back propagation is used to train the network. Long Short Term Memory (LSTM) architecture based script identification is reported by Adnan Ul-Hasan in [10].

DCT and distance transform based script identification from handwritten document is proposed by S.K. Md. Obaidullah et.al. [12]. Bangla, Roman, Devanagari and Oriya scripts are considered for performing experiments. Greedy attribute selection [GAS] approach for script identification at block level proposed by Md. Obaidullah et.al. [13]. Six Indic scripts namely, Devanagari, Bangla, Roman, Oriya, Urdu and Malayalam are considered for study. Feature vector based on texture (BRT, BDCT, BFFT and BDT) are used to classify bi-script and tri-scripts. Gabor filter and morphological re-construction is proposed by Nibaran Das et.al. [14] to identify the Indic scripts using MLP classifier. Prasanthkumar et.al. [15] proposed an automatic separation of text line followed by word segmentation and computed morphological and Gabor features. Classification is carried out using MLP, SVM, and KNN classifiers with the observation that MLP classifier yielded better results over SVM and KNN. Rajput et.al [16] presented line level script recognition using Gabor filter combined with DCT and wavelets. Nine Indian scripts were used for performing experiment, SVM classifier reported better recognition accuracy compared to the performance by KNN classifier.

Profile based script identification technique is presented by M.C. Padma et.al. [17]. Distinct features at top and bottom of the text line are considered for feature extraction. Features are extracted by calculating density ratio, pixel distribution, max pixel density at the bottom row of text line and density of connected components at the top of the text line. Classification is performed using KNN. Mallikarjun Hangarge et.al. [18] proposed a visual texture based methodology for handwritten script recognition at line level and block level. Horizontal projection profile is used to extract non-touching text line from document image. Morphological filters are utilized to extract 13 spatial spread features. KNN classifier is used to classify the script.

Compared to script identification from printed documents, the task of script identification from handwritten documents is difficult due to non uniform writing style, non uniform appearance of words and lines / uneven gap between words and lines and varying orientation of lines resulting in touching/ overlapping lines. Segmentation of such text lines are performed in our proposed work using connected component analysis.

Texture is a repeated pattern of the structure with regular intervals. It refers to surface characteristics and appearance of an object in terms of size, shape, density, arrangement, proportion of its elementary parts. Texture feature extraction refers to extracting texture features from objects of interest and is a key function in various image processing applications [19-24]. Several methods are presented in literature for texture based featured extraction (e.g. statistical, model based, and transform based textures). Local binary pattern (LBP) operator is a statistical based approach that describes the texture features in terms of smallest primitives called textons (or, histograms of texture elements). The operator produces different binary codes representing different types of curved edges, spots, flat areas, etc. and hence effective in object classification [30].

In this paper we present LBP based feature extraction for script type identification at line level. A total of 9 scripts are identified for study namely, Kannada, Hindi, Urdu, Malayalam, Punjabi, Telugu, Tamil, Oriya and English. Block diagram of the proposed work is shown in figure 3. Rest of the paper is presented as below. Section 2 describes data collection and preprocessing. Feature extraction and classification is presented in Section 3. Experimental results are discussed in section 4 and conclusion is given in section 5.

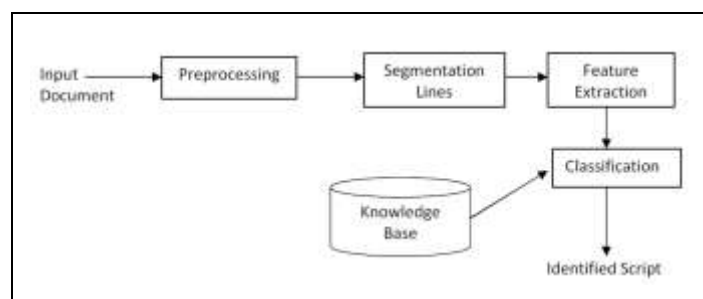


Figure 3: Block diagram of proposed system

II. DATA COLLECTION AND PREPROCESSING

Handwritten documents are collected from the persons with different age groups and professions. A total of 230 handwritten documents are collected. The collected handwritten documents are scanned using HP flatbed scanner. Details of the documents collected are presented in figure 4.

Script type	Number of document pages	Script type	Number of document pages
Kannada	40	Tamil	10
Hindi	22	Telugu	30
Malayalam	30	Urdu	32
Oriya	17	English	31
Punjabi	18		

Figure 4: Handwritten document image details

The scanned documents of gray scale are then converted to binary images by applying Otsu's thresholding method. Median filter is applied to remove salt and pepper noise. Profile based and connected component analysis is performed to segment the lines. The methodology used is described in our earlier work [6, 7]. Extracting the touching/overlapping lines is a challenging task. In order to extract touching lines, a bounding box is incorporated on every character of a line of both touching and overlapping. The midpoint of the all the bounding boxes is calculated and compared with average height of the line image and the components are labeled belonging to upper line or lower line. Example images depicting extraction of lines from the documents is shown in the figures 5 through 8. The proposed method successfully extracts the lines that appear touching/overlapping.

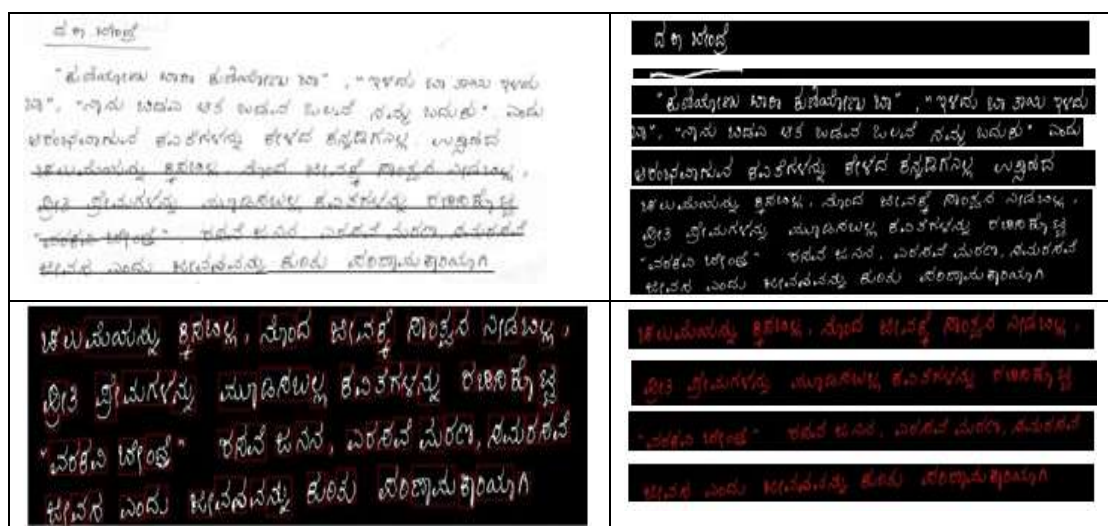


Figure 5: Segmented text line with little Orientation and Overlapping

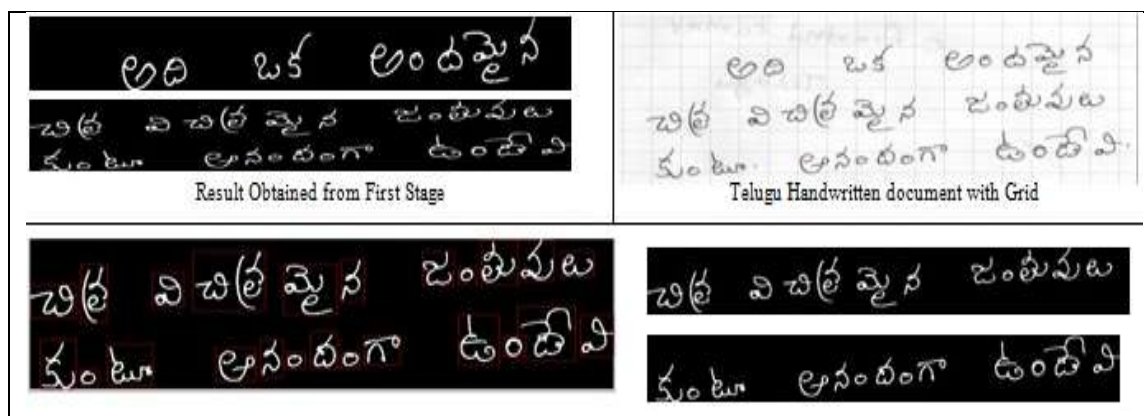


Figure 6: Segmentation of Multi-oriented text lines of Telugu Script

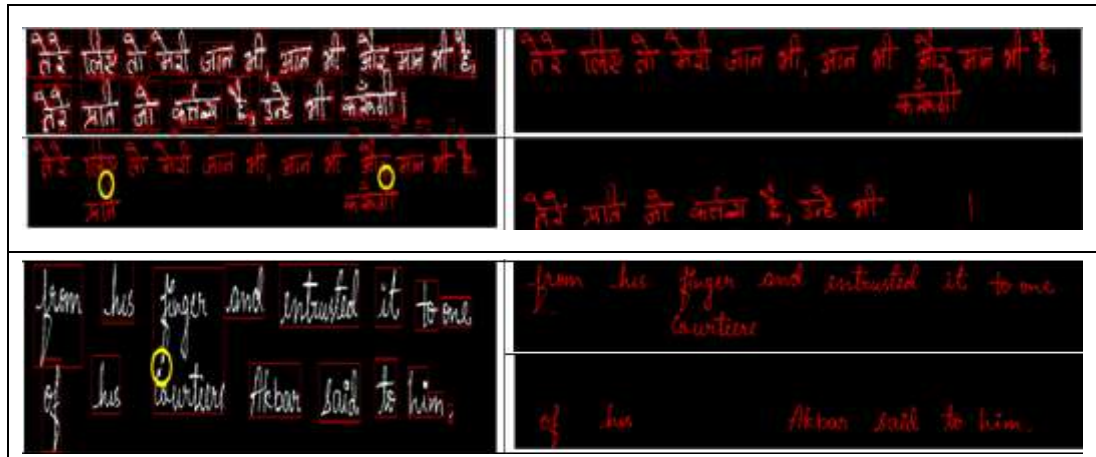


Figure 7: Segmentation of touching lines



Figure 8: Segmentation of touching lines



Figure 9: Samples images of text lines rejected

III. FEATURE EXTRACTION AND CLASSIFICATION

LBP operator [31] is used for feature extraction. LBP is a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. LBP operator produces a binary code by thresholding a 3x3 neighborhood by the gray value of its center. The histogram of these labels can then be used as a texture descriptor. Most important property is that LBP operator is robust to monotonic gray-scale changes caused and it has the advantage of simple implementation. Many related approaches based on the original implementation have been developed for texture and color texture segmentation.

The procedure for computing the binary code is as follows. In its basic form of implementation, LBP operator takes 3 x 3 neighborhood of a pixel and generates a binary 1 if the neighbor of the center pixel has the larger value than the center pixel. The operator generates a binary 0 if the neighbor is less than the center. For 8-neighborhood, an 8 digit binary number is generated which is represented as unsigned integer, making it a compact description. An example of binary code generation is given in figure 10.

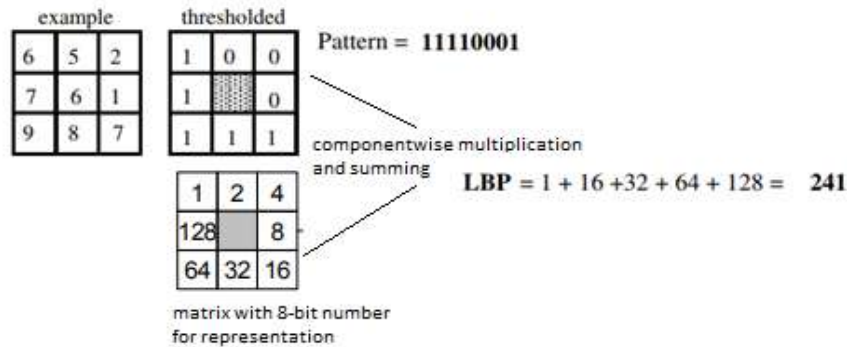


Figure 10: Calculating the LBP code

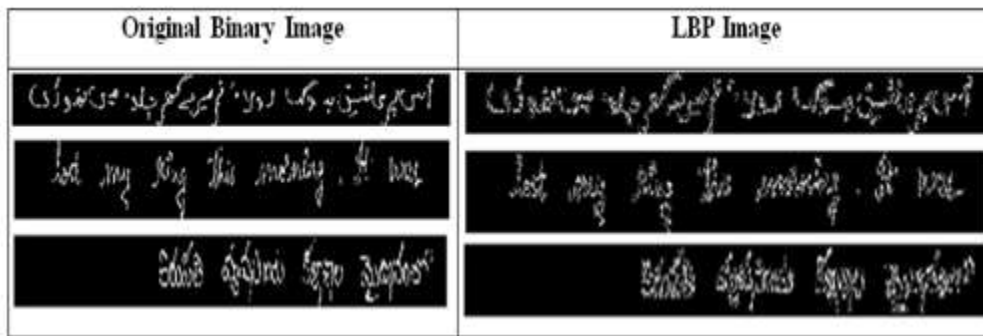


Figure 11: Sample Binary and LBP images

Feature extraction algorithm is presented below.

Input: pre-processed line image in binary

Output: LBP feature vector

1. Divide the image into 3x3 cells and for each cell compute the compact representation of LBP operator generated binary code.
2. Compute the histogram of each cell with combination of pixel smaller and greater than the centre pixel value.
3. Normalize the histogram to neighborhood pattern of (8, R), where R=1.
4. Feature vector of entire window is obtained by concatenating histogram of each cell having 59 uniforms pattern (out of 256).

Classification of the script type is carried out using KNN and SVM classifiers. The KNN is a lazy classifier that compares the test image feature vector with that of feature vector of all the images used for training and labels the test image to be of a specific script type using Euclidean distance measure [32]. The SVM classifier is trained to generate support vectors and the feature vector of test image is input to the trained SVM for script type identification. The working procedure of SVM can be found in [33].

IV. EXPERIMENTAL RESULTS

Experiments are carried out on scripts types namely, Kannada, Hindi, Urdu, Malayalam, Punjabi, Telugu, Tamil, Oriya and English respectively. A total of 230 handwritten documents are considered for experimentation. Preprocessing and segmentation of non-overlapping line is carried out using our earlier work [6, 7] and results are shown from fig. 5-8. Text lines with only 40% of occupancy of text are rejected by considering it as not a complete line. Sample images of text lines rejected are shown in figure 9. Figures 10 through 11 shows output samples images of feature extraction using LBP at line level proposed in this paper. 630 text line images are used to test the proposed method.

The overall recognition accuracy obtained using K-NN classifier is 92.63% and 94.91% accuracy is obtained using SVM classifier. Two-fold cross-validation is performed for computing the accuracy of the classification results. The details of the results obtained for bi-script documents are presented in Table 1. The results obtained for tri-script classification is shown in Table 2. SVM classifier performs better over KNN classifier.

Table1: Bi-Script Classification Using KNN and SVM classifiers

SL. NO	SCRIPT	KNN	SVM
1	Kannada	92.86	95.71
	English	98.57	100
2	Hindi	98.57	100
	English	98.57	98.57
3	Malayalam	92.86	95.71
	English	94.29	85.71
4	Oriya	92.86	98.57
	English	87.14	91.43
5	Punjabi	90	95.71
	English	90	90
6	Tamil	94.29	95.71
	English	88.57	85.71
7	Telugu	95.71	100
	English	94.21	95.71
8	Urdu	91.43	100
	English	90	87.14
Over All Accuracy		93.12%	94.73%

Table2: Tri-Script Classification Using KNN and SVM

SL. NO	SCRIPT	KNN	SVM
1	Kannada	95.71	92.86
	Hindi	98.57	100
	English	95.71	98.57
2	Malayalam	100	100
	Hindi	84.29	94.29
	English	94.29	90
3	Oriya	95.71	97.14
	Hindi	88.57	94.29
	English	87.14	91.43
4	Tamil	98.57	100
	Hindi	91.43	88.59
	English	72.86	81.43
5	Telugu	97.14	100
	Hindi	91.43	100
	English	92.86	95.71
6	Urdu	97.14	100
	Hindi	88.57	98.57
	English	88.57	88.57
Over All Accuracy		92.14%	95.08%

V. CONCLUSION

In this paper, an efficient approach for script type identification is proposed based upon the features extracted using LBP operator. Text lines are extracted using profile and connected component analysis. Our proposed text line extraction method successfully extracts touching/overlapping lines that appear in handwritten documents. Experiments are performed at bi-script and tri-script level using K-NN and SVM classifiers. Performance of SVM classifier is better over K-NN classifier in terms of overall recognition of script type

identification. The proposed work can be experimented on handwritten document images at block level and word level which is our future work.

REFERENCES

- [1]. Sahare, P., & Dhok, S. B. (2017). Script identification algorithms: a survey. *International Journal of Multimedia Information Retrieval*, 6(3), 211–232. doi:10.1007/s13735-017-0130-2.
- [2]. Miguel A. Ferrer, Aythami Morales and Umapada Pal “LBP Based Line-wise Script Identification”, 2013 12th International Conference on Document Analysis and Recognition, 1520-5363/13 \$26.00 © 2013 IEEE DOI 10.1109/ICDAR.2013.81.pp.369-373.
- [3]. K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, T. Yibulayin , “ Script Identification of Multi-Script Documents: a Survey ” , IEEE Access DOI 10.1109/ACCESS.2017.2689159 volume 5 2017, pp. 6546-6559.
- [4]. Sk Md Obaidullah, Chayan Halder2, K. C. Santosh, Nibaran Das, Kaushik Roy “Automatic Line-Level Script Identification From Handwritten Document Images - A Region-Wise Classification Framework For Indian Subcontinent” *Malaysian Journal of Computer Science*. Vol. 31(1), 2018 , pp 63-84.
- [5]. G G Rajput and Suryakant Baburao Ummappure “ Line-wise Script identification from handwritten document images using SIFT method” Proceedings of the International Conference on Intelligent Computing Systems (ICICS 2017 – Dec 15th -16th 2017) organized by Sona College of Technology, Salem, Tamilnadu, India Elsevier’s SSRN eLibrary – Journal of Information Systems & eBusiness Network -ISSN: 1556-5068 pp117-125.
- [6]. G. G. Rajput , Suryakant B. Ummappure and Preethi N. Patil, " Text Line Extraction from Handwritten Document images using Histogram and Connected Component Analysis," *International Journal of Computer Applications (0975 - 8887)* National conference on Digital Image and Signal Processing , D1SP 2015. Pp 11-17.
- [7]. G. G. Rajput , Suryakant B. Ummappure and Panditkumar Patil, "Separat ion of Touching or Overlapping Lines from Handwritten Document images using Histogram and Connected Component Analysis," *International Journal of Computer Applications (0975 8887)*National Conference on Digital Image and Signal Processing 2016.
- [8]. Sk Md Obaidullah , K. C. Santosh, Chayan Halder, Nibaran Das ,Kaushik Roy “Automatic Indic script identification from handwritten documents: page, block, line and word-level approach” *Int. J. Mach. Learn. & Cyber.* DOI 10.1007/s13042-017-0702-8 .
- [9]. Baoguang Shi,XiangBai , CongYao “Script identification in the wild via discriminative convolutional neural network” *Pattern Recognition* 52 (2016)448–458.
- [10]. Adnan Ul-Hasan, Muhammad Zeshan Afzal, Faisal Shafait, Marcus Liwicki, and Thomas M. Breuel “A Sequence Learning Approach for Multiple Script Identification”. 978-1-4799-1805-8/15/\$31.00 ©2015 IEEE pp 1046-1050.
- [11]. G. G. Rajput and Suryakant BaburaoUmmappure “Script Identification from Handwritten Documents using SIFT Method” *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPSCI-20 17)* 978-1-5386-0814-2/17/\$31.00 ©2017 IEEE, pp 520-526.
- [12]. Sk Md Obaidullah, Rownaqul Karim, Sujal Shaikh, Chayan Halder, Nibaran Das, KaushikRoy “Transform Based Approach for Indic Script Identification from Handwritten Document Images” 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 978-1-4673-6823-0/15/\$31.00 ©20 15 IEEE.
- [13]. Sk Md Obaidullah, Chayan Halder, Nibaran Das, KaushikRoy “Indic Script Identification from Handwritten Document Images – An Unconstrained Block-level Approach” , 2015 *IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)* 978-1-4799-8349-0/15/\$31.00 ©2015 IEEE pp 213-218.
- [14]. Sk Md Obaidullah, Nibaran Das, KaushikRoy “ Convolution Based Technique for Indic Script Identification from Handwritten Document Images” *I.J. Image, Graphics and Signal Processing*, 2015, 5, 49-57 Published Online April 2015 in 12. MECS (<http://www.mecs-press.org/>) DOI: 10.5815/ijgisp.2015.05.06.
- [15]. Prasanthkumar P V and Dileesh E D “Word level Script and Language identification for Unconstrained handwritten document images” 2014 3rd International Conference on Eco-friendly Computing and Communication Systems , 978-1-4799-7002-5/14 \$31.00 © 2014 IEEE DOI 10.1109/Eco-friendly.2014.78.
- [16]. Dr. G.G. Rajput, Anita H.B “Handwritten Script Recognition at Line Level – A Multiple Feature Based Approach” *International Journal of Engineering and Innovative Technology (IJEIT)* Volume 3, Issue 4, October 2013. Pp 90-95.
- [17]. M. C. PADMA, P. A. VIJAYA “SCRIPT IDENTIFICATION FROM TRILINGUAL DOCUMENTS USING PROFILE BASED FEATURES ” *International Journal of Computer Science and Applications, Technomathematics Research Foundation* Vol. 7 No. 4, pp. 16 - 33 , 2010.
- [18]. Mallikarjun Hangarge and B.V.Dhandra “Offline Handwritten Script Identification in Document Images” , *International Journal of Computer Applications (0975 – 8887)* Volume 4 – No.6, July 2010.
- [19]. Guo Xian Tan, Christian Viard-Gaudin, Alex C. Kot “Information Retrieval Model for Online Handwritten Script Identification” 2009 10th International Conference on Document Analysis and Recognition , 978-0-7695-3725-2/09 \$25.00 © 2009 IEEE , DOI 10.1109/ICDAR.2009.162 pp 336-340.
- [20]. M.C. Padma, P. A. Vijaya “Monothetic Separation of Telugu, Hindi and English Text Lines from a Multi Script Document” , *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009* , 978-1-4244-2794-9/09/\$25.00 ©2009 IEEE pp 4870-4875.
- [21]. U. Pal, B.B. Chaudhuri “Identification of different script lines from multi-script documents”, *Image and Vision Computing* 20 (2002) 945–954.
- [22]. U. Pal, S. Sinha and B. B. Chaudhuri “Multi-Script Line identification from Indian Documents”, *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR’03)* 0-7695-1960-1/03 \$17.00 © 2003 IEEE.
- [23]. S. Chanda, U. Pal and F. Kimura “Identification of Japanese and English Script from a Single Document Page” , *Seventh International Conference on Computer and Information Technology*, 0-7695-2983-6/07 \$25.00 © 2007 IEEE , DOI 10.1109/CIT.2007.109 pp 656-661.
- [24]. S. Ben Moussa, A. Zahour, A. Benabdelhafid, A.M. Alimi “Fractal-Based System for Arabic/Latin, Printed/Handwritten Script Identification” , 978-1-4244-2175-6/08/\$25.00 ©2008 IEEE.
- [25]. Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, Radek Grzeszczukand Bernd Girod “Robust Text Detection in Natural images with Edge-Enhanced Maximally Stable External Regions”.

- [26]. Álvaro González, Luis M. Bergasa , J. Javier Yebe “Text location in complex images”, Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) , 14 February 2013, Publisher: IEEE, Electronic ISBN: 978-4-9906441-0-9 Print ISBN: 978-1-4673-2216-4.
- [27]. Yao Li and Huchuan Lu , “Scene Text Detection via Stroke Width”, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan ,978-4-9906441-1-6 ©2012 IAPR pp. 681-684.
- [28]. Luk'a's Neumann Ji'r'i Matas , “Real-Time Scene Text Localization and Recognition” , 25th IEEE Conference on Computer Vision and Pattern Recognition,CVPR 2012, June 16-21, Providence, RI, USA.
- [29]. Gururaj Mukarambi, Satishkumar Mallapa and B.V.Dhandra “Script Identification from Camera Based Tri-Lingual Document”, 2017 IEEE 3rd International Conference on Sensing, Signal Processing and Security (ICSSS), 978-1-5090-4929-5©2017 IEEE, pp 214-217.
- [30]. Zhang, J., & Tan, T. (2002). Brief review of invariant texture analysis methods. Pattern Recognition, 35(3), 735–747. doi:10.1016/s0031-3203(01)00074-7.
- [31]. Matti Pietikäinen (2010) Local Binary Patterns. Scholarpedia, 5(3):9775., revision #183551
- [32]. Oliver Sutton ,”Introduction to k Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction”, February, 2012.
- [33]. B. Berwick, “An idiot ’ s guide to support vector machiens (SVMs), ”in 6.034 Artificial Intelligence–Recitations , MIT (2011).

Suryakanth Baburao Ummature "Script Identification from Handwritten Document Images Using LBP features "International Journal of Computational Engineering Research (IJCER), vol. 08, no. 09, 2018, pp 13-21