OPTICAL CHARACTER RECOGNITION SYSTEM FOR INDIAN LANGUAGES USING BHARATI SCRIPT BASED CNN

Mr. B. Hari Kumar[1], M. Bharathi[2], K. Poornima[3], M. Swathi[4], K. L. Sravanthi[5], UG Students[2,3,4,5],
Department of Electronics and Communication Engineering, Vignan's Institute of Engineering for Women
Mr. B. Hari Kumar[1], Assistant Professor, Department of Electronics and Communication Engineering, Vignan's Institute of Engineering for Women

ABSTRACT

There are different challenges in the field of computer vision. Two such problems are text detection and script identification in natural scene images. Deep neural networks (DNNs) have recently been used to solve these problems. Script identification is considered an important prerequisite of any end-to-end multilingual text reading system, which consists of two stages. The first stage is detection, and the second stage is identifying the language of the detected text. Text detection is a challengingtask because text can appear in different sizes, orientations, andfonts and canbe overlaid on complex backgrounds. The most common approach for text detection is to use a combination of computer vision techniques such as edge detection, blob analysis, and connected component analysis. Script identification is important because different scripts have different characteristics and may require different processing techniques for further analysis, such as optical character recognition (OCR). However, state-of-the-art CNN and DNN-based models have achieved high accuracy rates, with some studies reporting accuracy rates of over 95% for script identification tasks. The accuracy varies based on the different scripts.

Keywords: Optical character recognition, Convolution neural network, Deep neural network, Deep learning.

INRODUCTION

OCR is a system that provides full alpha-numeric recognition of printed characters at electronic speed by simply scanning the image. OCR systems are made up of both hardware and software. By using OCR, we can also edit the non-editable text. Optical Character Recognition (OCR) is the process that converts an image that contains text into machine-readable text. The Roman script is the common script for the Western European languages, which are expressed in English, and there are only 26 characters in English, so it becomes easy to remember. Most Indian scripts are abugida writing systems. When the vowel is present next to the consonant in a word, then it is written as diacritics on the consonant. Samyuktakshar, or composite character, is defined as a sequence or series of diacritics with consonants. Similar to the Roman script, the Bharati script can be used to express the majority of Indian languages. In this paper, the OCR system for recognizing Bharati script is presented. The main advantage of Bharati script is that it can serve as a common OCR for the majority of the Indian languages because it can be used as the common script for the majority of the Indian languages.

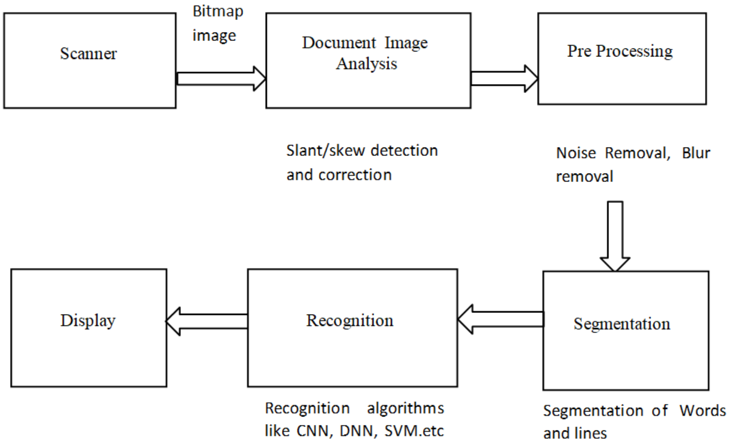

Scanner
Bitmap image
Document Image Analysis
Pre Processing
Slant/skew detection and correction
Noise Removal, Blur removal
Display
Recognition
Segmentation
Recognition algorithms like CNN, DNN, SVM.etc
Segmentation of Words and lines

Fig. 1. Block Diagram of optical character recognition

Analysis and Graphical Representation.

METHODS	TELUGU LANGUAGE ACCURACY
CNN	92.5
KNN	90
DNN	94.5
SVM	98

Table 1. Accuracy for Telugu language using different methods

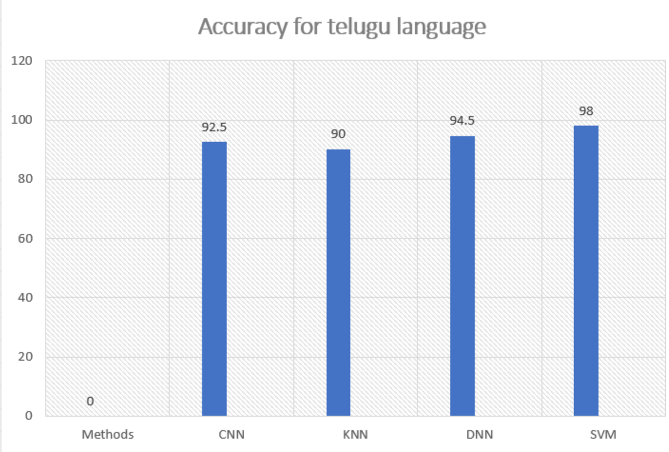


Fig 2. Graphical representation of accuracy for Telugu language using different methods

SCRIPT IDENTIFICATION METHODS

Printed or handwritten are the most commonly identified script documents. Along with these hybrid documents, this is addressed now. So according to this, content-type documents can be classified into three categories: printed, handwritten, and hybrid.

The identification of printed, handwritten, and hybrid documents is discussed below. For each document type, there are different methods used to perform script identification: page, paragraph, text block, text line, word, or character.

A) Script identification in printed documents

The main sources of printed documents are books, magazines, journals, dictionaries, etc.
1.Page/Paragraph/Text block Level Script Identification
Most research on printed documents' script identification has been carried out at the page level. The actual languages can be identified by the projection profiles of words and character shapes. Research on printed document script identification can also be conducted at the text-book level.

- Hochberg et al. [16] used cluster-based templates for discriminating 13 different scripts.
- Splitz [36] proposed a language identification scheme where the words of 26 different languages were classified into Latin-based scripts and Han-based scripts.
- Jie Ding et al. [11] presented a method that uses a combined analysis of several discriminating statistical features to categorize European language scripts.
- Pal and Chaudhuri [6] developed a system for identifying Bangla and Hindi scripts using a classification tree.
- Peake and Tan [24] proposed a method based on multiple Gabor filters and grey-level co-occurrence matrices to extract the texture features of five major scripts.

2 . Text-line Level Script Identification

In text line-level script identification, a textbook is first divided into lines.

- Pal and Chaudhuri [22] developed an automatic technique for separating text lines using script characteristics and shape-based features.
- Chanda et al. [3] proposed developing an automatic technique for the identification of Japanese and English script portions from a single line of a printed document.
- Padma et al. [21] developed a monothetic algorithm model to identify and separate Telugu, Hindi, and English text lines from a printed multilingual document.
- Prakash et al. [26] proposed a simple and efficient technique for script identification for Kannada, Hindi, and English text lines.
- Ferrer et al. [12] proposed a LBP-based line-wise script identification system to identify ten different scripts.

3 . World Level Script Identification

- Dhanya et al. [17] presented a successful method for identifying script at the word level in a bilingual document containing Roman and Tamil scripts.

- Jaeger et al. [17] used a Gabor filter analysis of textures and a multiple classifier to identify Arabic, Chinese, Hindi, and Korean scripts at the word level.
- Dhandra et al. [8, 9] proposed an automatic technique for script identification based on the morphological reconstruction of two printed scripts (at the word level): Telugu and Devnagari.
- Chanda et al. [4] proposed a SVM-based method for the identification of printed English and Thai scripts at the word-level from a single line of a document page.
- Chanda [5] proposed a SVM-based method for the identification of the Sinhala, Tamil, and English scripts from a single document page.

4: Character Level Script Identification

- Pal and Sarkar [23] used a combination of topological, contour, and water reservoir concept-based features to identify printed Urdu script.
- Rani et al. [27] carried out experiments on multi-font and multi-sized characters with Gabor features and gradient features to identify Gurumukhi and English scripts at the character level or numeral level.

B) Script identification in handwritten documents

- The script identification of handwritten documents is more challenging than the script identification of printed documents. The first study on handwritten script identification was carried out by Chaudhuri [7] and Hochberg [16].

1. Page/Paragraph/Text Block Level Script Identification

- Namboodiri and Jain [19] proposed an online handwritten script recognition system for classifying six major scripts at the global level.
- Hiremath et al. [15] identified a method based on texture features for script identification in a handwritten document.
- Ghosh and Shivaprasad [13] proposed a handwritten script identification method in which a "possibilistic" approach was used for cluster analysis.

2. Text-line Level Script Identification

- Namboodiri and Jain [19] proposed a method to classify words and lines into one of the six major scripts: Arabic, Cyrillic, Devanagari, Han, Hebrew, and Roman.

3. Word-Level Script Identification

A run-length smoothing algorithm was used to segment the document pages into lines and then into words. Fractal-based, busy-zone, and topological features were used along with the neural network classifier for script identification.

- Roy et al. [28] proposed a word-wise handwritten script identification method for Bangla and English script identification at the global level.
- Roy et al. [29] developed a technique for script separation of handwritten documents in Bangla, Roman, and Devanagari scripts.

- Zhou et al. [38] proposed a script separation technique for roman and Oriya scripts.
- Sarkar et al. [32] presented an automatic separation system for word-level script identification for Bengali and Devanagari.
- Dhandra et al. [10] used a two-stage approach. In the first stage, some global and local features were applied to identify the text words. In the second stage, the numerals written in different scripts were identified. Here Kannada, Devanagiri, and handwritten Roman documents were considered.

Roy et al. [30] proposed a handwritten script identification system for bi-script documents written in Persian and Roman scripts.

Obaidullah et al. [20] proposed a scheme to identify the six popular languages—Bangla, Devanagari, Malayalam, Urdu, Oriya, and roman scripts in Indian languages—and compared performance using different well-known classifiers.

Script identification in hybrid documents

Hybrid documents include printed and handwritten texts.

Ben Moussa et al. [2] proposed a multi-lingual automatic identification of Arabic and Latin.

Benjelil et al. [1] proposed an accurate system based on a steerable transform for Arabic and Latin script identification systems.

Saidani et al. [31] made a successful attempt to identify the Arabic or Latin script of a mach ine- printed or handwritten document at the word level.

Script identification in video frames and camera-based images

The extraction of images from the video frames or camera-base images has not been explored more till now. In video and camera-based script identification methods, the first step is the extraction of textual information, which is an important and very complex task.

Page/Paragraph/Text Block Level Script Identification

Glavata and Freisleben [14] presented, by using low-level features, an approach for discriminating between Latin and Ideographic scripts.

Zhao et al. [37] proposed a new spatial-gradient-based feature (SGF) for script identification at the block level for six scripts: Arabic, Chinese, English, Japanese, Korean, and Tamil.

Text-line Level Script Identification

Phanet et al. [25] proposed two features: smoothness and cursiveness, for video script identification at text-line level.

World Level Script Identification

Sharma et al. [33] used different classifiers like ANNs and SVMs for English, Bengali, and Hindi script identification.

Shiva kumara et al. [35] developed word-level script identification methods by using new gradient angular features for Arabic, Chinese, English, Japanese, Korean, and Tamil scripts.

Sharma et al. [34] presented a bag of visual words based on word-wise script identification from video images for five different south Indian languages.

Character-level script identification
Li and Tan [18] proposed a script identification based on statistical features technique to identify character-level English, Arabic, and Chinese scripts of camera-based images.

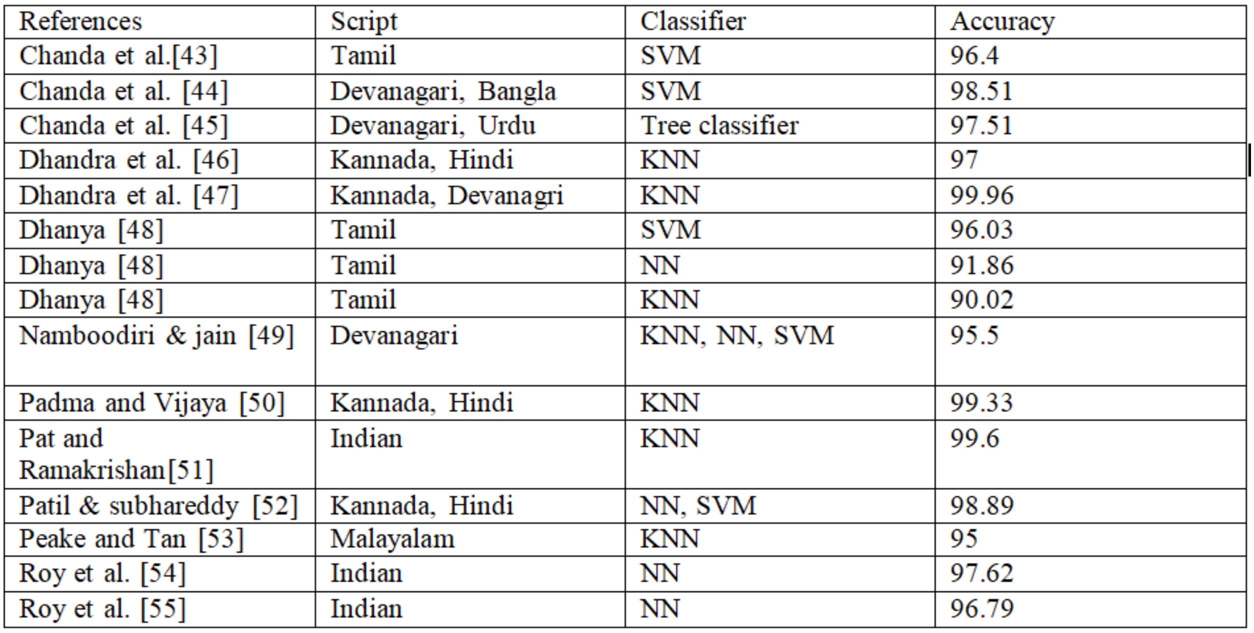
RELATED WORK

In the year 2007, Pal et al. described offline recognition of Telugu numerals but not of regular characters, which has very little work on offline recognition of Telugu script. Bindu Philip et al. (2009) [41] have suggested a technique for recognizing Malayalam characters in Malayalam text documents by using an SVM (support vector machine) method. And the recognition rates of the proposed algorithm are between 95.31 and 90.22. Pal et al. (2012) report results for only the pure vowels (V) and consonants (C), and the combination of both CV for offline handwritten character recognition of Telugu is not possible. And they also consider very small data sets. A Multilingual OCR System for Indian Scripts by S. R. Mahadeva Prasanna, K. B. Raja, and K. V. N. Sunitha [42]: This paper presents a multilingual OCR system for Indian scripts, including the Bharati script. The system uses a combination of feature extraction techniques and a neural network-based classifier to recognize characters from different scripts with high accuracy.

CONVOLUTIONAL NEURAL NETWORK

The convolutional neural network (CNN) is a subtype of neural network that is mainly used for applications in speech and image recognition. CNN requires very little pre-processing data compared to other learning algorithms. CNN automatically detects important features without any human supervision. CNN is a supervised type of deep learning, most commonly used in image recognition and computer vision. The name "Convolution" is derived from a mathematical operation, a specialized kind of linear operation that uses the mathematical operation instead of matrix multiplication in at least one layer of CNN. There are three layers that are involved in CNN.
1. Convolution Layer: The convolution layer is one of the fundamental layers in a convolutional neural network (CNN). Its primary function is to apply a set of filters, also known as kernels or weights, to the input image or feature map. The convolution operation involves sliding the filter over the input image, computing the element-wise multiplication between the filter and the corresponding portion of the input, and summing up the results. This produces a feature map that highlights certain features in the input image.
2.Pooling Layer: This layer reduces the spatial size of the feature map by performing a down- sampling operation, which helps reduce the amount of computation required in the network.
3.Fully Connected Layer: This layer takes the output of the previous layers and flattens it into a single vector, which is then passed through a series of fully connected (dense) layers to produce the final output.

SCRIPT IDENTIFICATION ANALYSIS



References	Script	Classifier	Accuracy
Chanda et al.[43]	Tamil	SVM	96.4
Chanda et al. [44]	Devanagari, Bangla	SVM	98.51
Chanda et al. [45]	Devanagari, Urdu	Tree classifier	97.51
Dhandra et al. [46]	Kannada, Hindi	KNN	97
Dhandra et al. [47]	Kannada, Devanagri	KNN	99.96
Dhanya [48]	Tamil	SVM	96.03
Dhanya [48]	Tamil	NN	91.86
Dhanya [48]	Tamil	KNN	90.02
Namboodiri & jain [49]	Devanagari	KNN, NN, SVM	95.5
Padma and Vijaya [50]	Kannada, Hindi	KNN	99.33
Pat and Ramakrishan[51]	Indian	KNN	99.6
Patil & subhareddy [52]	Kannada, Hindi	NN, SVM	98.89
Peake and Tan [53]	Malayalam	KNN	95
Roy et al. [54]	Indian	NN	97.62
Roy et al. [55]	Indian	NN	96.79

Table 2. Script identification analysis of different Indian languages.

RESULT AND ANALYSIS

In the references we have considered, only 50% of the noise is removed using the CNN method. Bharati script identification is generally considered to have a high accuracy rate. Depending on the specific process we use, the expected accuracy rate would be around 95–97%. Our future work in this direction emphasizes the development of more enriched OCR models and large datasets for all the Indian languages.

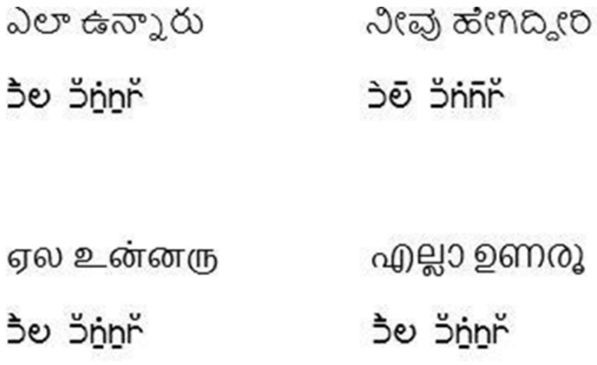


Fig 3. The representation of Telugu, Kannada, Tamil and Malayalam text in Bharati Script

CONCLUSION

Overall, developing an efficient multilingual OCR system for Indian languages through the use of Bharati script requires careful consideration of the different approaches and factors involved. With the right approach and attention to detail, it is possible to develop an OCR system that can accurately recognize and digitize text in different Indian languages, making it easier to access and analyze large amounts of information.

REFERENCES

[1] S.A. Angadi, M.M. Kodabagi, "A fuzzy approach for word level script identification of text in low resolution display board images using wavelet features", Proc. ICACCI, Aug. 22-25, 2013, pp.1804 – 1811.

[2]S. Ben Moussa, A. Zahour, A. Benabdelhafid, A.M. Alimi, "Fractal-based system for Arabic/Latin, printed/handwritten script identification", Proc. ICPR, Tampa, Dec. 8-11, pp. 1–4.

[3]S. Chanda, U. Pal, F. Kimura, "Identification of Japanese and English script from a single document page", Proc. IEEE-CIT,2007, pp.656-661.

[4]S. Chanda, O.R. Terrades, U. Pal, "SVM Based Scheme for Thai and English Script Identification", Proc. ICDAR, Parana, 2007, pp. 551-555.

[5]S. Chanda, S. Pal, U. Pal, "Word-wise Sinhala Tamil and English Script Identification Using Gaussian Kernel SVM", Proc. ICPR, Tampa, Dec.8-11, 2008, pp. 1-4.

[6]B.B. Chaudhuri, U. Pal, "An OCR system to read two Indian language scripts : Bangla and Devnagari(Hindi)",ICDAR,Ulm,1997,pp. 1011-1015.

[7]B.B. Chaudhuri, "On multi-script OCR system evaluation", Proc. Int. Workshop on Performance Evaluation Issues in Multi-Lingual OCR,1999.

[8]B.V. Dhandra, H. Maliikarjun, R. Hegadi, V.S. Malemath, "Word-wise script identification based on morphological reconstruction in printed bilingual documents", Proc. IET Int. . on Visual Information Engineering (VIE), Bangalore, Sept. 26-28, 2006, pp. 389-393.

[9]B.V. Dhandra, H. Mallikarjun, R. Hegadi, V.S. Malemath, "Word-wise Script Identification from Bilingual Documents Based on Morphological Reconstruction", Proc. Int. Conf. on Digital Inform ation Management, Bangalore, Dec. 6, 2006, pp. 389 – 394.

[10]B.V. Dhandra, M. Hangarge,"Global and local features based handwritten text words and numerals script identification", Proc. ICCIMA, Dec. 13-15, 2007, pp. 471-475.

[11]J. Ding, L. Lam, C.Y. Suen, "Classification of oriental and European scripts by using characteristic features", Proc. ICDAR, Ulm, 1997, pp.1023-1027

[12]M.A. Ferrer, A. Morales, U. Pal, "LBP based line-wise script identification", Proc. ICDAR, Washington DC, 2013, pp. 369-373.

[13]D. Ghosh, A.P. Shivaprasad, “Handwritten script identification using possibilistic approach for cluster analysis”, Journal of the Indian Institute of Science, vol.80, no. 3, 2000.
[14]J. Gllavata, B. Freisleben, “Script Recognition in Images with Complex Backgrounds”, Proc. ISSPIT, Athens, Dec. 21, 2005, pp.589-594.
[15]P.S. Hiremath,S. Shivashankar, J.D. Pujari, V. Mouneswara, “Script Identification in a handwritten document image using texture features”,Proc. IACC, Patiala, Feb. 19-20, 2010, pp. 110-114.
[16]J. Hochberg, P. Kelly, T. Thomas,L. Kerns, “Automatic script identification from images using cluster-based templates”, IEEE TPAMI, vol. 19, no. 2, 1997, pp.176-181.
[17]S. Jaeger, H. Ma, D. Doermann, “Identifying Script on Word-Level with Informational Confidence”, Proc. ICDAR, 2005, pp. 416-420.
[18]L. Li, C.L. Tan, “Script Identification of Camera-based Images”, Proc ICPR, Tampa FL, Dec. 8-11, 2008, pp. 1-4.
[19]A.M. Namboodiri, A.K. Jain, “On-line Script Recognition”, IEEE TPAMI,2004, vol. 26, no. 1, pp. 124- 130.
[20]S.M. Obaidullah, K. Roy, N.Das, “Comparison of different classifiers for script identification from handwritten document”, Proc. ISPCC, Sept.26-28, 2013, pp. 1 - 6.
[21]M.C. Padma, P.A. Vijaya, “Monothetic Separation of Telugu. Hindi and English Text Lines from a Multi Script Document”, Proc. SMC, Oct.11-14, 2009, pp. 4870 - 4875.
[22]U. Pal, B.B. Chaudhuri, “Script line separation from Indian multi-Script documents”, Proc. ICDAR, Bangalore, 1999, pp. 406-409.
[23]U. Pal, A. Sarkar, “Recognition of Printed Urdu Script”, Proc. ICDAR, Bangalore, 2003, pp. 1183-1187.
[24]G.S. Peake, T.N. Tan, “Script and language identification from document images”, LNCS vol. 1352, 8th BMVC, 1997, pp. 230-233.
[25]T.Q. Phan, P. Shiva kumara, Z. Ding, S. Lu, and C. L. Tan, “Video script identification based on text lines”, Proc. ICDAR, 2011, pp. 1240-1244.
[26]K.A. Prakash, G. Rajesh, U.A. Dinesh, M. Krishnamoorthi, N.V. Subha Reddy, “Text Line Script Identification for a Trilingual Document”, Proc. ICCCNT, 2010, pp.1-3.
[27]R. Rani, R. Dhir, G.S. Lehal, “Script Identification of Pre-Segmented Multi-Font Characters and Digits”, Proc. ICDAR, Washington DC, Aug.25-28, 2013, pp. 1150–1154.
[28]K. Roy, U. Pal, B.B. Chaudhuri, “Neural Network based Word-wise Handwritten Script Identification for Indian Postal Automation”, Proc . ICISIP, Jan. 4-7, 2005, pp. 240-245.

[29]K. Roy, K. Majumder, “Trilingual Script Separation of Handwritten Postal Document”, Proc. ICVGIP, Dec. 16-19, 2008, pp. 693-700.

[30]K.Roy,A.Alaei,U.Pal,“Wordwise Handwritten Persian and Roman Script Identification”, Proc. ICFHR, Kolkata, Nov. 16-18. 2010, pp. 628-633.

[31]A. Saidani, A.K. Echi, A. Belaid, “Identification of Machine-printed and Handwritten Words in Arabic and Latin Scripts”, Proc. ICDAR, Washington DC, Aug. 25-28, 2013, pp. 798 – 802.

[32]R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasi puri, D.K. Basu, “Word level Script Identification from Bangla and Devanagari Handwritten Texts mixed with Roman Script”, Journal of Computing, Vol. 2, No. 2, 2010,pp.103-108.

[33]N. Sharma, S. Chanda, U.Pal, M. Blumenstein, “Word-wise Script Identification from Video Frames”, Proc. ICDAR, Washington DC, Aug.25-28, 2013, pp. 867 – 871.

[34]N. Sharma, R. Mandal, R. Sharma, U. Pal, M. Blumenstein, “Bag- of Visual Words for Word-Wise Video Script Identification: A Study ”,Proc. IJCNN, Killarney, Ireland, July 12-17, 2015, pp. 1-7.

[35]P. Shiva kumara, N. Sharma, U. Pal, M. Blumenstein, “Gradient Angular-Features for Word-wise Video Script Identification”, in Proc . ICPR, Stockholm, Aug. 24-28, 2014, pp. 3098-3103.

[36]A.L. Spitz, “Determination of the Script and Language Content of Document Images”, IEEE T-PAMI, Vol. 19, No. 3, 1997, 235-245.

[37]D. Zhao, P. Shiva kumara, S.J. Lu, C.L. Tan, “New Spatial-Gradient Features for Video Script Identification”, Proc. DAS, 2012, pp. 38 – 42.

[38]L. Zhou, Y. Lu, C.L. Tan, “Bangla/English Script Identification Based on Analysis of Connected Component Profiles”, LNCS, vol. 3872, 2006,pp. 243-254.

[39]B. Hari Kumar, P. Chitra “Survey Paper Of Script Identification Of Telugu Language Using OCR” International Journal of Electronics and Communication Engineering (IJECE) ISSN(P): 2278-9901; ISSN(E): 2278-991X Vol. 8, Issue 3, Apr-May 2019; 15-20.

[40]G. Srinivas Rao, Mohammed Imanuddin, B. Hari Kumar. ”Script Identification of Telugu, English and Hindi Document Image” International Journal of Advanced Engineering and Global Technology(IJAEGT). Vol-2, Issue-2, February 2014.

[41]B. Philip and R. D. Sudhakar Samuel. “An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers” International Journal of Recent Trends in Engineering, Issue. 1, Vol. 1, May 2009.

[42] A Multilingual OCR system for Indian Scripts by S.R. Mahadeva Prasanna, K.B. Raja and K.V.N. Sunitha.

[43]S. Chanda, S. Pal, U. Pal, “World-Wise Sinhala Tamil and English Script Identification Using Gaussian Kernel SVM”, Proc. ICPR, Tampa, Dec. 8-11,2008,pp.1-4.

S. Chanda, S. Pal, K. Franke, U. Pal, “Two-stage Approach for Word-wise Script Identification”, Proc. ICDAR,2009, pp. 926-930

S. Chanda, U. Pal, “English, Devanagari and Urdu text identification”, Proc. ICDAR, 2005, pp. 538- 545

B.V. Dhandra, H. Mallikarjun, R. Hegadi, V.S. Malemath, “Word-wise script identification based on morphological reconstruction in printed bilingual documents”, Proc. IET Int. . on Visual Information Engineering (VIE), Bangalore, Sept. 26-28, 2006, pp. 389-393

B.V. Dhandra, M. Hangarge, “Global and local features based handwritten text words and numerals script identification”, Proc. ICCIMA, Dec. 13-15, 2007, pp. 471-475

D. Dhanya, A.G. Ramakrishnan, “Script Identification in printed bilingual documents”, LNCS vol. 2423, 2002, pp. 13-24

A.M. Namboodiri, A.K. Jain, “On-line Script Recognition”, IEEE TPAMI,2004, vol. 26, no. 1, pp. 124-130

M.C. Padma, P.A. Vijaya, “Entropy Based Texture Features Useful for Automatic Script Identification”, Int. J. Computer Science and Engineering, Vol. 2, No. 2, 2010, pp. 115-120

S.B. Patil, N.V. Subha Reddy, “Neural network based system for script identification in Indian documents”, Sadhana,2002, vol.27, no.1,pp.83-97.

P.B. Pati, A.G. Ramakrishnan, “Word level multi-script identification”, Pattern Recognition Letters, 2008, vol. 29, no. 9, pp. 1218-1229

G.S. Peake, T.N. Tan, “Script and language identification from document images”, LNCS vol. 1352, 8th BMVC, 1997, pp. 230-233.

K. Roy, U. Pal, B.B. Chaudhuri, “Neural Network based Word-wise Handwritten Script Identification for Indian Postal Automation”, Proc. ICISIP, Jan. 4-7, 2005, pp. 240-245

K. Roy, K. Majumder, “Trilingual Script Separation of Handwritten Postal Document”, Proc. ICVGIP, Dec. 16-19, 2008, pp. 693-700.