International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Comprehensive Survey of Script Identification in Indian Multi-Script Documents

Jagin M. Patel[1] and Bharat C. Patel[2]
[1]M. K. Institute of Computer Studies, Bharuch, Gujarat, India
[2]Smt. Tanuben and Dr. Manubhai Trivedi College of Information Science, Surat, Gujarat, India

Article Info
Volume 8, Issue 2
Page Number : 535-546

Publication Issue :
March-April-2022

Article History
Accepted: 10 April 2022
Published: 30 April 2022

ABSTRACT

Multilingual character recognition plays a crucial role in various applications, including document analysis, text-to-speech synthesis, machine translation, and optical character recognition (OCR). The diversity and complexity of Indian scripts pose significant challenges for accurate character recognition, as these scripts exhibit variations in shape, size, and structure across different languages and writing styles. This survey paper provides an overview of the current state of research and advancements in the field of multilingual character recognition for Indian scripts. The diversity and complexity of Indian scripts present unique challenges in accurately recognizing characters across different languages and writing styles. This survey aims to summarize the existing literature, methodologies, and techniques employed in multilingual character recognition for Indian scripts.

Keywords : Indian scripts, Multi-lingual, Script Identification, OCR

I. INTRODUCTION

Multi-script identification refers to the process of automatically recognizing and determining the scripts or writing systems used in a given text or document. With the increasing globalization and interconnectedness of the world, it is not uncommon to encounter multilingual or multi-script content, where different writing systems coexist within a single document or text.

Bilingual or Multilingual script identification determines the script used in a given text, especially when it contains multiple languages or scripts. This identification is important in various applications. Multilingual script identification is a crucial step in language identification systems, which aim to automatically determine the language(s) used in a given text. Language identification is essential for many natural language processing tasks, such as machine translation, information retrieval, and text-to-speech synthesis. Multilingual script identification helps in accurately identifying the scripts involved and subsequently improves the accuracy of language identification.

India is a linguistically diverse country with multiple official languages and numerous regional languages. India has a rich linguistic landscape, with languages written in different scripts like Devanagari (Hindi),

Tamil, Telugu, Bengali, Gujarati, etc. In India, documents may contain multilingual languages: one is regional, second is English and/or Hindi. In scenarios where documents or texts contain multiple languages, such as government documents, legal contracts, or multinational company databases, bilingual script identification assists in automatically identifying the script of each language within the document.

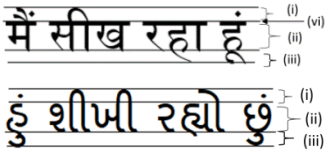
The scope of these domains extends across various applications including postal addresses, cheques, application forms, ticket booking, data entry forms and government forms. These documents often contain text written in multiple languages, and the challenge lies in accurately identifying and extracting the multilingual data. However, errors may arise during the processing of such data. Automatic detection of scripts in multilingual environments is a difficult research issue over the past two decades [1].

Characteristics of Indian scripts

Indian language scripts hold a rich cultural and historical significance and play a vital role in preserving the linguistic diversity of the Indian subcontinent. India is a multilingual nation with many officially recognized languages and numerous regional languages. Each of these languages has its unique script. Indian scripts utilize various writing systems, including syllabic.

In most Indian languages, a text line structure can be divided into three zones, as shown in Fig. 1. The upper zone refers to the area above the head-line, whereas the middle zone refers to the area below the head-line.

In many scripts, characters have a horizontal line at the upper part. This is known as "sirorekha" or the head-line. It is a feature that is used to separate one text line from another. Some scripts have no concept of head-lines but have vertical line-like layouts.



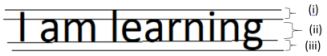


Fig.1: Example of zoning: (a) Hindi (a) Gujarati (c) English line.

(i) upper zone

(ii) Middle zone

(iii) Lower zone

(iv) head-line or Header-line or Sirorekha

The characters in the English script can be uppercase or lowercase. However, the case idea is missing from Indian scripts. Because different scripts' characters have varied shapes, a segmentation approach that works for one script might not work for another.

II. Script identification features and methods

The majority of script identification research focuses on printed [2,3,4,5,6,7, 8, 9,10,11,12,13,14] or handwritten [15,16,17,18,19,20,21,22,23] scripts in documents.

In addition to optical scanners, cameras [24, 25] and camcorders can also be used to acquire documents.

Many studies focused on individual Indian scripts or languages, such as Devanagari, Tamil, or Kannada, rather than considering a comprehensive multilingual approach. This limited the generalizability and scalability of the recognition systems across different Indian scripts.

In 1997, Spitz [26] carried out research on automatic script identification within the Han script class (Chinese, Japanese, Korean) and 23 Latin-based languages which used a method built on character shape codes.

In 2004, Pal and Chaudhuri [27] published a survey for Indian script identification and mentioned there are six writing systems namely, Logographic system, Syllabic system, Alphabetic System, Abjads, bugidas, and Featural system.

A method for script segmentation of one script may not work for another since the character forms of other scripts differ. For instance, in some scripts, such as Bangla and Devnagari, the majority of the

characters have a headline or horizontal line at the top, whereas, in Gujarati and Oriya, the characters lack headlines but have vertical line-like structures.

[12] used statistical and topological features and [28] used Gabor features derived from connected components. Traditional machine learning algorithms such as k-nearest Neighbors (k-NN) and Support Vector Machines (SVM) were commonly employed as classifiers for multilingual character recognition in Indian scripts. These algorithms were effective to some extent but often struggled to handle the variations and complexities present in different scripts and languages.

To build a multilingual character recognition system, we would need a combination of techniques that can handle those scripts. This may involve training separate models for each character recognition and integrating them into a single system. The models should be trained on a diverse dataset that includes examples of characters from multiple languages.

There are various types of features used for multi-script identification including Structural features (e.g. Circularity, Rectangularity, Component-Based, Freeman Chain Code based Feature etc.), Log-Gabor filtering, Global features (e.g. Stroke density - Vertical stroke, horizontal stroke), Pixel Density, local features (e.g. Aspect ratio, Extent, Eccentricity etc.)

The two basic categories of script identification techniques are local and global techniques [29]. In order to identify the script, local techniques examine a list of linked components (Line, Word, and Char) in the document. Character segmentation or connected component analysis is the key factor influencing script categorization. Global techniques, on the other hand, analyze text blocks or regions. Local approaches take longer than the global strategy.

Pati et. al. [30] proposed methods for page layout analysis as well as for script recognition in bilingual text documents at the word level. Many researchers proposed multi-script identification methods, but they have advantages and disadvantages.

Mohanty et. al. [31] proposed a method of bilingual (English-Oriya) script identification which loss accuracy by training two scripts in one system, some Oriya characters were mistaken for Romans, accuracy rate decreased for the thin and small-size characters. Dhanya et. al. [5] made assumption that a word must contain a minimum of four characters. The method proposed by Dhandra et. al. [6] failed to identify (i) mark '|' and broken sirorekha, (ii) touched and broken components. This method has also a problem with Arial Black font of size more than 16 points. The method proposed by Dhandra et. al. [32] failed to identify (i) mark '|' and broken sirorekha (ii) touched and broken components. This method Misclassified Boldface Tamil words. The method proposed by Chaudhari et. al. [4] has difficulty to prepare N-grams because unarranged characters exist in two scripts. They also categorised errors into Substitute, Rejection, Run on, split character and delete errors.

Obaidullah et. al. [33] provides a method for identifying 10 official Indian scripts—Bangla, Devnagari, Roman, Oriya, Urdu, Gujarati, Telegu, Kannada, Malayalam, and Kashmiri—from printed documents. For this effort, a total of 459 document pages are taken into account, and a 62 dimensional feature set is computed. Finally, an average recognition rate of 98.9% is discovered using a basic logistic classifier with 5-fold cross-validation. They used features such as Circularity, Rectangularity, Freeman Chain Code based Feature (Fig 2), Component Based Feature, mathematical and morphological features etc.

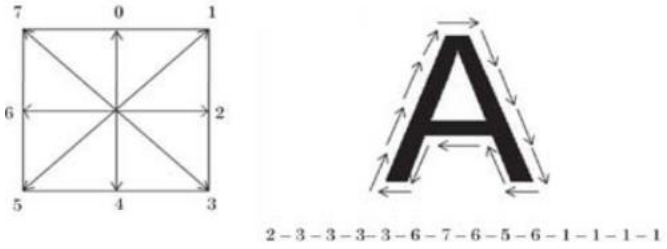


Fig. 2: Freeman chain code and an example for the alphabet "A"

The ability to differentiate one script from another depends on the fact that each script has a distinct

spatial distribution and visual characteristics. Therefore, the fundamental challenge in script recognition is to develop a method to extract these features/attributes from a given text and then categorise the script of the document in accordance with those findings. For Indic scripts, several script identification methods have been described. They can be categorised into three categories: page level, line level and Word level script identification.

Page level script identification

Large chunks of input text are needed for the script identification methods at the page level so that there is enough information to reveal the script's characteristics.

Dhandra et. al [34] proposed a method for English, Hindi, Kannada, and Urdu script identification in printed documents which is based on the morphological operation. It used a feature extractor which consists of two stages. During the first stage, morphological operation is performed using line structuring element on a document in four directions viz. horizontal, vertical, right and left diagonal. Based on the average height of all connected image components, the structuring element's length is fixed. During the second stage, average pixel distribution is found. A nearest-neighbor analysis is used for script classification.

K. Roy et al. [35] developed an algorithm to identify document level scripts from handwritten documents written in six scripts viz, Bangla, Devnagari, Malayalam, Urdu, Oriya and Roman. This algorithm derived 46 features based on circularity, fractal-based features, component analysis, small component, Freeman chaincode etc. Multi-Layer Perceptron (MLP) is used for classification purposes. A total of 160 documents were employed in the experiment, of which 32 were written in Bangla, 30 in Devnagari, 20 in Malayalam, 16 in Urdu, 30 in Oriya, and 32 in Roman character.

Padma and Vijaya [36] presented an approach for South Indian Script viz. English, Telugu, Malayalam, Kannada, Tamil, Urdu, and Hindi Identification using Haar Wavelet Features. This approach used coefficient sub bands and generate Gray level co-occurrence matrix. After that Haralick texture features are derived. In this approach, 2100 text images were used for learning while 1400 text images were used for testing. It achieved an average success of 99.68%. The image size was 600x600.

Again, Padma and Vijaya [37] presented an approach for script identification using Haralick texture features. But this time this approach identified 10 Indian scripts viz. Bangla, Devanagari, Roman (English), Gujarati, Malayalam, Oriya, Tamil, Telugu, Kannada and Urdu. In this approach, 2500 text images were used for testing. It achieved an average success of 98.24%. The image size was 600x600.

Although many script recognition methods have demonstrated good performance, they can only be used with papers that have been mechanically printed. The document images must therefore be preprocessed. The texture-based script identification approach put forward in [38] addresses this. Singhal et. al. [38] proposed an algorithm for classifying hand-written documents containing script viz. English, Hindi, Bangla and Telugu. The algorithm used de-noising, thinning, pruning, *m*-connectivity, and normalization during the preprocessing stage. After that, multi-channel Gabor filter is applied to extract Texture features. In the final step, script classification is done by applying fuzzy classification.

Hiremath et al. [39] developed an approach to script identification from handwritten documents. This approach uses texture features which are derived from co-occurrence histograms of sub-bands. This system achieved an accuracy of 97.5%. The image size was 256x256.

Text line-level script identification

Joshi et. al. [40] presented a scheme for ten Indian Script Identification from Documents. This scheme

uses hierarchical classification and characteristics taken from the log-Gabor filter bank. It is constructed with multiple orientations and at optimal scale. During the first stage, global features are used to group scripts into five classes. Sub-classification is carried out in the second stage utilizing features relevant to the script. All features are extracted globally from a given text block. The scheme has been tested on 10 Indian scripts, and it was discovered to be reasonably insensitive to changes in font size as well as robust to skew produced during scanning.

Pal and Chaudhuri [10] presented an algorithm for Script Line Separation from Indian Multi-Script documents by means of script characteristics and shape based structures. They used features such as Horizontal projection profile, Vertical run length distribution, Distribution of lowermost points of the components, Position of vertical line with respect to its bounding box, Distribution of vertical component above meanline, and Horizontal run information. This algorithm is independent of character size, font and case insensitivity.

Aithalet. al. [13] proposed an algorithm in which three scripts are separated using a horizontal projection profile. Each text line's horizontal projection profile is used to guide the feature extraction process. The system's knowledge base is built using 15 separate document images that together include roughly 450 text lines. To classify the script for a new text line, the required features are taken from the horizontal projection profile and compared with the knowledge base that has been previously saved.

Multiple feature-based approaches were presented by Rajput and Anita [41] to identify the script type from handwritten manuscripts. Here, eight well-known Indian scripts are taken into account. Gabor filters, Discrete Cosine Transform, and Daubechies Wavelets are used to extract features. Experiments are conducted to evaluate the proposed system's line-level recognition accuracy for bilingual and, later, trilingual scripts. At the line level, it achieved 100% accuracy for bi-scripts. K-NN Classifier is used for classification.

Word level

Script recognition at the word level in a multi-script document is typically more challenging than text line level script identification. This happens because of not sufficient information available from only a few characters in a word.

Pal et. al. [12] make use of horizontal profiles, statistical, topological, and stroke-based features and a tree based classifier is used for twelve Indian language scripts. Patil and Ramakrishnan [42] developed an algorithm for identifying scripts from printed documents at the word level for eleven Indian scripts viz. Bengali, English, Gujarati, Hindi, Kannada, Tamil, Telugu, Malayalam, Odiya, Punjabi, and Urdu. It used Gabor and discrete cosine transform (DCT) features.

Pal and Chaudhuri[43] describe an automated word segmentation method that can distinguish between Roman, Bangla, and Devnagari characters found in a single document. The method uses a tree structure and first uses the 'headline' feature to divide Roman script words. In Roman, there is no headline, yet it is prevalent in Bangla and Devnagari. Then, utilizing certain finer aspects of the character set, words in Bangla and Devnagari are separated. The system achieved 96.09% overall accuracy.

In one research, Sinha et. al. [44], proposed a reliable method for word-by-word extracting and identification from Indian doublet. Here, the document is divided first into lines, and the lines are subsequently divided into words. Individual script words are recognized from the documents using several topological and structural attributes such as number of loops, headline feature (fig.3(a)), water reservoir concept-based features (fig. 3(b)), profile features(fig.3(c), etc. Tested on 24210 words with various doublets, the suggested strategy averaged more than 97% accuracy. Here, they consider five major Indian doublet documents: Devnagari-English,

Devnagari-English, Telugu-English, Bangla-English, Malayalam-English, and Gujarati-English.

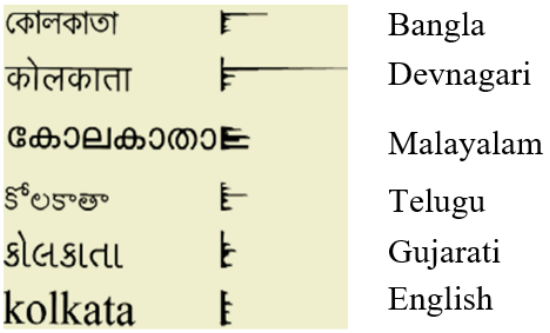


Fig. 3(a). Row-wise longest horizontal run is shown in six different script words.

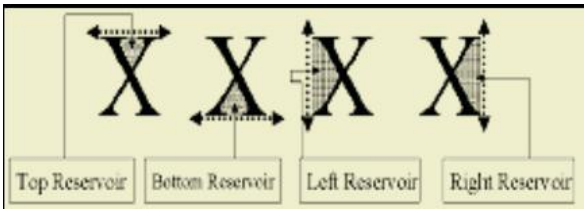


Fig. 3(b) reservoirs for English character 'X

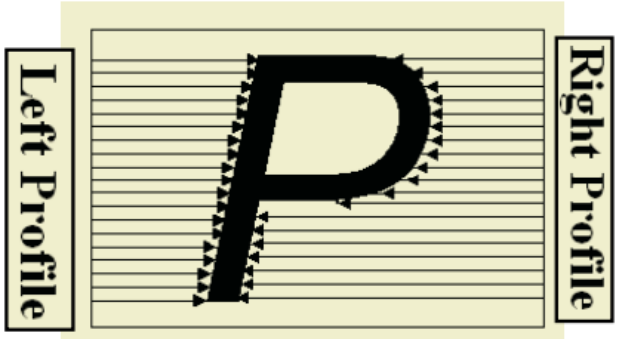


Fig. 3(c). Vertical line detection approach for italics character.

Joshiet. al. [28] proposed three different schemes to recognise Indian scripts. The first method is based on a representation of the horizontal profile of the textual blocks in the frequency domain. The other two approaches make use of linked elements that were taken directly from the textual region. A connected component is only handled if its height is higher than or equal to three-fourths or less than one-fourth of the typical character height in the document picture. Each linked component is represented in the training data by 12 Gabor feature values and a script label. Four important Indian language scripts-Devanagari, Roman (English), Telugu, and Malayalam—have been classified using this approach.

III. Comparative analysis

Some of the work in script recognition is summarised in Table 1. This table lists the various script characteristics that were employed by various scholars. However, only a few scripts were used in their studies to achieve the findings they reported. This raises the question of how these script capabilities will function in scripts that are not included in respective works. Investigating the suggested script identification feature is crucial.

Table 1 : Script recognition summary

Ref. No.	Author (s)	Languages	Features used	Page/ Word / Line /character	Classifier Used
[04]	Chaudha ri et. al.	Bangla - Devnagari	Vertical lie, boundary box width, presence of signature in the upper zone	Line, word, charact er	tree classifier, run length template matching
[05]	Dhanya et. al.	English - Tamil	Spatial spread features, Directional features	Word	SVM, NN and KNN

[06]	Dhandra et. al.	Kannada - Devnagari and English numeral	Eccentricity, aspect ratio, and horizontal and vertical stroke	Word	KNN
[24]	Sharma et. al.	English, Bengali and Hindi	Zernike moments, Gabor and gradient	Word	SVM
[25]	Phan et. al.	English-Chinese, English-Tamil, Chinese-Tamil-English, Chinese-Tamil.	Smoothness and cursiveness of the upper and lower lines, PCA	Line	KNN
[31]	Mohanty et. al.	English - Oriya	Analyze Text Segment, used horizontal histogram, Baird Algorithm used for skew correction	All	SVM
[32]	Dhandra et. al.	Kannada, Tamil and Devnagari with English Numerals	Discriminating features	Word	KNN
[33]	Obaidullah et. al.	Bangla, Devnagari, Roman, Oriya, Urdu, Gujarati, Telegu, Kannada, Malayalam and Kashmiri.	(i) Structural Feature, (ii) Mathematical Feature and (iii) Morphological Feature		Logistic classifier with 5 fold cross validation
[35]	Roy et. al.	Bangla, Devnagari, Malayalam, Urdu, Oriya and Roman	On circularity, fractal dimension, component analysis, small component, Freeman chain code	Page	MLP
[36]	Padma and Vijaya	English, Telugu, Malayalam, Kannada, Tamil, Urdu, and Hindi	Gray level co-occurrence matrix, Harr wavelet	Page	KNN
[37]	Padma and Vijaya	Bangla, Devanagari, Roman (English), Gujarati, Malayalam, Oriya, Tamil, Telugu, Kannada and Urdu	Gray level co-occurrence matrix, Harr wavelet	Page	KNN
[38]	Hassan et. al.	Hindi-English, Bangla-English	Texture and Shape Based	Word	SVM & AdaBoost
[38]	Singhal et. al.	English, Hindi, Bangla and Telugu	Visual appearance of the text image	Page	A quantitative measure for dissimilarity
[39]	Hiremath et. al.	English, Kannada, Tamil, Urdu, Telugu, Bengali, Hindi, and Malayalam	Texture feature, DWT	Page	KNN
[40]	Joshi et. al.	Bangla, Devanagari, Roman(English), Gurumukhi, Gujarati, Malayalam, Oriya, Tamil, Kannada and Urdu	Log-Gabor Filtering, Oriented local energy responses, Oriented local energy responses, Horizontal profile:,	Line	Quadratic Bayes normal classifies, Neural network based classifier, K-Nearest Neighbor Classifier, SVM, Parzen density based classifier
[41]	Rajput and Anita	eight script	Gabor filters, DCT and Daubechies Wavelets	Line	KNN

[42]	Patil and Ramakrishnan	Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Odiya, Punjabi, Tamil, Telugu and Urdu,	Gabor and Discrete cosine transform (DCT)	Word	Nearest neighbor and SVM
[44]	Sinha et. al.	Devnagari-English, Devnagari-English, Telugu-English, Bangla-English, Malayalam-English, and Gujarati-English.	Shirorekha, feature, Distribution of vertical stroke feature, Water reservoir principle based feature, Shift Feature Below Shirorekha, Left and right profile, Deviation feature	Word	Tree classifier
[46]	Rani et. al.	English - Gurumukhi	Structural, Gabor and Discrete Cosine Transforms(DCT)	Word	SVM, KNN, Parzen Probabilistic Neural Network (PNN)
[47]	Singh et. al.	Devnagari - Gurumukhi	Use the density of extracted vertical and horizontal strokes		
[48]	Sreejith et. al.	Hindi - Sanskrit	N-gram profile	-	Similarity Measure count
[49]	Sarkar et. al.	Bangla-Roman, Devanagari-roman	Horizontalness, segmentation-based, foreground background transition	Word	Multilayer perceptron, back propagation algorithm
[50]	Patil and Ramakrishnan	Kannada , Devanagari and Roman	Gabor and Discrete cosine transform (DCT)	Word	Nearest neighbor and SVM
[51]	Hangare and Dhandra	Devanagari, English, Urdu	Vertical and Horizontal Stroke Density, Right and left diagonal stroke density	Word	KNN
[52]	Chaudhari and Gulati	English-Gujarati	Statistical approach	Digit	KNN
[53]	Chaudhari and Gulati	English-Gujarati	Horizontal projection, spatial spread of pixels	Line-wise	KNN
[54]	Bhunia et. al.	Indic script	handcrafted feature	Word	multi-modal deep network
[55]	Ghosh et. al.	Devanagari and Bengali	horizontal zoning	Word	LSTM and BLSTM RNN

Applications

Multilingual character recognition for Indian scripts has various practical applications across different domains. Here are some notable applications:

(i) Identity Verification: It can be employed in identity verification systems that rely on documents containing Indian script characters. It helps authenticate individuals by accurately recognizing characters in documents such as passports, identification cards, and driving licenses.

(ii) OCR: OCR technology for multilingual character recognition enables the automatic extraction of text from images or scanned documents. It is used in applications such as text recognition in images, automated data entry, and text-to-speech conversion.

(iii) Postal and Address Recognition: Postal services depend on the efficient and precise recognition of addresses and postal codes in Indian scripts. The

automatic sorting, routing, and delivery of mail is made possible with the use of multilingual character recognition, which improves postal operations.

(iv) Handwritten Text Recognition: It helps in transcribing and understanding handwritten text in Indian scripts. It can be used to digitise handwritten notes, forms, and old documents. It can also help people with disabilities access written content.

(v) Data mining and analysis: Multilingual character recognition makes it possible to extract text from a variety of sources, including user-generated content in Indian scripts, internet articles, and social media posts. This helps with analysis and insights particular to languages.

(vi) Document Digitization: Multi-script script identification will be valuable for digitizing and preserving historical documents written in various Indian scripts.

This field is continually evolving, and new applications are emerging as researchers and developers explore its potential in various domains.

IV. CONCLUSION

This research has focused on the development of a robust multilingual character recognition system specifically designed for Indian scripts.

The research investigates different approaches for feature extraction, to capture the distinguishing characteristics of characters in each script.

By providing a comprehensive overview of the current research landscape, methodologies, challenges, and future prospects, this survey serves as a valuable resource for researchers, practitioners, and professionals interested in multilingual character recognition for Indian scripts. It highlights the need for further research and development to overcome the challenges and improve the accuracy and efficiency of multilingual character recognition systems in the Indian context.

Looking ahead, further advancements can be made in multilingual character recognition for Indian scripts. Future research should explore additional Indian scripts, expanding the scope of the system to cover a more comprehensive range of languages.

This research paves the way for continued progress in multilingual character recognition for Indian scripts, promoting advancements in language technology and facilitating efficient communication and understanding across diverse linguistic communities.

V. REFERENCES

[1]. B.B. Chaudhuri, U. Pal, A complete printed Bangla OCR system, Pattern Recognit. 31 (1998) 531–549.

[2]. Chanda, Sukalpa, Srikanta Pal, and Umapada Pal. "Word-wise sinhala tamil and english script identification using gaussian kernel svm." In 2008 19th International Conference on Pattern Recognition, pp. 1-4. IEEE, 2008.

[3]. Chanda, Sukalpa, Srikanta Pal, Katrin Franke, and Umapada Pal. "Two-stage approach for word-wise script identification." In 2009 10th International Conference on Document Analysis and Recognition, pp. 926-930. IEEE, 2009.

[4]. Chaudhuri, B. B., and Umapada Pal. "An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)." In Proceedings of the fourth international conference on document analysis and recognition, vol. 2, pp. 1011-1015. IEEE, 1997.

[5]. Dhanya, D., and A. G. Ramakrishnan. "Script identification in printed bilingual documents." In Document Analysis Systems V: 5th International Workshop, DAS 2002 Princeton, NJ, USA, August 19–21, 2002 Proceedings 5, pp. 13-24. Springer Berlin Heidelberg, 2002.

[6]. Dhandra, B. V., H. Mallikarjun, Ravindra Hegadi, and V. S. Malemath. "Word-wise script identification from bilingual documents based on morphological reconstruction." In 2006 1st International Conference on Digital Information Management, pp. 389-394. IEEE, 2006.

[7]. Ghosh, Shamita, and Bidyut B. Chaudhuri. "Composite script identification and orientation detection for indian text images." In 2011 International Conference on Document Analysis and Recognition, pp. 294-298. IEEE, 2011.

[8]. Jaeger, Stefan, Huanfeng Ma, and David Doermann. "Identifying script on word-level with informational confidence." In Eighth international conference on document analysis and recognition (ICDAR'05), pp. 416-420. IEEE, 2005.

[9]. Padma, M. C., and P. A. Vijaya. "Monothetic separation of Telugu, Hindi and English text lines from a multi script document." In 2009 IEEE International Conference on Systems, Man and Cybernetics, pp. 4870-4875. IEEE, 2009.

[10]. Pal, U., and B. B. Chaudhuri. "Script line separation from Indian multi-script documents." IETE Journal of Research 49, no. 1 (2003): 3-11.

[11]. Pal, U., and B. B. Chaudhuri. "Automatic identification of english, chinese, arabic, devnagari and bangla script line." In Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 790-794. IEEE, 2001.

[12]. Pal, Umapada, Suranjit Sinha, and B. B. Chaudhuri. "Multi-script line identification from Indian documents." In Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., vol. 3, pp. 880-880. IEEE Computer Society, 2003.

[13]. Aithal, Prakash K., G. Rajesh, Dinesh U. Acharya, and NV Krishnamoorthi M. Subbareddy. "Text line script identification for a tri-lingual document." In 2010 Second International conference on Computing, Communication and Networking Technologies, pp. 1-3. IEEE, 2010.

[14]. Rani, Rajneesh, Renu Dhir, and Gurpreet Singh Lehal. "Script identification of pre-segmented multi-font characters and digits." In 2013 12th international conference on document analysis and recognition, pp. 1150-1154. IEEE, 2013.

[15]. Dhandra, B. V., and Mallikarjun Hangarge. "Global and local features based handwritten text words and numerals script identification." In International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), vol. 2, pp. 471-475. IEEE, 2007.

[16]. Hiremath, P. S., Jagdeesh D. Pujari, S. Shivashankar, and V. Mouneswara. "Script identification in a handwritten document image using texture features." In 2010 IEEE 2nd International Advance Computing Conference (IACC), pp. 110-114. IEEE, 2010.

[17]. Namboodiri, Anoop M., and Anil K. Jain. "Online script recognition." In 2002 International Conference on Pattern Recognition, vol. 3, pp. 736-739. IEEE, 2002.

[18]. Obaidullah, Sk Md, Kaushik Roy, and Nibaran Das. "Comparison of different classifiers for script identification from handwritten document." In 2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC), pp. 1-6. IEEE, 2013.

[19]. Pal, Umapada, Nabin Sharma, Tetsushi Wakabayashi, and Fumitaka Kimura. "Handwritten numeral recognition of six popular Indian scripts." In Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, pp. 749-753. IEEE, 2007.

[20]. Roy, K., U. Pal, and B. B. Chaudhuri. "Neural network based word-wise handwritten script identification system for Indian postal automation." In Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005., pp. 240-245. IEEE, 2005.

[21]. Roy, Kaushik, and Kinshuk Majumder. "Trilingual script separation of handwritten postal document." In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 693-700. IEEE, 2008.

[22]. Roy, K., S. Kundu Das, and Sk Md Obaidullah. "Script identification from handwritten document." In 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 66-69. IEEE, 2011.

[23]. Zhou, Lijun, Yue Lu, and Chew Lim Tan. "Bangla/English script identification based on analysis of connected component profiles." In Document Analysis Systems VII: 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings 7, pp. 243-254. Springer Berlin Heidelberg, 2006.

[24]. Sharma, Nabin, Sukalpa Chanda, Umapada Pal, and Michael Blumenstein. "Word-wise script identification from video frames." In 2013 12th International Conference on Document Analysis and Recognition, pp. 867-871. IEEE, 2013.

[25]. Phan, Trung Quy, Palaiahnakote Shivakumara, Zhang Ding, Shijian Lu, and Chew Lim Tan. "Video script identification based on text lines." In 2011 international conference on document analysis and recognition, pp. 1240-1244. IEEE, 2011.

[26]. Spitz, A. Lawrence. "Determination of the script and language content of document images." IEEE Transactions on Pattern Analysis and Machine Intelligence 19, no. 3 (1997): 235-245.

[27]. Pal, Umapada, and B. B. Chaudhuri. "Indian script character recognition: a survey." pattern Recognition 37, no. 9 (2004): 1887-1899.

[28]. S. Chaudhury and R. Sheth., Trainable script identification strategies for Indian languages. Fifth International Conference on Document Analysis and Recognition (1999) 657–660.

[29]. Joshi, Gopal Datt, Saurabh Garg, and Jayanthi Sivaswamy. "Script identification from Indian documents." In Document Analysis Systems, vol. 7, pp. 255-267. 2006.

[30]. Pati, Peeta Basa, S. Sabari Raju, Nishikanta Pati, and A. G. Ramakrishnan. "Gabor filters for document analysis in Indian bilingual documents." In International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of, pp. 123-126. IEEE, 2004.

[31]. Mohanty, Sanghamitra, and HN Das Bebartta. "A novel approach for bilingual (english-oriya) script identification and recognition in a printed document." International Journal of Image Processing (IJIP) 4, no. 2 (2010): 175.

[32]. Dhandra, B. V., Mallikarjun Hangarge, Ravindra Hegadi, and V. S. Malemath. "Word level script identification in bilingual documents through discriminating features." In 2007 International Conference on Signal Processing, Communications and Networking, pp. 630-635. IEEE, 2007.

[33]. Obaidullah, Sk Md, Anamika Mondal, and Kaushik Roy. "Structural feature based approach for script identification from printed Indian document." In 2014 International Conference on Signal Processing and Integrated Networks (SPIN), pp. 120-124. IEEE, 2014.

[34]. Dhandra, B. V., P. Nagabhushan, Mallikarjun Hangarge, Ravindra Hegadi, and V. S. Malemath. "Script identification based on morphological reconstruction in document images." In 18th International Conference on Pattern Recognition (ICPR'06), vol. 2, pp. 950-953. IEEE, 2006.

[35]. Roy, K., S. Kundu Das, and Sk Md Obaidullah. "Script identification from handwritten document." In 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 66-69. IEEE, 2011.

[36]. Padma, M. C., and P. A. Vijaya. "Global approach for script identification using wavelet packet based features." International Journal of Signal Processing, Image Processing and Pattern Recognition 3, no. 3 (2010): 29-40.

[37]. Padma, M. C., and P. A. Vijaya. "Wavelet packet based texture features for automatic script identification." Int. J. Image Process 4, no. 1 (2010): 53-65.

[38]. Singhal, Vivek, Nishant Navin, and Debashish Ghosh. "Script-based classification of hand-written text documents in a multilingual environment." In Proceedings. Seventeenth Workshop on Parallel and Distributed Simulation, pp. 47-54. IEEE, 2003.

[39]. Hiremath, P. S., Jagdeesh D. Pujari, S. Shivashankar, and V. Mouneswara. "Script identification in a handwritten document image using texture features." In 2010 IEEE 2nd

International Advance Computing Conference (IACC), pp. 110-114. IEEE, 2010.

[40]. Joshi, Gopal Datt, Saurabh Garg, and Jayanthi Sivaswamy. "Script identification from Indian documents." In Document Analysis Systems, vol. 7, pp. 255-267. 2006.

[41]. Rajput, Ganapatsingh G., and H. B. Anita. "Handwritten script recognition using DCT, gabor filter and wavelet features at line level." Soft Computing Techniques in Vision Science (2012): 33-43.

[42]. Pati, Peeta Basa, and A. G. Ramakrishnan. "Word level multi-script identification." Pattern Recognition Letters 29, no. 9 (2008): 1218-1229.

[43]. Pal, Umapada, and B. B. Chaudhuri. "Automatic separation of words in multi-lingual multi-script Indian documents." In Proceedings of the fourth international conference on document analysis and recognition, vol. 2, pp. 576-579. IEEE, 1997.

[44]. Sinha, Suranjit, Umapada Pal, and B. B. Chaudhuri. "Word–wise script identification from indian documents." In Document Analysis Systems VI: 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004. Proceedings 6, pp. 310-321. Springer Berlin Heidelberg, 2004.

[45]. Hassan, Ehtesham, Ritu Garg, Santanu Chaudhury, and M. Gopal. "Script based text identification: a multi-level architecture." In Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pp. 1-8. 2011.

[46]. R. Rani, R. Dhir, and G. S. Lehal, "Performance Analysis of Feature Extractors and Classifiers for Script Recognition of English and Gurmukhi Words," Proceeding of the ACM workshop on Document Analysis and Recognition, pp. 30-36, 2012.

[47]. Singh, Sukhvir, Anil Kumar, Dinesh Kr Shaw, and D. Ghosh. "Script separation in machine printed bilingual (Devnagari and Gurumukhi) documents using morphological approach." In 2014 Twentieth National Conference on Communications (NCC), pp. 1-5. IEEE, 2014.

[48]. Sreejith, C., M. Indu, and PC Reghu Raj. "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts." In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-4. IEEE, 2013.

[49]. Sarkar, Ram, Nibaran Das, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, and Dipak Kumar Basu. "Word level script identification from bangla and devanagri handwritten texts mixed with roman script." arXiv preprint arXiv:1002.4007 (2010).

[50]. Pati, Peeta Basa, and A. G. Ramakrishnan. "Word level multi-script identification." Pattern Recognition Letters 29, no. 9 (2008): 1218-1229.

[51]. Hangarge, Mallikarjun, and B. V. Dhandra. "Offline handwritten script identification in document images." International Journal of Computer Applications 4, no. 6 (2010): 6-10.

[52]. Chaudhari, Shailesh A., and Ravi M. Gulati. "An OCR for separation and identification of mixed English—Gujarati digits using kNN classifier." In 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), pp. 190-193. IEEE, 2013.

[53]. Chaudhari, Shailesh A., and Ravi M. Gulati. "Script Identification from Bilingual Gujarati-English Documents." International Journal of Computer Applications 93, no. 17 (2014).

[54]. Bhunia, A.K., Mukherjee, S., Sain, A., Bhunia, A.K., Roy, P.P., Pal, U.: Indic handwritten script identification using offline-online multi-modal deep network. Inf. Fusion 57, 1–14 (2020)

[55]. Ghosh, R., Vamshi, C., Kumar, P.: Rnn based online handwritten word recognition in Devanagari and Bengali scripts using horizontal zoning. Pattern Recogn. 92, 203–218 (2019)

Cite this article as :

Jagin M. Patel, Bharat C. Patel, "A Comprehensive Survey of Script Identification in Indian Multi-Script Documents", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.535-546, March-April-2022.