International Journal on Document Analysis and Recognition (IJDAR) (2021) 24:97–121
https://doi.org/10.1007/s10032-021-00365-5

ORIGINAL PAPER

Offline script recognition from handwritten and printed multilingual documents: a survey

Deepak Sinwar[1] · Vijaypal Singh Dhaka[1] · Nitesh Pradhan[2] · Saumya Pandey[1]

Received: 3 April 2020 / Revised: 8 March 2021 / Accepted: 9 March 2021 / Published online: 22 March 2021

Abstract
Script recognition has many real-life applications like optical character recognition, document archiving, writer identification, searching within the documents, etc. Automatic script recognition from multilingual documents is a stimulating task, where the system must identify and recognize several types of scripts that can be available on a single page. In offline script recognition, printed or handwritten documents are firstly scanned followed by the process of script recognition, whereas in online script recognition documents are already in soft-copy form. Most of the script recognition techniques presented by researchers so far are based on traditional image processing frameworks. But nowadays, it is observed that Deep Learning-based techniques are more capable of achieving a script recognition task efficiently as well as accurately. This paper provides a comprehensive survey of various techniques available for identification and recognition of multilingual scripts from the last few decades that are mainly focused on Indic scripts. However, some potential non-Indic script identification works are also incorporated for ease of understanding. We hope that this survey can act as a compendium as well as provide future directions to researchers for developing generic OCRs.

Keywords Indic script identification · Script recognition · Support vector machine · Artificial neural network · Multi-layer perceptron · Nearest neighbor · Multilingual · Handwritten · k-NN

1 Introduction

The script is used to define the writing system with the help of several graphics forms. Generally, it refers to a pattern of writing based on a certain character set. Day by day, information technology is increasing its role in our daily life and our trust over the digitization of routine services in all fields of our life is rapidly increasing. In the direction of paperless solutions of historical document archives, the development of Optical Character Recognizer (OCR) needs to be faster as well as optimized. OCR development has been the oldest area of research and investigation, but to date, we could not develop a multilingual, generic, and robust system that can recognize several scripts written on a page accurately. However, plenty of OCRs proved to be efficient in script recognition tasks, but most of them are limited to some specific scripts only. Nowadays, the most challenging task is to recognize scripts from multilingual documents in which contents are generally written in more than one script/ language. Hence, the problem of developing the generic script recognizer is one of the hardest problems of the research domain. Multi-script recognition majorly involves the identification of individual scripts followed by script recognition task based on several features. Manual script identification seems insignificant; that is why the demand for automatic script recognizer is increasing day by day. On the other hand, multilingual script identification is foremost required for the development of generic OCRs to overcome the availability of limited language OCRs. For the last two decades, researchers are working continuously on automatic script identification [25]. Though it has been

✉ Deepak Sinwar
deepak.sinwar@gmail.com

Vijaypal Singh Dhaka
prof.dhaka@gmail.com

Nitesh Pradhan
nitesh.pradhan@jaipur.manipal.edu

Saumya Pandey
saumya.april1@gmail.com

[1] Department of Computer and Communication Engineering, Manipal University Jaipur, Dehmi Kalan Jaipur, Rajasthan 303007, India

[2] Department of Computer Science and Engineering, Manipal University Jaipur, Dehmi Kalan Jaipur, Rajasthan 303007, India

observed that few of the research was focused on the identification of isolated characters rather than the identification of the whole script. In this paper, we are trying to present a constructive analysis of several works published for multilingual script identification which were majorly focused on Indic scripts. However, few non-Indic script identification approaches are also discussed for providing an in-depth analysis of script identification task. As compared to roman scripts, the Indic script is constituted by two-dimensional constituent symbols; that is why the process of Indic script identification is relatively complex as compared to roman. In general, the process of recognizing handwritten as well as printed scripts involves a predefined process, viz. preprocessing, segmentation, feature extraction, and recognition [25] as shown in Fig. 2. Preprocessing is used to improve the quality of the image by removing noise elements if any. Plenty of preprocessing techniques (noise removal, skew correction, binarization, feature extraction, header line removal, etc.) exist in the research literature, but the choice of preprocessing method depends on the application for which the image is used. In script identification, the aim of preprocessing is to remove the document image noise that is generally introduced during scanning, text acquisition, transmission, etc. A common form of distortion called skew often gets incorporated while copying or scanning a document. On the other hand, binarization helps in converting an image into a two-level (black and white) images so that the computation cost of processing becomes cheaper. Feature extraction on the other hand is one of the basic components of the script recognition system that extracts useful patterns from text images. Sometimes, it is also used to reduce the total number of features required for accomplishing the script recognition task by preparing a minimum subset of features called feature set.

From the past decade, interest in research has turned toward the Indian multilingual script identification. India is a country with 22 scheduled languages and 99 non-scheduled languages. As per the number of scripts that are concerned, India is a country with 13 scripts that can be roughly classified into two main classes, viz. scripts with 'matra' and scripts without 'matra' [89]. Matra is a structural property of the script with a horizontal line on the upper part of the characters. Nowadays, it becomes the hottest challenge for researchers to recognize several Indian scripts written on a page and translate them for the required purposes, i.e., digital transformation, recognition, storage, etc. Day by day, research in this field, especially automatic script identification is increasing, and researchers are competing at national as well as international levels. Automatic identification of the script is classified into two categories, viz. local and global approaches [46]. The local approach extracts the features from connected components such as characters, words, and lines after segmenting the document. The main drawback of this approach is that there is no common segmentation process that exists that can incorporate all scripts. On the other hand, the global approach analyzes the regions containing more than one line. So, at least two text lines need to be analyzed for the global segmentation approach. The work presented in this paper is a narrative survey of various script identification techniques which were mainly developed for Indic script identification. Generally, script identification can be classified according to their structures, connections, and writing styles. Out of these, one may identify the connected components and then based on their shapes and structures the remaining analysis can be made. Such kind of connected components can be found easily in Indic scripts. In general, script recognition is divided into two classes as follows:

1. Online script recognition
2. Offline script recognition

In online script recognition, the process of recognizing a script is usually done on various online systems (without keyboard), i.e., Personal Digital Assistant (PDA), touch screen panels, graphics tablet, and other handheld devices [85], such as mobile phones. This process requires very efficient algorithms to accomplish the task. First, the algorithm needs to know about the type of script being written in the document, because sometimes documents are written in multiple scripts. The process becomes very complex when a single document contains the text of more than one language/ script. After analysis of a script, the task of recognizing words and characters begins, whereas, in the case of offline script recognition, the process of script recognition takes place after scanning printed/ handwritten documents. The process of identifying script from printed documents is generally referred to as Printed Script Identification (PSI), and the process of identifying scripts from handwritten documents is referred to as Handwritten Script Identification (HSI). After scanning, the documents are converted into digital representation in the form of pixels for extracting useful features. This task is a little bit complex as compared to online script recognition because the segmentation of background from the foreground is easy in the case of online recognition.

In offline recognition, the efficiency of segmentation mainly depends on the quality of the documents. If a scanned document is of good quality, then the process becomes faster; otherwise, it needs lots of pre-processing computations to discriminate between foreground and background. Nowadays, most of the work in script recognition is focusing on offline script recognition instead of the online one. After identifying the corresponding recognition system, the process can be subdivided into different phases as shown in Fig. 1.

As per the characters are concerned, they may be classified as either cursive or non-cursive. Both cursive and non-cursive letters can be available in either of two types, viz. handwritten and printed documents. In the case of online script recog-

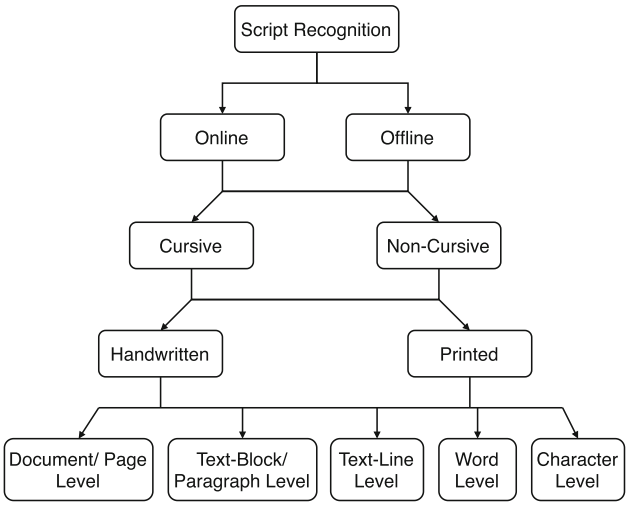

Fig. 1 Classification of online as well as offline script recognition

nition, the documents are generally handwritten, whereas offline script recognition deals with both types. The script recognition can be carried out and further broken down into different levels of granularity as follows:

1. Document/ page level
2. Text line level
3. Text block/ paragraph level
4. Word level
5. Character level

Some researchers have classified the script recognition task at different levels of granularity. Keserwani *et al.* [64] classified the scripts recognition process into two broad categories, viz. holistic word-level approach and patch-level approach. In the holistic word-level approach, the word is considered as a single indivisible entity, whereas in the patch-level approach image patches are considered for script recognition.

The present survey shows the comprehensive analytical study of some potential multilingual script identification tools and techniques mainly of offline script recognition so far. In short, the key contributions of this paper are as follows:

- In-depth review of several research articles that are mainly focused on Indic script identification on different granularity levels, classifiers, and accuracy rates.
- Comparison with other script identification-based state-of-the-art survey articles.
- Presented several public data sets for Indic script identification.
- Highlighted several open research challenges and future directions for script identification.

Fig. 2 General script recognition process [164]

The subsequent sections of this paper are organized as follows. The general script recognition process is briefed in section 2, followed by an introduction to different writing systems in section 3. Various traditional script identification methods are surveyed in section 4 at different granularity levels along with discussion on each level. Section 5 presents some deep learning-based script identification approaches. Few public data sets for Indic script identification are presented in section 6. Section 7 presents the summary of some script identification-based survey articles. Section 8 is about discussion, challenges, and future directions for Indic script identification. Finally, section 9 concludes the survey.

2 Script Recognition Process

Generally, script recognition can be accomplished by traditional image processing techniques that follow four major steps, viz. Preprocessing, Segmentation, Feature Extraction, and Classification as illustrated in Fig. 2. In addition to this, several Machine Learning and Deep Learning techniques may be utilized for achieving the overall task. Sometimes the process of script identification is blended with language identification. Bashir *et al.* [9] presented the task of language identification using text categorization that can be solved using computational as well as statistical approaches, whereas in script identification, the major task is to first identify the text region from an image or video followed by classification of scripts. This section presents the brief about the general script recognition process into two main classes, viz. traditional image processing and modern deep learning techniques.

2.1 Traditional image processing and machine learning techniques

In this technique, scanned images are firstly preprocessed for removal of noise (if any) to refine their quality followed by segmentation, feature extraction, and finally classification. Segmentation is the process of dividing the larger image into several smaller segments for easy as well as efficient computations. Numerous techniques have been proposed by researchers to do the segmentation, e.g., horizontal segmentation, vertical segmentation, threshold-based segmentation, edge detection, pixel-wise Support Vector Machine (SVM) [172], region growing, etc. On the other hand, feature extraction is needed to extract the relevant and required features instead of processing the whole image at one go. Various

techniques have been proposed by researchers for feature extractions, i.e., horizontal and vertical histogram, topological features, curvature features [28], component-based features [142], wavelet-based features, etc. A review of various feature extraction techniques for script recognition is provided in subsequent sections along with different script recognition techniques. The last phase in script recognition is the classification of extracted features. Generally, the process of classification is done with the help of machine learning techniques. Like feature extraction, numerous classifiers have been devised to do the task. Some of the famous classifiers are Support Vector Machines, K-Nearest Neighbor (k-NN), Multilayer Perceptron (MLP), Feed Forward Neural Networks, etc. These four steps of the process are summarized in the following subsections as follows:

2.1.1 Preprocessing

The script recognition process starts with the preprocessing of images. There may be chances that the captured images may contain some noise, so it becomes vital to remove the noise from these captured images. Also preprocessing is used to enhance the image quality by applying numerous filters/ transformations. There may be various reasons for the noise in the image. Noise may come during image acquisition/ transmission; for example, skew a kind of noise that may occur in the image during acquisition or transferring [164]. There are lots of preprocessing strategies in the research literature, viz. binarization, skew/ slant correction, scaling, etc. The choice of a strategy depends on the application area. Some of the pre-processing techniques are explained as follows:

1. **Binarization** The process of converting a 256-level image into two levels is called binarization. To binarize an image, first of all we have to set a threshold limit; if pixel values of the image are less than the threshold specified, the value of that particular pixel will be set to 1; otherwise, it will be set to 0 [164]. We may also perform script recognition without performing the binarization process also. Shi *et al.* [158] developed a novel method using Convolutional Neural Network (CNN) to perform the process of script recognition without including the binarization process. But in traditional image processing methodologies, without binarization, the task of extracting features becomes relatively complex.
2. **Normalization** To overcome the effect of noise, we may normalize the intensity values of the pixels to a specified range. This process is carried out by setting the intensity value to the average values of the surrounding pixels [132].
3. **Skeleton** This technique is used to enhance the quality of images that are suffering from various issues, i.e., blurring, low-resolution, complex backgrounds, etc. [154]. With the help of this technique, one can easily extract the structural features of the image that plays a vital role in the script recognition. The process of obtaining the skeleton of the image begins with the binarization process followed by the clustering of foregrounds and backgrounds [154].
4. **Morphological Operations** Binary images sometimes suffer from many imperfections such as distortion. Morphological operations are generally applied to remove those imperfections from the images [2]. Basic morphological operations include erosion, dilation, and compound operations.

2.1.2 Segmentation

The next step after performing pre-processing on the text images is to divide the image into smaller portions called segments. An efficient segmentation technique results in improved accuracy of OCRs. Numerous techniques have been proposed in the research literature on segmentation; some of them are summarized as follows:

1. Line and word segmentation: While processing the whole page/ document, firstly the text needs to be divided into text blocks that are further subdivided into lines, words, and characters. The idea behind the subdivision is to find the valleys by counting the black pixels in a row [18]. The process continues in both horizontal and vertical directions for finding out text lines and words, respectively.
2. Character Segmentation: After finalizing the text line and word boundaries, the process of finalizing character boundaries begins. Chanda *et al.* [18] have used the headline feature for this purpose. Generally, this feature is not available in English like languages, but we may find it in some other scripts such as Devanagari and Bangla. The concept behind finding the character is the presence of a headline, if we are unable to locate the headline it means the corresponding character is isolated. Sharma *et al.* [157] proposed a new method for segmenting text from video frames using four gradient-based features.

2.1.3 Feature extraction

Features are nothing but the trait inputs to the classifiers. In script recognition, the feature extraction process provides the required and relevant features to the classifiers after applying various techniques. Features ought to be easy to compute as well as needs to be robust [88]. Obaidullah *et al.* [88] categorized features into three main categories as follows:

1. Mathematical / Abstract features
2. Structural Features
3. Script dependent features

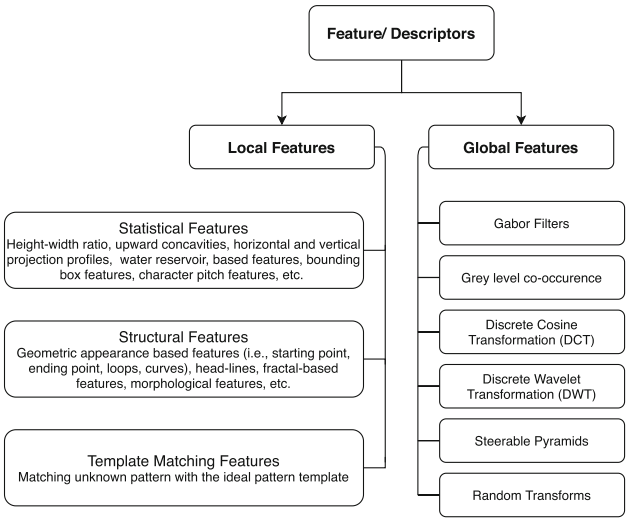

Fig. 3 Classification of different features for script recognition process [170]

Some famous features used by feature extraction techniques are: Gabor features, component-based features, Discrete Wavelet Transform (DWT)-based features, structural-/shape-based features, principle stroke features, Haar features, horizontal and vertical projection-based features, histogram-based features, spatial features, density distributions, water reservoir principle-based feature, fractal-based features, Curvature Scale Space (CSS) features, the position of concavities, optical densities, the center of gravity features and many more are summarized in the subsequent sections. Ubul *et al.* [170] classified feature extraction into two main classes, viz. local features and global features as shown in Fig. 3. Local features are based on the shape and structure of a character, whereas global features can be steerable pyramids, Gabor, transformations, etc. In general, global feature extraction is needed for object detection, whereas local feature extraction is required for object recognition. Sometimes features and descriptors are used interchangeably in image identification and recognition.

2.1.4 Classification

Classification deals with the process of classifying inputs into different classes based on various features. It is sometimes called the heart of the pattern recognition process. The efficiency and accuracy of a general script recognition process depend on both features and classifiers. Numerous classifiers have been proposed in the research literature about classification problem such as Support Vector Machine, Multilayer Perceptron, k-Nearest Neighbor, Naïve Bayes, rule-based classifiers, tree-based classifiers (i.e., decision trees, Random Forest (RF), etc.), cross-validation, hierarchical classifiers, Artificial Neural Network (ANN), etc. In this paper, authors have reviewed and compared the accuracy of these classifiers on different types of data sets in the subsequent sections.

2.2 Deep learning

Due to various advancements in Deep Learning (DL) techniques, feature engineering has witnessed a paradigm shift [66]. In this paradigm shift, feature extraction is generally carried out during the training phase. Plenty of deep learning-based techniques exists in the research literature, but CNNs proved to be appropriate for feature extraction for specific tasks. Unlike traditional feature extraction process which involves lots of human interventions, deep learning automatically extracts useful features and speeds up the entire process [169]. Also it is more advanced and has tremendous learning capabilities. It enables systems to learn from observations and train the proposed model automatically. However, some human intervention is also involved in fine-tuning to achieve accurate identification and recognition. Many DL models have been invented by different researchers, i.e., Convolutional Neural Network, Auto Encoders, Restricted Boltzmann Machine (RBM), Deep Belief Networks (DBN), etc. From the last decade, CNN proves to be one of the most popular tools for image recognition. CNN has the facility to automatically extract complex features from images that are further used to train the system for recognition. The general architecture of CNN consists of five basic layers:

1. Input layer: used to input the data in the form of images.
2. Convolutional layers: used for generating feature maps by applying a series of filters. These filters are useful for pattern recognition, edges detection, color changes, etc.
3. Non-linear functions: most real-world situations need CNN to work upon non-linearity. To achieve this non-linearity in CNN, the Rectified Linear Unit (ReLU) function is most widely used. It is a component-wise operation that is applied on every pixel of an image to replace the negative pixels in the feature map to zero. Some other non-linear functions, i.e., *Sigmoid* and *Tanh* can also be used for achieving non-linearity, but ReLU is most widely used for this purpose because of its better performance.
4. Pooling layers: used to reduce the dimensions of feature maps without removing the important information from the data. Pooling can be the max, min, average, sum, etc.
5. Fully connected layer: It is a multi-layer perceptron neural network, that uses *Softmax* as an activation function. This layer is called fully connected because all the neurons of the previous layer are connected to every neuron of the next layer.

3 Writing systems

As of today, throughout the world, six script writing systems are in use [40]. Therefore, all systems have one or more scripts and can be used with one or more languages.

1. **Logographic system** In a logographic system, full words are represented using characters or symbols, for example, Korean, Japanese, and Chinese scripts. These types of scripts have multiple short strokes as well as appearance-based visual features that clearly distinguish them from other Asian and Western scripts.
2. **Syllabic system** In a syllabic writing system each symbol represents a syllable or phonetic sound. Japanese writing systems are generally based on a mix of both syllabic and logographic systems. The symbols used to represent syllables are referred to as *Kanas*.
3. **Alphabetic System** In alphabetic systems, a set of some characters generally referred to as alphabets are used to make phonemes of a language. Armenian, Greek, Cyrillic, and Latin are the most prominent alphabetic systems. The alphabetic system is originated in Greek and then spread over the globe. Many languages are based on alphabetic systems, for example, Italian, English, Spanish, German, Portuguese, etc. Alphabetic systems have been extended to some other writing systems, i.e., Cyrillic writing systems that are quite similar to Latin but with a different character set [40]. Some Asian and Eastern European writing systems had adopted the Cyrillic writing system, i.e., Russian, Macedonian, Ukrainian, and Bulgarian.
4. **Abjads** Abjads consist of symbols for consonantal sounds that follow the right to left writing pattern along with a text line. Hebrew, Arabic, Urdu, Farsi, Sindhi are some famous examples of the Abjads writing system. The words in Abjads are cursive long strokes consisting of few dots. The process of script identification for Abjads is relatively simpler as compared to other scripts because it is easier to recognize the long cursive strokes with dots.
5. **Abugidas** Abugidas is originated from the ancient Brahmic writing system and all the south-east Asian and Indian writing systems are originated from there only. It is like alphabetic systems in which words are written by combining characters with the help of a straight horizontal line called Shirorekha. Shirorekha also helps in distinguishing Brahmic scripts from other scripts. Devanagari, Gurumukhi, Bengali, Manipuri, Oriya, and Gujarati are some examples of Brahmic scripts used in the Indian subcontinent.
6. **Featural system** In this type of writing system, features are represented by characters for making phonemes. Korean Hangul writing system is one of the best examples of featural systems where logographic Hanja is mixed with featural Hangul [40].

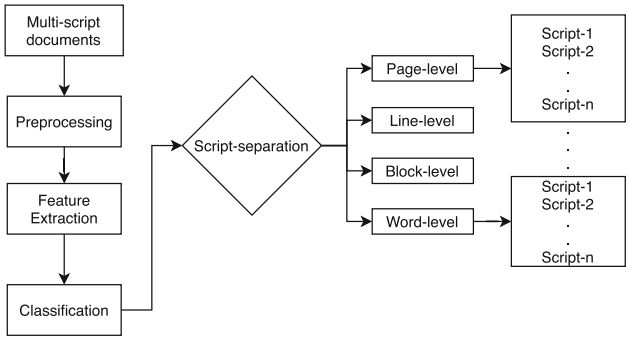

Fig. 4 Block diagram of multi-level script identification from eleven Indic scripts [101]

These six writing systems consist of various similar shaped symbol scripts that may sometimes lead to confusion during script identification. However, each writing system consists of different spatial characters, diacritics, or ligatures which helps in distinguishing one writing system from others.

4 Traditional script identification methods

As discussed in Section 2, script recognition methods can be categorized based on different types such as document/ page level, text block level, text line level, word level, and character level. These methods can be applied to either handwritten or printed documents. However, sometimes documents may consist of a mixture of both handwritten and printed scripts. The process of recognizing mixed script documents is generally referred to as hybrid script recognition. In general, the script recognition process can be categorized and applied to different types, i.e., printed vs. handwritten, cursive vs. non-cursive level, online vs. offline, document level, text block level, text line level, word level, and character level. However, some authors have also presented some multi-level script identification approaches in which the same document is considered at different levels, viz. block, line, word, character, etc. Obaidullah *et al.* [101] presented a multi-level script identification approach based on both script-dependent and script-independent features. Their objective was to test the performance of the script identification method when input is provided at different levels of granularity as shown in Fig. 4.

Input data consisting of eleven Indic scripts are provided in the form of a multi-script document to their proposed system followed by preprocessing/ segmentation to achieve further levels, i.e., block, line, and word. For classification, MLP and Random Forest classifiers were employed for comparing the performance at different levels. Experimental results on their own data set showed the better performance of MLP classifier in almost all situations, whereas the line level provides the most consistent results as compared to other levels.

In the following sub-sections, the authors reviewed and presented the summary of several works done by researchers for script identification at different levels of granularity.

4.1 Document-/ Page-level script identification

In document-level script identification, the process of script identification is carried out on the whole document at once. After processing, the document is further divided into pages, paragraphs, text lines, words, and characters so that the recognition of the exact letter takes place. Some researchers distinguish the process of script recognition at the document level and page level, but in general, the technical specifications are almost the same. That is why some researchers use the terms document-level script recognition and page-level script recognition interchangeably. The first step in page-level script identification is to detect the text area from the page. This process can be carried out by segmenting the page into two different segments, viz. text and non-text segments [121]. For this purpose, several texture segmentation techniques exist in research literature such as Discrete Wavelet Transform, Multi-channel filtering, histograms, etc. After identifying the texture part on a page, the next step is to extract the required features that play an important role in script identification. After extraction, these features are stored in a feature library for achieving the script identification [50].

Plenty of work has been carried out for document-/ page-level script identification by different researchers. As discussed above the document-level script identification can be carried out on both handwritten and printed documents. Some of the work in this area concerning handwritten and printed documents are summarized in Tables 1 and 2, respectively. K. Roy *et al.* [142] performed the document-level script recognition task on Devanagari, Bangla, Malayalam, Roman, Urdu, Oriya scripts by extracting component-based, topological-based, and fractal features. The component-based feature extraction technique was employed for extracting these features followed by MLP classifier and was able to achieve an overall classification accuracy of 84.21%. Obaidullah *et al.* [88] also employed a component analysis technique for extracting structural features followed by classification using MLP and achieved an accuracy of 92.8%. In [95] they presented a Visual Analytic-based Feature Fusion Framework (VA-FFF) for script identification from multi-lingual documents over Roman, Devanagari, Bangla, Oriya, and Urdu scripts. Using Structural Appearance (SA) and Directional Morphological Filter (DMF), they collected 54 features. DMF has the advantage over SA in feature identification when the structural properties of graphemes of two different scripts are the same. Using MLP classifier the average accuracy rate in the case of the VA-FFF approach was 95.6 as compared to 93.4 and 92.3 of SA and DMF, respectively. On the other hand, Obaidullah *et al.* [97] presented the page-level script identification from 11 Indic scripts using state-of-the-art classifiers based on feature combination framework over Structured Visual Appearance (SVA) and Directional Stroke Identification (DSI). Experimental analysis on their PHDIndic_11 data set reported more than 99% identification accuracy on both bi-script and tri-script cases. Table 1 represents the comparative study of some document-/ page-level script identification approaches based on feature extraction, classification, and accuracy rate.

Generally, texture-based feature extraction is common amongst several feature extraction techniques. It is evident that different scripts have different textures, so to devise a script identification system one must keep in mind the type of script and its texture features. In [88], a system for Indic script identification from handwritten documents is embodied. They have categorized all features into three categories, viz. Mathematical, Structural, and Script-based features. Mahmoud *et al.* [74] presented an optimal threshold selection based on the discriminant criterion [102]. They have used Hidden Markov Model (HMM) for the identification of scripts from handwritten documents with an overall classification accuracy of 51.2%.

On the other hand, Khoddami *et al.* [65] have used the concept of Curvature Scale Space features for representing different script levels. They first applied the feature extraction at the component-level and then extended the same to line-level and page-level granularity. Based on the k-NN classification mechanism, they were able to achieve the classification accuracy of multilingual documents up to 99%.

4.1.1 Discussion

Generally, the page-level script identification is a relatively complex task and have several challenges such as less amount of text is present on the page for script identification, the document is available in multiple scripts, the document is available in both handwritten and printed scripts, the spatial relationship among texture patterns, availability of noise [163], etc. To efficiently detect the script from a whole page or document, one must consider these challenges and find an optimal solution. Most of the work for document-/ page-level script identification is focused on texture-, statistical-, and structural-based feature extraction techniques followed by classification using MLP and k-NN classifiers. Extracting features with the help of other feature extraction techniques followed by classification using other classifiers, viz. Neural Networks, SVM, RBF, Decision Trees, etc. can be considered as future work for multilingual Indic script documents.

4.2 Text line-based script identification

In text line-based script identification, a single text line may be either of a single script or can be multi-lingual. In both

Table 1 Document-/ page-level script identification from handwritten documents

References	Features extracted	Feature Extraction/ Decomposition technique	Classifier used	Identification accuracy rate (best/ average)
[142]	Structural	Component-based feature extraction	MLP	84.20%
[88]	Structural	Component analysis	MLP	92.80%
[48]	Stroke	Morphological filters	k-NN	99.20%
[91]	Spatial	DWT	MLP	82.20%
[98]	Statistical	Transformation methods	MLP	94.30%
[99]	Structural	Component analysis	MLP, RF, Logistic Model Tree, etc.	91.20%
[142]	Fractal/ Component	Component analysis	MLP	89.48%
[150]	Statistical	MLP	MLP	99.29%
[165]	Texture	Gabor filtering	Multi-prototype classifier	91.60%
[50]	Texture	Co-occurrence histogram	k-NN	97.50%
[52]	Connected components	Novel approach	Linear Discriminant Analysis	88.00%
[165]	Texture features	Gabor filter	Fuzzy classification	91.60%
[50]	Texture	Histograms, Wavelet	k-NN	97.50%

Table 2 Document-/ page-level Script identification from printed documents

References	Features Extracted	Feature Extraction/Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[80]	Structural	Horizontal histogram	SVM	99.80%
[79]	Reference document vectors	Document vector similarity	Document vector similarity	95.63%
[71]	Document vectors	Word shape coding scheme	K-means clustering	91.56%
[105]	Texture	Wavelets, Haar	k-NN	99.33%
[110]	Principle strokes	Structural	Feature-based tree classifier	96.09%
[8]	Statistical	Profile coefficient method	Coefficient-based classifier	96.20%
[15]	Wavelet log Co-occurrence	Wavelet log co-occurrence	EM, GMM	95.12%
[59]	Statistical	Histograms	Novel approach	96.70%
[42]	Spatial	Various techniques	Tree, k-NN, SVM	100.00%
[43]	Low level	Normalization, Text localization	k-NN	89.00%
[51]	Gabor	Histograms	k-NN	98.00%
[53]	Structural	Connected component	Hierarchical clustering	98.00%
[34]	Stroke	Morphological reconstruction and pixel distribution	Morphological reconstruction, NN Classifier	97.00%
[106]	Texture	Wavelet / Haar	k-NN	99.63%
[108]	Texture	Wavelet/ Haar	k-NN	98.24%
[65]	Curvature Scale Space	Renormalized CSS and local maximums of CSS contours	Improved k-NN, Voting algorithm	100.00%
[68]	Statistical	Density distributors and entropy models	Voting-based classification, Novel algorithm	98.16%
[72]	Structural	Log–Gabor filter	Log–Gabor filter	98.10%

cases, text lines are segmented for achieving script identification. Most of the line-based script identification work is focused on printed documents instead of handwritten documents. Obaidullah *et al.* [96] proposed an automatic line-level script identification from handwritten documents for eight Indic scripts. They combined multiple features to form a generic script identifier. Structural-, directional-, and texture-based features were employed for this purpose, and the performance of each feature extraction technique was validated individually as well as collectively. In addition to multiple feature extraction, they employed multiple classifiers for script identification, viz. MLP, SVM, RF, and Fuzzy Unordered Rule Induction Algorithm (FURIA). The performance of MLP was observed to be outperforming in terms of average accuracy rate amongst all classifiers under consideration on their data set consisting of 2034 text lines from different sources. M.C. Padma *et al.* [107], presented the script identification from tri-lingual documents using profile-based features. They have devised separate algorithms for identification of top and bottom profiles for discriminating features of text lines of three languages under consideration. On the other hand, Marti and Bunke [79] presented a database of English handwritten text lines using projection profiles. U. Pal *et al.* [112] have presented an automated script identification for printed line-level documents. They first distinguished scripts based on 'Matra' features so that Bangla and Devanagari scripts are in one group and remaining scripts are in different groups. After initial classification, both Bangla and Devanagari scripts were differentiated using basic structural features. On the other hand, Chinese text lines were identified using vertical runs of pixels as well as run length smoothing [44], and further English text lines were distinguished from Arabic using some statistical features based on the water reservoir [109]. On the other hand, R. Gopakumar *et al.* [45] extracted structural features by incorporating the concept of zoning. They have conducted experiments on few South-Indian scripts along with English and Hindi. The process begins by identifying text lines from the document images followed by structural feature extraction. Kohavi [57] has discussed the importance of wrapper model over filter model for the subset selection. They have integrated three different elements in designing these methods, viz. procedure for searching features, function for evaluating features, and finally a classifier for identification. Tables 3 and 4 show some more work over text line-based script identification from handwritten and printed documents, respectively.

It is evident that feature selection does not depend on the learning methodology of the classifier because it uses the properties of the data itself. The main concept of a feature selection algorithm is to perform a searching methodology for extracting required features using an evaluation function. G.S. Rao *et al.* [137] also presented a method for text line-based script identification from a printed tri-lingual document of Hindi, Telugu, and English. Their main concept is to analyze the text lines using top and bottom profile features. For validating their approach, they had conducted experiments on a data set containing 300-text lines obtained of varying font types and sizes. The overall accuracy of their methods was found to be 99% using the k-NN classifier. In [39], Ferrer *et al.* proposed a novel method for text line-based script identification. For extracting stroke-based features, Local Binary Pattern (LBP) was utilized. The basic idea was to estimate the distribution using several regions (top, bottom, middle-upper, and lower middle) of the text line. Experimental results on public data sets using Least Square Support Vector Machine (LSSVM)-based classification provides 90% script identification accuracy. Most of the work on the line-level script identification is focused on global feature extraction followed by classification. In addition to the global feature extraction, the script identification process should also incorporate various optimization factors, i.e., feature dimensions, complexity, etc. that contribute to increasing the performance of the script identification process [89]. Obaidullah *et al.* [89] considered these optimization factors for separating scripts into two main classes, viz. with 'matra' and without 'matra.' Fractal Geometry Analysis (FGA), Canny Edge Detector (CED), and Morphological Line Transform (MLT) were considered for feature extraction followed by script classification using three state-of-the-art classifiers, viz. MLP, Bayes-Net (BN), and RF. Experimental analysis of all feature extraction and classification techniques under consideration is carried out on a handmade data set consisting of 1204 line-level document images. Experimental results indicate higher performance in the case of FGA feature extraction along with RF classifiers. The work presented by them is an extension of their earlier work [90].

4.2.1 Discussion

It is observed that only few works are available for handwritten script identification from text lines as compared to printed documents. In the case of printed documents, statistical and structural feature extraction techniques were mainly employed. Generally, the text line script identification suffers from several open research issues that need to be considered, i.e., incorrect segmentation of characters from connected words, short lines containing less number of characters, variety of text styles and sizes in a single text line, single feature extraction do not work properly for multi-lingual text lines, etc. Considering these open research issues while developing a generic text line-based OCR for Indic script identification is highly solicited.

Table 3 Text line-based script identification from handwritten documents

References	Features Extracted	Feature Extraction/Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[107]	Structure	Top and bottom profiles	k-NN	93.67%
[79]	Text lines	Novel approaches	Various approaches	98.80%
[84]	Spatial, Stroke	Various approaches	NN, SVM, Bayes, Tree	95.50%
[83]	Fractal	FSLaD	k-NN, RBF	98.72%
[89]	Matra	FGA, CFD, LT	MLP, BN, RF	95.68%

Table 4 Text line-based script identification from printed documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[112]	Structural	Modified run length smoothing	Novel approach	97.33%
[45]	Structural	Zone-based feature extraction	k-NN and SVM	100.00%
[103]	Geometrical	Novel algorithm	Novel algorithm	99.67%
[118]	Spatial	Contour tracing, Water reservoir	Tree classifier	97.52%
[67]	Matra	Novel approach	Hierarchical, Template matching	94.70%
[4,5]	Statistical	Horizontal projection profile	Rule-based classifier	99.83%
[113]	Principal stroke	Horizontal and Vertical profile	Segmentation	98.50%
[58]	Statistical	Log-Gabor filters	Hierarchical classifier	97.11%
[133]	Multiple features	Gabor filter with DCT & DWT	k-NN, SVM	96.70%
[56]	Structural	PCA	k-NN	95.00%
[69]	Structural	Hamming distance	Statistical technique	91.00%
[104]	Structural	Novel approach	Monothetic model	98.50%
[111]	Structural	Novel algorithm based on RLSA	Separation scheme	97.33%
[127]	Structural	PCA	k-NN	83.70%
[139]	Statistical, Structural	Novel approach	Novel approach	96.05%
[3]	Structural	Subset feature selection	k-NN	97.00%
[68]	Position of concavities, Optical density	Density distributors and Word unigram entropy	Voting-based classification and Novel algorithm	98.16%

4.3 Text block-level script identification

Text block-level script identification works by segmenting the whole document into equal sized text blocks consisting of many text lines. The size of text blocks may vary and sometimes needs padding if characters are on the boundary of a text block [100]. The presence of more than one word in a single text line can be considered as a special case of text block-based script identification. Sometimes paragraphs or parts of paragraphs are also considered as a part of text blocks. Developing a generalized text block-level OCR seems hard especially for multi-lingual countries like India. So, instead of developing a generalized OCR, it is preferred to develop script dependent OCRs in the case of text block-level script identification [87]. Block-level script identification plays a vital role in developing script dependent OCR. Obaidullah *et al.* [87] proposed an automatic handwritten script identification based on block-level script identification for six popular Indic scripts, viz. Devanagari, Oriya, Malayalam, Bangla, Urdu, and Roman. For constructing a feature vector, texture-based, statistical-based, and transformation-based techniques were employed. MLP classifier on several combinations of scripts under consideration provides promising script classification results. Table 5 shows some of the

Table 5 Text block-level script identification from printed and handwritten documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[120]	Rotation-invariant	Gabor filter, Steerability property	FFNN	98.50%
[125]	Texture	Gabor filters, Gray-level co-occurrence matrices	k-NN	95.00%
[21]	Chain-code histogram	Corresponding approach	SVM	99.85%
[26]	Gabor/ Texture	Gabor filter	Gabor filter	100.00%
[63]	Morphological	Morphological	k-NN	88.50%

work carried out by different researchers for text block-level script recognition.

On the other hand, Pan *et al.* [120] worked in this area by creating a Gabor filter bank using rotational invariant features followed by steerability properties to reduce the computation cost of feature extraction. For script classification, a Feed-Forward Neural Network (FFNN) was employed that has provided 98.5% classification accuracy. In [77], Manjula and Hegadi presented the script identification of Kannada and Hindi text using Edge Direction Histogram (EDH) and Maximized Mutual Information (MMI)-based feature extraction followed by classification using k-NN and SVM. Experimental analysis on 400 input images shown 100% classification accuracy using SVM as compared to 99.8% using the k-NN classifier.

4.3.1 Discussion

Only a few works are available for text block-based script identification from Indic scripts. The problem of developing script dependent text block-based OCRs for Indic scripts is still in the infancy stage that needs to be considered as a challenging task to solve some open research issues, viz. selection of appropriate feature selector and classifier, unavailability of decision parameters for uniform segmentation, availability of multiple scripts in a single text block, etc. Researchers need to consider these issues to develop generic multi-lingual text block-based OCRs for Indic scripts in the future.

4.4 Word-level script identification

Word-level script identification is more difficult as compared to the text line level because of several reasons, i.e., words may not be available in the form of straight text lines, some characters may be missing or incomplete in a word, skewed words, presence of noise, etc. The task of word-level script identification is one of the challenging tasks in the area of document analysis and recognition. Plenty of researchers tried to solve the problem of word-level script identification; few of them are summarized in this section. U. Pal *et al.* [110] described a word segmentation-based approach for automatic script identification from printed Indic documents. Their work was focused on separating Bangla, Devanagari, and Roman scripts. First, the paragraphs are subdivided into text lines and text blocks followed by word-level segmentation using histogram-based features. The computation of text lines and word segmentation is based on horizontal and vertical scanning, respectively. Generally, Bangla and Devanagari scripts are difficult to separate because of their structural similarity. For separating Bangla with Devanagari script, stroke-based as well as projection profiles were used. The accuracy rates of their different script identification mechanism were 94.88%, 95.28%, and 98.12% for Bangla, Devanagari, and Roman, respectively, with an overall accuracy of 96.09% [110]. Shivakumara *et al.* [160] presented the task of word-wise script identification using Gradient Angle Features (GAF) for video text lines. For the segmentation of words from video, they have extracted gradient directional features. They aim to study the cursiveness properties of Potential Text Candidates (PTC) from video frames using GAFs. On the other hand, Pal *et al.* [115] presented a methodology for pin-code recognition from Indian multilingual documents. In general, due to a variety of handwriting styles digits of pin-code got overlapped with others that creates a problem in segmentation. To overcome this problem, they have incorporated the top and bottom water reservoir-based concept for achieving over-segmentation instead of under segmentation and performed classification of digits using dynamic programming. Experimental results on Bangla, Devanagari, and English showed classification accuracy up to 97.18%.

S. Sinha *et al.* [166] proposed a technique for printed documents to identify the scripts using word-wise script identification. They have considered pairs of Indic scripts from a data set of 7500-words and a data set of 24210 words for testing. The average accuracy rate of these word-wise

Table 6 Word-level script identification from handwritten documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[141]	Matra, Water reservoir	Statistical methods	Tree classifier	89.00%
[144]	Structural	RLSA, DAB	MLP	97.62%
[143]	Structural	Novel approach based on RLSA	MLP	96.79%
[140]	Structural	Novel algorithm	MLP, SVM, k-NN, etc.	99.20%
[148]	Various techniques	Co-occurrence matrix of oriented gradients	k-NN	99.85%
[147]	Structural	NA	Naïve Bayes	98.40%
[164]	Various techniques	Approximation	Various classifiers	95.35%
[10]	Texture	Steerable pyramid transform	k-NN	99.00%
[11]	Texture	Pyramid sub bands	k-NN	99.00%
[33]	Structural	Morphological erosion, Average pixel distribution	k-NN	98.12%
[32]	Global and Local	Regional descriptors, Morphological filters	k-NN	99.00%
[49]	Horizontal, Diagonal	Directional DCT	k-NN	96.42%

Indic script pairs was 97.92%. Another word-wise identification approach has been presented by S. Chanda *et al.* [19] from printed documents containing Sinhala, English, and Tamil scripts. They have used structural, topological and water-reservoir principles (as described in [118])-based features extraction techniques. The accuracy rate of their script identification mechanism was observed to be 98%. Tables 6 and 7 show some other word-wise script identification techniques containing different feature extraction and classification techniques along with their accuracy rates.

4.4.1 Discussion

Word-level script identification is one of the famous script identification approaches. Most of the work in handwritten word-level script identification is focused on analyzing the structural as well as the statistical properties of the words. As per the classifier is concerned, only k-NN and MLP were mainly employed and proved to provide satisfactory accuracy rates for word-level script identification. In the case of printed word-level script identification, Gradient and Gabor feature extraction were mainly involved followed by SVM-based classification. For a successful word-level generic OCR, there is a vital need to consider the modern deep learning methodologies along with considerations of some open research issues, viz. availability of non-uniform sized and styled words, skewed words, presence of noise, segmentation of connected words, multilingual words, etc.

4.5 Character-level script identification

Character-level script identification works on the principle of extracting features from different characters. As shown in Table 8, Pal *et al.* [117] have used various dimensional features for recognition of numerals from six popular Indian scripts. With the help of fivefold cross-validation, they have obtained classification accuracy near to 99%. Raghunandan *et al.* [131] presented two novel approaches for text identification from multi-script video and image. They first identified the text regions from video scenes using Iterative Nearest Neighbor Symmetry (INNS) followed by text pair identification using Mutual Nearest Neighbor Pair (MNNP). For identification of a rectangular window for arbitrary orientated words, an angular relationship was computed between fused band and sub-band. Wavelet features were extracted from each window followed by SVM classification. On the other hand, Razzak *et al.* [138] presented script identification from Urdu and Arabic scripts. It is one of the challenging tasks to recognize Urdu like scripts where the characters are found to be more cursive than other scripts. With the help of directional features and fuzzy rules, they have obtained the script recognition accuracy to be 96.3%. For printed English and Chinese characters, Xiao-Rong Lin *et al.* [70] presented

Table 7 Word-level script identification from printed documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[121]	Script dependent	Gabor filter	Multi-channel filtering	99.56%
[122]	Texture	Gabor filter	LDA, k-NN	94.70%
[35]	Directional, Spatial	Gabor filter	SVM, NN, k-NN	96.03%
[73]	Feature vector	Gabor, Wavelet	Multi-class classifier	92.08%
[119]	Texture	LBP	BEMD-LBP, Wavelet, LBPV	95.41%
[123]	Structural, Gabor	Novel	k-NN, SVM, LD	99.60%
[124]	Discriminative	Novel	Feed-Forward Neural Network	96.23%
[154]	Texture, Gradient	Gabor, Gradient filters	SVM	87.50%
[156]	Gradient, Texture	LBP, HOG, Gradient Local auto correlation	SVM, ANN	94.25%
[155]	Gradient, Texture	Bag-of-features, Spatial pyramid matching	SVM	99.22%
[159]	Discriminative	LLC, CNN-patch, GLCM	Multi-stage spatially sensitive pooling network	94.40%
[158]	Discriminative	VLFeat	Discriminative CNN (DisCNN)	88.60%
[160]	Gradient-angular	Novel (GAF)	k-NN	88.20%
[161]	Structural	Gradient histograms	Gradient spatial structural features	83.00%
[6]	Wavelet	Wavelet, Bounding box, Binarization	Fuzzy inference scheme	94.33%
[23]	Structural	Horizontal projection profile	SVM	99.36%
[19]	Structural	Histogram, Segmentation, smoothing	SVM	98.30%
[20]	Structural	Various techniques	Tree classifier	98.09%
[29]	Spatial	Novel approach	Heuristic-based approach	93.00%
[55]	Gabor	Gabor filter analysis	k-NN, SVM	97.51%

a method based on Decision Tree-SVM with an accuracy of 99.6%. Some of the work in the field of handwritten as well as printed character-level script recognition is also summarized in Tables 8 and 9, respectively.

Halder *et al.* [47], presented a character-level handwritten script identification from a large volume of 53250 Bangla characters. They have modified and evaluated three transformation-based feature extraction techniques, viz. Fast Fourier Transform (FFT), Gray-Level Co-occurrence Matrix (GLCM), and Discrete Cosine Transform (DCT). FFT is generally preferred to transform pixel values to a periodic sequence of complex numbers. GLCM on the other hand statistically computes the occurrences of gray-level pixels in an image. Whereas DCT computes image coefficients with the help of cosine function as compared to coefficient calculation using both sine and cosine functions in FFT. Using SVM classification, they were able to achieve a writer identification accuracy of 98.62% by combining the three transformations. On the other hand, Sharma and Dhaka [31,151–153] presented potential results for the recognition of handwritten English characters and numerals using feed-forward neural networks along with some image processing techniques. Manjula and Hegadi [78] presented an LBP-based approach for feature extraction followed by classification using cubic SVM and weighted k-NN from multi-lingual documents containing Oriya and English scripts. Experimental results on 3000-character data shown higher classification accuracy of 96.5% using the weighted k-NN approach.

Table 8 Character-level script identification from handwritten documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[117]	Directional	Directional, Bounding box, Gaussian filter	Modified quadratic classifier	99.56%
[138]	Shape	Feature purification process	Fuzzy Rule-based approach	96.30%
[162]	Structural	Binarization	Feed-Forward NN	96.15%
[27]	Structural	Binarization	Feed-forward NN	99.10%

4.5.1 Discussion

As per the work in character-level script identification is concerned, mostly the work was focused on extracting structural and statistical features using neural network and SVM-based classification. The main problem with character-level script identification is the correct segmentation of characters from different scripts. To date, no segmentation technique proved to be a generalized one that can be applied to a variety of scripts. There may be chances that one segmentation technique works very well with one script but unable to accurately segment the characters of different scripts. For developing a generic OCR for Indic script identification based on character-level script identification there it is vital to develop a generic segmentation technique that can be applied to multiple scripts at once.

5 Script identification using deep learning-based approaches

Deep learning-based script identification is gaining popularity nowadays because of their several advantages as compared to traditional image processing techniques. On the other hand, it is observed that for both online and offline modalities, researchers adopted separate approaches for script classification. Nowadays, the process of hybrid script recognition that combines both modalities is also gaining popularity. Bhunia *et al.* [14] proposed a novel multi-model deep network that takes into account both online and offline modalities for script identification. They first converted handwritten data from both modalities and then fed the results to their deep network for avoiding the need of two separate script identification modules as shown in Fig. 5. Experimental results on six Indic scripts along with Roman script outperformed traditional script recognition methods. They showed that character-level training can achieve promising classification accuracy as compared to traditional word-level approaches. Carbune *et al.* [16] described an online handwriting approach using Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN). Their system is fast as well as supports multiple languages. The system is currently being used at Google and supporting 102 languages in 26 scripts. Zheng *et al.* [174] presented an approach to extract displacement features from pooling layers for text recognition. Generally, traditional pooling layers are intended to fetch the maximum value from the pooling window without keeping track of the location, from where the maximum value arises. But in their work, they have implemented the concept of displacement features for retaining the spatial information. Displacement features keep track of the position of maximum value from the pooling window.

Bhunia *et al.* [13] proposed a novel script identification method based on LSTM for the extraction of local and global features. They have evaluated their work on four multilingual data sets and achieved script identification accuracy up to 97.75%. The problem of word spotting is also gaining popularity nowadays. In this regard, Sudholt and Fink [167] proposed PHOCNet, a deep CNN for word spotting in handwritten documents. On the other hand, Ghosh *et al.* [41], presented online word recognition from Devanagari and Bengali scripts using RNN. They divided each word horizontally into three zones, viz. upper, middle, and lower, followed by the extraction of structural features. The structural features are then fed into LSTM and BLSTM versions of RNN for training and testing purposes. Evaluation of their method on large scale data sets shown the script recognition accuracy to be 99.50% and 95.24% on Devanagari and Bengali scripts, respectively. Keserwani *et al.* [64] presented the task of word-level script identification using a zero-shot learning-based approach. For extracting stroke sequence, they have used the pre-trained VGG-16 model that accentuates both global and sequential features. Ansari *et al.* [7] presented a novel approach based on CNN for the identification of natural scene text. For identification of text and non-text regions, LBP and T-HOG (Histogram of Oriented Gradients (HOG)-based texture descriptor) feature sets are combined with SVM. After the detection of text regions, training is being carried out by CNN. Evaluation of multiple data sets shown promising results as compared to other state-of-the-art approaches. Rabbya *et al.* [130] presented a lightweight CNN model for offline handwritten characters from Bangla script. Evaluation

Table 9 Character-level script identification from printed documents

References	Features Extracted	Feature Extraction/ Decomposition Technique	Classifier used	Identification accuracy rate (best/ average)
[70]	Statistical	Decision Tree-SVM	DT-SVM	99.60%
[116]	Water reservoir, Topological, etc.	Projection profile	Feature-based tree classifier	97.80%
[126]	Structural	Novel approach	Distance-based classifier	98.89%
[128]	Gabor	Novel approach	k-NN	96.50%
[136]	Gabor	Gabor Filter	SVM	98.90%
[135]	Structural	Novel, Feature vector, Gabor filter	SVM, k-NN, NN	99.40%
[22]	Statistical	Various techniques	Tree classifier	98.79%
[18]	Gradient	Various techniques	SVM	98.51%
[17]	Zernike moment, Gradient	Feature vector, Morphology, PCA	SVM	81.39%
[24]	Structural features	Statistical	k-NN	99.23%

of three data sets shown the character recognition accuracy up to 98%. Mane and Kulkarni [75] have proposed a customized CNN that can learn features automatically and can predict the class of Marathi numerals, whereas Samanta *et al.* [149] presented a Hidden Markov Model-based word recognition scheme for handwritten cursive characters. Ukil *et al.* [171] presented a deep learning-based word-level script identification from 11 Indic scripts (PHDIndic_11 [97]). Instead of using handcrafted features, they have selected features using different CNNs followed by script identification tasks using a MLP classifier. Their main concept behind extracting the features is to incorporate two- and three-layered CNNs on varying scales of input images. After extraction and merging of relevant features, script identification is performed and has provided promising results over other state-of-the-art techniques. DeepWriter on the other hand is a deep multi-stream CNN for recognizing writers developed by Xing and Qiao [173]. The structure of DeepWriter is like AlexNet, in which handwritten patches are resized while maintaining aspect ratio. Testing and training on IAM and HWDB1.1 shown writer identification accuracy up to 99.01%. On the other hand, Jaderberg *et al.* [54] presented a state-of-the-art approach for text recognition from natural scene images and text-based image retrieval. Various text recognition and spotting data sets as well as enough literature have been presented in their work with promising results. Morera *et al.* [81] addressed the recognition of gender and handedness using CNN offline handwriting. Handedness is considered as binary classification, in which the prediction of 'right-handed' or 'left-handed' persons are predicted with the help of their handwriting. They have evaluated their novel approach on real-world data sets and presented significant results.

5.1 Discussion

With the advancements in deep learning approaches, the process of script recognition becomes easier nowadays. DL provides various advantages over traditional image processing frameworks for example, automatic feature extraction, ability to process unstructured data, no data labeling is required, etc. But in addition to the advantages, DL suffers from several drawbacks too, such as the huge amount of data is needed for achieving high accuracy, over-fitting, lack of transparency (sometimes hard to distinguish what imaging features are used to determine the output), requires high quality images, requires high end machines for complex computations, etc. Despite these drawbacks, in script recognition DL suffers from some other challenges as well, viz. difficulties in processing non-uniform images (different words may be of different sizes), cross-language feature extraction and zone segmentation, variety of hand-writings, etc. These challenges need to be addressed by researchers while developing generic OCRs for Indic scripts using deep learning-based approaches.

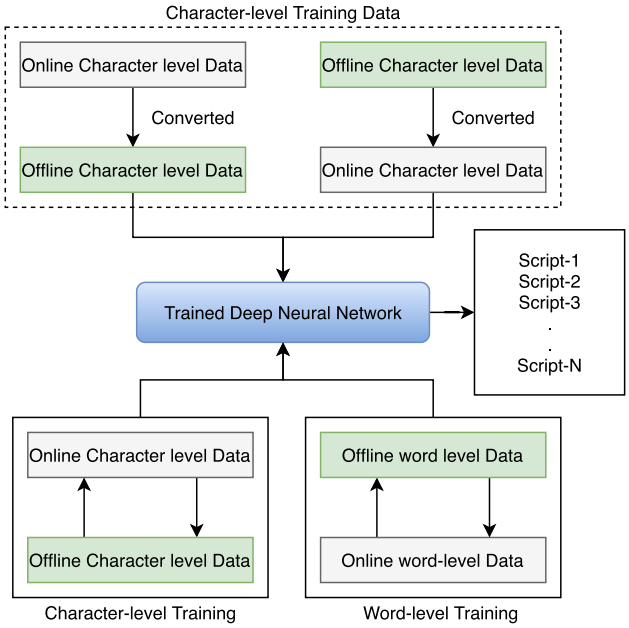

Fig. 5 Multi-model (offline–online) script identification framework [14]

6 Some public data sets for Indic script identification

Generally, the biggest challenge in undertaking the script identification task is the availability of public data sets. This section provides researchers with the sources of some famous public data sets available for carrying out the script identification task. We have identified some data sets that are freely available for the public to conduct a script identification task as shown in Table 10.

7 Summary of few other script identification survey articles

While preparing this survey article, we have investigated several other survey articles of this nature as highlighted in this section. It is observed that very few survey articles are available for Indic script identification from both handwritten and printed modalities. At the end of this section, we will present the major contributions of our work.

1. **Bashir *et al.* [9] (2018, 109 sources, 1986–2015)** Presented a summary of 109 research articles of the past few decades by categorizing them into two main cat-

Table 10 Some public data sets for Indic script identification

Source	Modality	Level	Description
[94]	Printed	Word	26,000 words of 13 Indic scripts
[97]	Handwritten	Document	1458 documents of 11 official Indic scripts
[12]	Handwritten, Scene Character	Character	Several offline and online databases of Bangla, Devanagari and Oriya
[146]	Handwritten	Word, Character	Several databases of Bangla and Devanagari scripts
[1]	Handwritten, Printed	Document, text line, word, character	Various data sets for Indic script identification
[93]	Handwritten	Word	Numeral data sets over Bangla, Devanagari, Roman and Urdu scripts
[37,38]	Handwritten	Word	Devanagari and Telugu data sets
[145]	Scene text	Word, Character	Scene text data for Telugu, Malayalam and Devanagari scripts
[36]	Handwritten	Character	Devanagari 5137 numerals and 20305 characters database
[30]	Handwritten	Character	55278 Bangla characters database
[168]	Handwritten	Word	265 Tamil city names database
[92]	Handwritten	Word	5659 word database of Devanagari, Urdu, Bangla, and Roman scripts
[86]	Handwritten	Word	100,000 words, each of Kannada and Tamil

egories, viz. International scripts and National (Indic) scripts.

2. **Obaidullah** *et al.* [100] **(2018, 14 sources, 1986–2017)**
   A comprehensive survey of handwritten multi-script document images of Indic scripts. Presented a summary of 74 research articles and categorized them based on modalities, feature extraction, and classification techniques. In addition to these, they have also presented some Indic script data sets and few challenges to Indic scripts identification.
3. **Ubul *et al.*** [170] **(2017, 134 sources, 1989–2016)**
   Presented survey of multi-script documents into two main categories, viz. printed, handwritten and hybrid at different levels of granularity (i.e., document/ page, text block, text line, word, and character). Also presented various features for script identification into global and local classes. In addition to these, they have presented several classification techniques, some public data sets and finally analyzed the performance of several research articles based on scripts, features, classifiers, and accuracy rates.
4. **Manjula and Hegadi** [76] **(2017, 18 sources, 1995–2014)**
   Presented a survey of multilingual document analysis based on Indic scripts. Highlighted implementation techniques, drawbacks, and accuracy rates of 13 research articles. Included a total of 18 bibliographic articles.
5. **Singh *et al.*** [164] **(2014, 97 sources, 1973–2014)**
   Presented a state-of-the-art survey of multilingual offline Indic documents. Classified work of few decades into two main classes, viz. structure-based and visual appearance-based script identification, at different levels of granularity. In addition to this, they have provided a comparative study of several works based on methodology, script types, modalities, granularities, and script identification accuracies. Finally, they have provided the scope of future directions in developing offline script identification-based OCRs for Indic scripts.

6. **Ramteke and Rane** [134] **(2012, 64 sources, 1973–2011)** Presented a summary of several research articles focused on offline handwritten script identification of Devanagari script. Classified various approaches based on script identification process, viz. preprocessing, feature extraction, and classification. Finally, presented some discussions and issues with future directions.
7. **Ghosh *et al.*** [40] **(2010, 85 sources, 1986-2006)** Presented a summary of several writing systems and script recognition methodologies. Classified various approaches into structure-based and appearance-based script identifications at each level of granularity, i.e., page, text block, text line, word, and character level. Also presented a comparative study of few approaches of structure-based and appearance-based approaches concerning modalities, features, classifier, script types, and best accuracy rate. In addition to these, they have also discussed online script/ video text recognition along with issues in developing multiscript OCRs. A total of 85 bibliographic articles were used in this review article.
8. **Pal and Chaudhuri** [114] **(2004, 121 sources, 1969–2003)** Presented a review of OCRs for printed Indian script recognition. At that time there were only a few OCRs were available for handwritten Indic scripts. They have classified OCRs for Devanagari, Bangla, Tamil, Telugu, Oriya, Gurumukhi, Gujarati, and Kannada. At last, they have presented some future directions for developing OCRs for poor quality documents, multi-fonts, Bi-Script/ multi-scripts, handwritten characters, and visually handicapped persons.
9. **Mori *et al.*** [82] **(1992, 193 sources, 1957–1991)** Presented the development and historical perspective of OCRs. The development of OCRs is further classified into two main categories, viz. template matching and structural analysis. In addition to this, they have presented the commercialization of OCRs along with their generations. At last, few learning algorithms were discussed with concluding remarks for future research directions.
10. **Plamondon** *et al.* [129] **(1989, 180 sources, 1929–1988)** Survey of automatic signature and writer identification. Their major focus was to discuss preprocessing, feature extraction techniques, and comparison and analysis of various state of the art. Also, they have presented several issues and challenges in signature verification/ writer identification.

7.1 Highlights of this survey article

In this paper, we have examined and summarized 174 sources for handwritten as well as printed script identification. We aimed to focus the work done in the past few decades on Indic script identification using both traditional image processing approaches as well as modern deep learning-based approaches on different levels of granularity. The summary of our contributions incorporated in this survey article is mentioned as follows:

1. In-dept review of several research articles on different granularity levels, classifiers, and accuracy rates.
2. Comparison with other script identification-based state-of-the-art survey articles.
3. Presented few public data sets for Indic script identification.
4. Highlighted several open research challenges and future directions for script identification.

8 Discussion, challenges, and future directions

Several attempts have been made in the research literature about script identification from online as well as offline documents. This paper has presented a survey of existing work for offline script identification from handwritten as well as printed documents. To date, only a few survey articles are available for script identification that has addressed all modalities. Some survey articles are focused on specific modalities of script identification. A comprehensive survey of script identification approaches focused mainly on handwritten documents of Indic scripts only is presented by Obaidullah *et al.* [100]. They have presented the existing script identification literature of both online and offline approaches. In addition to these approaches, they have presented some public data sets that are focused on Indic handwritten scripts along with features, classifiers, and challenges. After reviewing several works on script identification, many difficulties in achieving this task are observed and the same has been reported in this section. Generally, researchers are facing more difficulties in identifying handwritten scripts as compared to printed scripts [97] mainly due to:

1. Different writing styles of individuals.
2. Lack of symmetry among handwritten characters.
3. Skewed words.
4. Irregular spaces between words.
5. Non-uniform sized words/ characters.
6. Selection of feature sets for diversified scripts in a multilingual environment.

The issue of skew detection and correction is also reported by Jundale and Hegadi [60]. Skew is a form of distortion in words that may occur due to several reasons, i.e., noise during scanning a document for OCR, writing behavior, etc. For a good OCR system, it is foremost required to perform skew detection and correction [62]. It is quite difficult to detect

and correct skew for handwritten words as compared to the printed one. Generally, skew can be of three types, viz. global skew that is uniform to the whole document which generally arises during scanning of a document; local skew that is only present in a single text line; and multiple skews that appears in multiple text lines [62]. Only a few works are available till now for skew detection and correction [60–62] on handwritten Indic scripts with promising accuracy rates. OCRs are generally script dependent; hence, developing a generic skew detector and corrector for Indic handwritten scripts is in demand and challenging nowadays. Out of several works summarized in this paper, no generic script identifier has been observed. However, few of the works proved themselves to be optimum one, but still, the journey towards generic script identifier needs many milestones.

9 Conclusion

Over the last few decades, researchers have proposed numerous techniques for the identification of scripts from multilingual documents at a different level of granularity, i.e., document level, paragraph level, text line level, word level, and character level. This paper has examined the work done in the past two decades for script identification at different granularity levels. We have also presented comparative studies using a systematic survey of contributions made by different researchers in the field of multilingual script identification based on these levels. Some researchers have presented hybrid approaches by combining several approaches for both feature extraction and classification. Only a few works are available for handwritten Indic script identification using recent state-of-the-art techniques, i.e., Deep Learning. However, with the advancements in Deep Learning paradigms, the task of script identification becomes much easier as well as accurate. Most of the work done in recent years is based on traditional classifiers especially k-NN, SVM, and MLP. It is observed that MLP proved to be efficient amongst them script identification at different levels. The average accuracy rate is considered as one of the main criteria for presenting the comparative studies at different levels. However, it is hard to compare different script identification methods due to the diversity, size, and types of data set incorporated into different approaches. It is observed that variation in script identification accuracy of different approaches depends on several criteria, viz. input samples, levels of script identification, feature extraction techniques, etc. On the other hand, only a few works are reported for Indic script identification for both handwritten and printed offline documents. It can be concluded that there has been an instant need to work on multi-lingual script identification, especially for Indic scripts using deep learning-based paradigms. In addition to the above, developing generalized OCRs for the Indian subcontinent will be highly appreciable in the future.

References

1. Center for microprocessor application for training education and research (cmater. https://code.google.com/archive/p/cmaterdb/
2. Morphological image processing. https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic4.htm
3. Ablavsky, V., Stevens, M.R.: Automatic feature selection with applications to script identification of degraded documents. In: ICDAR, pp. 750–754. Citeseer (2003)
4. Acharya, D.U., Gopakumar, R., Aithal, P.K.: Multi-script line identification system for indian languages. J. Comput. **2**(11), 107–111 (2010)
5. Aithal, P.K., Rajesh, G., Acharya, D.U., Subbareddy, N.K.M.: 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)
6. Angadi, S.A., Kodabagi, M.: A fuzzy approach for word level script identification of text in low resolution display board images using wavelet features. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1804–1811. IEEE (2013)
7. Ansari, G.J., Shah, J.H., Yasmin, M., Sharif, M., Fernandes, S.L.: A novel machine learning approach for scene text extraction. Future Gener. Comput. Syst. **87**, 328–340 (2018)
8. Bashir, R., Quadri, S.: Identification of kashmiri script in a bilingual document image. In: 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), pp. 575–579. IEEE (2013)
9. Bashir, R., Quadri, S., Giri, K.J.: Script identification: a review. Int. J. Inf. Technol. pp. 1–15 (2018)
10. Benjelil, M., Kanoun, S., Mullot, R., Alimi, A.M.: Arabic and latin script identification in printed and handwritten types based on steerable pyramid features. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 591–595. IEEE (2009)
11. Benjelil, M., Mullot, R., Alimi, A.M.: Language and script identification based on steerable pyramid features. In: 2012 International Conference on Frontiers in Handwriting Recognition, pp. 716–721. IEEE (2012)
12. Bhattacharya, U.: Indian scripts character database (isical). https://www.isical.ac.in/~ujjwal/download/database.html
13. Bhunia, A.K., Konwer, A., Bhunia, A.K., Bhowmick, A., Roy, P.P., Pal, U.: Script identification in natural scene image and video frames using an attention based convolutional-lstm network. Pattern Recogn. **85**, 172–184 (2019)
14. 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)
15. Busch, A., Boles, W.W., Sridharan, S.: Texture for script identification. IEEE Trans. Pattern Anal. Mach. Intell. **27**(11), 1720–1732 (2005). https://doi.org/10.1109/TPAMI.2005.227
16. Carbune, V., Gonnet, P., Deselaers, T., Rowley, H.A., Daryin, A., Calvo, M., Wang, L.L., Keysers, D., Feuz, S., Gervais, P.: Fast multi-language lstm-based online handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR) pp. 1–14 (2020)
17. Chanda, S., Franke, K., Pal, U.: Identification of indic scripts on torn-documents. In: 2011 International Conference on Document Analysis and Recognition, pp. 713–717. IEEE (2011)

18. Chanda, S., Pal, S., Franke, K., Pal, U.: Two-stage approach for word-wise script identification. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 926–930. IEEE (2009)
19. Chanda, S., Pal, S., Pal, U.: 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)
20. Chanda, S., Pal, U.: English, devanagari and urdu text identification. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 538–545. Citeseer (2005)
21. Chanda, S., Pal, U., Franke, K., Kimura, F.: Script identification–a han and roman script perspective. In: 2010 20th International Conference on Pattern Recognition, pp. 2708–2711. IEEE (2010)
22. Chanda, S., Pal, U., Kimura, F.: Identification of japanese and english script from a single document page. In: 7th IEEE International Conference on Computer and Information Technology (CIT 2007), pp. 656–661. IEEE (2007)
23. Chanda, S., Terrades, O.R., Pal, U.: Svm based scheme for thai and english script identification. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 1, pp. 551–555. IEEE (2007)
24. Chaudhari, S.A., Gulati, R.M.: 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)
25. Chaudhuri, B., Pal, U.: A complete printed bangla ocr system. Pattern Recogn. **31**(5), 531–549 (1998)
26. Chaudhury, S., Sheth, R.: Trainable script identification strategies for indian languages. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318), pp. 657–660. IEEE (1999)
27. Choudhary, A., Ahlawat, S., Rishi, R., Dhaka, V.S.: Performance analysis of feed forward mlp with various activation functions for handwritten numerals recognition. In: 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), vol. 5, pp. 852–856. IEEE (2010)
28. Dalal, S., Malik, L.: A survey for feature extraction methods in handwritten script identification. Int. J. Simul. Syst. Sci. Technol. **10**, 1–7 (2009)
29. Das, M.S., Rani, D.S., Reddy, C.: Heuristic based script identification from multilingual text documents. In: 2012 1st International Conference on Recent Advances in Information Technology (RAIT), pp. 487–492. IEEE (2012)
30. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A benchmark image database of isolated bangla handwritten compound characters. IJDAR **17**(4), 413–431 (2014)
31. Dhaka, V., et al.: Offline language-free writer identification based on speeded-up robust features. Int. J. Eng. **28**(7), 984–994 (2015)
32. Dhandra, B., Hangarge, M.: 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)
33. Dhandra, B., Mallikarjun, H., Hegadi, R., Malemath, V.: Word-wise script identification based on morphological reconstruction in printed bilingual documents (2006)
34. Dhandra, B., Nagabhushan, P., Hangarge, M., Hegadi, R., Malemath, V.: 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. Dhanya, D., Ramakrishnan, A.: Script identification in printed bilingual documents. In: International Workshop on Document Analysis Systems, pp. 13–24. Springer (2002)
36. Dongre, V.J., Mankar, V.H.: Development of comprehensive devnagari numeral and character database for offline handwritten character recognition. Appl. Comput. Intell. Soft Comput. **2012**, (2012)
37. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Offline handwriting recognition on devanagari using a new benchmark dataset. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 25–30. IEEE (2018)
38. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Towards spotting and recognition of handwritten words in indic scripts. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 32–37. IEEE (2018)
39. Ferrer, M.A., Morales, A., Pal, U.: Lbp based line-wise script identification. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 369–373. IEEE (2013)
40. Ghosh, D., Dube, T., Shivaprasad, A.: Script recognition-a review. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2142–2161 (2010)
41. 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)
42. Ghosh, S., Chaudhuri, B.B.: Composite script identification and orientation detection for indian text images. In: 2011 International Conference on Document Analysis and Recognition, pp. 294–298. IEEE (2011)
43. Gllavata, J., Freisleben, B.: Script recognition in images with complex backgrounds. In: Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005., pp. 589–594. IEEE (2005)
44. Gonzalez, R.C., Woods, R.E.: Digital image processing (2002)
45. Gopakumar, R., Subbareddy, N., Makkithaya, K., Acharya, D.U.: Script identification from multilingual indian documents using structural features. J. Comput. **2**(7), 106–111 (2010)
46. Guru, D., Ravikumar, M., Harish, B.: A review on offline handwritten script identification. Int. J. Comput. Appl. **975**, 8878 (2012)
47. Halder, C., Obaidullah, S.M., Roy, K.: Offline writer identification from isolated characters using textural features. In: Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015, pp. 221–231. Springer (2016)
48. Hangarge, M., Dhandra, B.: Offline handwritten script identification in document images. Int. J. Comput. Appl. **4**(6), 6–10 (2010)
49. Hangarge, M., Santosh, K., Pardeshi, R.: Directional discrete cosine transform for handwritten script identification. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 344–348. IEEE (2013)
50. Hiremath, P., Pujari, J.D., Shivashankar, S., Mouneswara, V.: Script identification in a handwritten document image using texture features. In: 2010 IEEE 2nd International Advance Computing Conference (IACC), pp. 110–114. IEEE (2010)
51. Hiremath, P.S., Shivashankar, S.: Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image. Pattern Recogn. Lett. **29**(9), 1182–1189 (2008)
52. Hochberg, J., Bowers, K., Cannon, M., Kelly, P.: Script and language identification for handwritten document images. Int. J. Doc. Anal. Recogn. **2**(2–3), 45–52 (1999)
53. Hochberg, J., Kelly, P., Thomas, T., Kerns, L.: Automatic script identification from document images using cluster-based templates. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 176–181 (1997)
54. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vision **116**(1), 1–20 (2016)
55. Jaeger, S., Ma, H., Doermann, D.: Identifying script on word-level with informational confidence. In: Eighth International Confer-

ence on Document Analysis and Recognition (ICDAR'05), pp. 416–420. IEEE (2005)
56. Jindal, M., Hemrajani, N.: Script identification for printed document images at text-line level using dct and pca. IOSR J. Comput. Eng. **12**(5), 97–102 (2013)
57. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Machine Learning Proceedings 1994, pp. 121–129. Elsevier (1994)
58. Joshi, G.D., Garg, S., Sivaswamy, J.: Script identification from indian documents. In: International Workshop on Document Analysis Systems, pp. 255–267. Springer (2006)
59. Juan Cheng, Xijian Ping, Guanwei Zhou, Yang Yang: Script identification of document image analysis. In: First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06), vol. 3, pp. 178–181 (2006). https://doi.org/10.1109/ICICIC.2006.518
60. Jundale, T.A., Hegadi, R.S.: Skew detection and correction of devanagari script using hough transform. Proc. Comput. Sci. **45**, 305–311 (2015)
61. Jundale, T.A., Hegadi, R.S.: Skew detection of devanagari script using pixels of axes-parallel rectangle and linear regression. In: 2015 International Conference on Energy Systems and Applications, pp. 480–484. IEEE (2015)
62. Jundale, T.A., Hegadi, R.S.: Skew detection and correction of devanagari script using interval halving method. In: International Conference on Recent Trends in Image Processing and Pattern Recognition, pp. 28–38. Springer (2016)
63. Kanoun, S., Ennaji, A., LeCourtier, Y., Alimi, A.M.: Script and nature differentiation for arabic and latin text images. In: Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 309–313. IEEE (2002)
64. Keserwani, P., De, K., Roy, P.P., Pal, U.: Zero shot learning based script identification in the wild. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 987–992. IEEE (2019)
65. Khoddami, M., Behrad, A.: Farsi and latin script identification using curvature scale space features. In: 10th Symposium on Neural Network Applications in Electrical Engineering, pp. 213–217. IEEE (2010)
66. Krishnan, P., Jawahar, C.: Hwnet v2: an efficient word image representation for handwritten documents. IJDAR **22**(4), 387–405 (2019)
67. Kumar, B., Bera, A., Patnaik, T.: Line based robust script identification for indianlanguages. Int. J. Inf. Electron. Eng. **2**(2), 189 (2012)
68. Lee, D.S., Nohl, C.R., Baird, H.S.: Language identification in complex, unoriented, and degraded document images. In: Document Analysis Systems II, pp. 17–39. World Scientific (1998)
69. Li, L., Tan, C.L.: Script identification of camera-based images. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
70. Lin, X.R., Guo, C.Y., Chang, F.: Classifying textual components of bilingual documents with decision-tree support vector machines. In: 2011 International Conference on Document Analysis and Recognition, pp. 498–502. IEEE (2011)
71. Lu, S., Tan, C.L.: Automatic detection of document script and orientation. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 1, pp. 237–241. IEEE (2007)
72. Luqman, H., Mahmoud, S.A., Awaida, S.: Kafd arabic font database. Pattern Recogn. **47**(6), 2231–2240 (2014)
73. Ma, H., Doermann, D.S.: Gabor filter based multi-class classifier for scanned document images. In: ICDAR, vol. 3, p. 968. Citeseer (2003)
74. Mahmoud, S.A., Ahmad, I., Alshayeb, M., Al-Khatib, W.G., Parvez, M.T., Fink, G.A., Märgner, V., El Abed, H.: Khatt: Arabic offline handwritten text database. In: 2012 International Conference on Frontiers in Handwriting Recognition, pp. 449–454. IEEE (2012)
75. Mane, D., Kulkarni, U.: Visualizing and understanding customized convolutional neural network for recognition of handwritten marathi numerals. Proc. Comput. Sci. **132**, 1123–1137 (2018)
76. Manjula, S., Hegadi, R.S.: A review on multilingual document analysis in indian context. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 519–522. IEEE (2016)
77. Manjula, S., Hegadi, R.S.: Identification and classification of multilingual document using maximized mutual information. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 1679–1682. IEEE (2017)
78. Manjula, S., Hegadi, R.S.: Recognition of oriya and english languages based on lbp features. In: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–3. IEEE (2017)
79. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recogn. **5**(1), 39–46 (2002)
80. Mohanty, S., Bebartta, H.D.: A novel approach for bilingual (english-oriya) script identification and recognition in a printed document. IJIP **4**(2), 175 (2010)
81. Morera, Á., Sánchez, Á., Vélez, J.F., Moreno, A.B.: Gender and handedness prediction from offline handwriting using convolutional neural networks. Complexity **2018**, (2018)
82. Mori, S., Suen, C.Y., Yamamoto, K.: Historical review of ocr research and development. Proc. IEEE **80**(7), 1029–1058 (1992)
83. Moussa, S.B., Zahour, A., Benabdelhafid, A., Alimi, A.M.: Fractal-based system for arabic/latin, printed/handwritten script identification. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
84. Namboodiri, A.M., Jain, A.K.: Online script recognition. In: Object recognition supported by user interaction for service robots, vol. 3, pp. 736–739. IEEE (2002)
85. Namboodiri, A.M., Jain, A.K.: Online handwritten script recognition. IEEE Trans. Pattern Anal. Mach. Intell. **26**(1), 124–130 (2004). https://doi.org/10.1109/TPAMI.2004.1261096
86. Nethravathi, B., Archana, C., Shashikiran, K., Ramakrishnan, A.G., Kumar, V.: Creation of a huge annotated database for tamil and kannada ohr. In: 2010 12th International Conference on Frontiers in Handwriting Recognition, pp. 415–420. IEEE (2010)
87. Obaidullah, S.M., Das, N., Halder, C., Roy, K.: Indic script identification from handwritten document images–an unconstrained block-level approach. In: 2015 IEEE 2nd international conference on recent trends in information systems (ReTIS), pp. 213–218. IEEE (2015)
88. Obaidullah, S.M., Das, S.K., Roy, K.: A system for handwritten script identification from indian document. J. Pattern Recogn. Res. **8**(1), 1–12 (2013)
89. Obaidullah, S.M., Goswami, C., Santosh, K., Das, N., Halder, C., Roy, K.: Separating indic scripts with matra for effective handwritten script identification in multi-script documents. Int. J. Pattern Recognit Artif Intell. **31**(05), 1753003 (2017)
90. Obaidullah, S.M., Goswami, C., Santosh, K., Halder, C., Das, N., Roy, K.: Separating indic scripts with 'matra'–a precursor to script identification in multi-script documents. In: Proceedings of International Conference on Computer Vision and Image Processing, pp. 205–214. Springer (2017)
91. Obaidullah, S.M., Halder, C., Das, N., Roy, K.: Numeral script identification from handwritten document images. Proc. Comput. Sci. **54**, 585–594 (2015)

92. Obaidullah, S.M., Halder, C., Das, N., Roy, K.: A corpus of word-level offline handwritten numeral images from official indic scripts. In: Proceedings of the Second International Conference on Computer and Communication Technologies, pp. 703–711. Springer (2016)
93. Obaidullah, S.M., Halder, C., Das, N., Roy, K.: A new dataset of word-level offline handwritten numeral images from four official indic scripts and its benchmarking using image transform fusion. Int. J. Intell. Eng. Inf. **4**(1), 1–20 (2016)
94. Obaidullah, S.M., Halder, C., Das, N., Roy, K.: Pwdb_13: A corpus of word-level printed document images from thirteen official indic scripts. In: Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015, pp. 233–242. Springer (2016)
95. Obaidullah, S.M., Halder, C., Das, N., Roy, K.: Visual analytic-based technique for handwritten indic script identification–a greedy heuristic feature fusion framework. In: Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015, pp. 211–219. Springer (2016)
96. Obaidullah, S.M., Halder, C., Santosh, K., Das, N., Roy, K.: Automatic line-level script identification from handwritten document images-a region-wise classification framework for indian subcontinent. Malays. J. Comput. Sci. **31**(1), 63–84 (2018)
97. Obaidullah, S.M., Halder, C., Santosh, K., Das, N., Roy, K.: Phdindic_11: page-level handwritten document image dataset of 11 official indic scripts for script identification. Multimedia Tools Appl. **77**(2), 1643–1678 (2018)
98. Obaidullah, S.M., Karim, R., Shaikh, S., Halder, C., Das, N., Roy, K.: Transform based approach for indic script identification from handwritten document images. In: 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), pp. 1–7. IEEE (2015)
99. Obaidullah, S.M., Roy, K., Das, N.: 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)
100. Obaidullah, S.M., Santosh, K., Das, N., Halder, C., Roy, K.: Handwritten indic script identification in multi-script document images: a survey. Int. J. Pattern Recognit Artif Intell. **32**(10), 1856012 (2018)
101. Obaidullah, S.M., Santosh, K., Halder, C., Das, N., Roy, K.: Automatic indic script identification from handwritten documents: page, block, line and word-level approach. Int. J. Mach. Learn. Cybernet. **10**(1), 87–106 (2019)
102. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
103. Padma, M., Vijaya, P.: Identification of telugu devanagari and english scripts using discriminating. J. Comput. Sci. **1**, 64–78 (2009)
104. Padma, M., Vijaya, P.: 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)
105. Padma, M., Vijaya, P.: Entropy based texture features useful for automatic script identification. Int. J. Comput. Sci. Eng. **2**(02), 115–120 (2010)
106. Padma, M., Vijaya, P.: Global approach for script identification using wavelet packet based features. Int. J. Signal Process. Image Process. Pattern Recogn. **3**(3), 29–40 (2010)
107. Padma, M., Vijaya, P.: Script identification from trilingual documents using profile based features. IJCSA **7**(4), 16–33 (2010)
108. Padma, M., Vijaya, P.: Wavelet packet based texture features for automatic script identification. Int. J. Image Process **4**(1), 53–65 (2010)
109. Pal, U., Belaıd, A., Choisy, C.: Touching numeral segmentation using water reservoir concept. Pattern Recogn. Lett. **24**(1–3), 261–272 (2003)
110. Pal, U., Chaudhuri, B.: Automatic separation of words in multilingual multi-script indian documents. In: Proceedings of the fourth international conference on document analysis and recognition, vol. 2, pp. 576–579. IEEE (1997)
111. Pal, U., Chaudhuri, B.: 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)
112. Pal, U., Chaudhuri, B.: Identification of different script lines from multi-script documents. Image Vis. Comput. **20**(13–14), 945–954 (2002)
113. Pal, U., Chaudhuri, B.: Script line separation from indian multi-script documents. IETE J. Res. **49**(1), 3–11 (2003)
114. Pal, U., Chaudhuri, B.: Indian script character recognition: a survey. pattern Recognition **37**(9), 1887–1899 (2004)
115. Pal, U., Roy, R.K., Roy, K., Kimura, F.: Indian multi-script full pin-code string recognition for postal automation. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 456–460 (2009). https://doi.org/10.1109/ICDAR.2009.171
116. Pal, U., Sarkar, A.: Recognition of printed urdu script. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 1183–1187. Citeseer (2003)
117. Pal, U., Sharma, N., Wakabayashi, T., Kimura, F.: 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)
118. Pal, U., Sinha, S., Chaudhuri, B.: Multi-script line identification from indian documents. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 880–884. IEEE (2003)
119. Pan, J., Tang, Y.: A rotation-robust script identification based on bemd and lbp. In: 2011 International Conference on Wavelet Analysis and Pattern Recognition, pp. 165–170. IEEE (2011)
120. Pan, W., Suen, C.Y., Bui, T.D.: Script identification using steerable gabor filters. In: Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 883–887. IEEE (2005)
121. Pati, P.B., Raju, S.S., Pati, N., Ramakrishnan, A.: 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)
122. Pati, P.B., Ramakrishnan, A.: Hvs inspired system for script identification in indian multi-script documents. In: International Workshop on Document Analysis Systems, pp. 380–389. Springer (2006)
123. Pati, P.B., Ramakrishnan, A.: Word level multi-script identification. Pattern Recogn. Lett. **29**(9), 1218–1229 (2008)
124. Patil, S.B., Subbareddy, N.: Neural network based system for script identification in indian documents. Sadhana **27**(1), 83–97 (2002)
125. Peake, G., Tan, T.: Script and language identification from document images. In: Proceedings Workshop on Document Image Analysis (DIA'97), pp. 10–17. IEEE (1997)
126. Peng, L., Liu, C., Ding, X., Wang, H.: Multilingual document recognition research and its application in china. In: Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp. 7–pp. IEEE (2006)
127. Phan, T.Q., Shivakumara, P., Ding, Z., Lu, S., Tan, C.L.: Video script identification based on text lines. In: 2011 International Conference on Document Analysis and Recognition, pp. 1240–1244. IEEE (2011)
128. Philip, B., Samuel, R.S.: A novel bilingual ocr for printed malayalam-english text based on gabor features and dominant sin-

gular values. In: 2009 International Conference on Digital Image Processing, pp. 361–365. IEEE (2009)
129. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification-the state of the art. Pattern Recogn. **22**(2), 107–131 (1989)
130. Rabby, A.S.A., Haque, S., Islam, S., Abujar, S., Hossain, S.A.: Bornonet: Bangla handwritten characters recognition using convolutional neural network. Proc. Comput. Sci. **143**, 528–535 (2018)
131. Raghunandan, K., Shivakumara, P., Roy, S., Kumar, G.H., Pal, U., Lu, T.: Multi-script-oriented text detection and recognition in video/scene/born digital images. IEEE Trans. Circuits Syst. Video Technol. **29**(4), 1145–1162 (2018)
132. Rai, H., Yadav, A.: Iris recognition using combined support vector machine and hamming distance approach. Expert Syst. Appl. **41**(2), 588–593 (2014)
133. Rajput, G., Anita, H.: Handwritten script recognition at line level-a multiple feature based approach. Int. J. Eng. Innov. Technol. **3**(4), 90–95 (2013)
134. Ramteke, A.S., Rane, M.E.: A survey on offline recognition of handwritten devanagari script. Int. J. Sci. Eng. Res. **3**(5), (2012)
135. Rani, R., Dhir, R., Lehal, G.S.: Performance analysis of feature extractors and classifiers for script recognition of english and gurmukhi words. In: Proceeding of the workshop on Document Analysis and Recognition, pp. 30–36 (2012)
136. Rani, R., Dhir, R., Lehal, G.S.: 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)
137. Rao, G.S., Imanuddin, M., Harikumar, B.: Script identification of telugu, english and hindi document image. Int. J. Adv. Eng. Global Technol **2**(2), 443–452 (2014)
138. Razzak, M.I., Hussain, S., Sher, M.: Numeral recognition for urdu script in unconstrained environment. In: 2009 International Conference on Emerging Technologies, pp. 44–47. IEEE (2009)
139. Rezaee, H., Geravanchizadeh, M., Razzazi, F.: Automatic language identification of bilingual english and farsi scripts. In: 2009 International Conference on Application of Information and Communication Technologies, pp. 1–4. IEEE (2009)
140. Roy, K., Alaei, A., Pal, U.: Word-wise handwritten persian and roman script identification. In: 2010 12th International Conference on Frontiers in Handwriting Recognition, pp. 628–633. IEEE (2010)
141. Roy, K., Banerjee, A., Pal, U.: A system for word-wise handwritten script identification for indian postal automation. In: Proceedings of the IEEE INDICON 2004. First India Annual Conference, 2004., pp. 266–271 (2004). https://doi.org/10.1109/INDICO.2004.1497753
142. Roy, K., Das, S.K., Obaidullah, S.M.: Script identification from handwritten document. In: 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 66–69. IEEE (2011)
143. Roy, K., Majumder, K.: Trilingual script separation of handwritten postal document. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 693–700. IEEE (2008)
144. Roy, K., Pal, U., Chaudhuri, B.: 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)
145. Roy, P.P.: Center for visual information technology (cvit) - international institute of information technology, gachibowli, hyderabad. https://cvit.iiit.ac.in/research/resources
146. Roy, P.P.: Pattern recognition, image processing and machine learning (parimal) iit roorkee. http://parimal.iitr.ac.in/dataset
147. Saïdani, A., Echi, A.K., Belaid, A.: Identification of machine-printed and handwritten words in arabic and latin scripts. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 798–802. IEEE (2013)
148. Saidani, A., Kacem, A., Belaid, A.: Co-occurrence matrix of oriented gradients for word script and nature identification. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 16–20. IEEE (2015)
149. Samanta, O., Roy, A., Parui, S.K., Bhattacharya, U.: An hmm framework based on spherical-linear features for online cursive handwriting recognition. Inf. Sci. **441**, 133–151 (2018)
150. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D.K.: Word level script identification from bangla and devanagri handwritten texts mixed with roman script. arXiv preprint arXiv:1002.4007 (2010)
151. Sharma, M.K., Dhaka, V.P.: Offline scripting-free author identification based on speeded-up robust features. International Journal on Document Analysis and Recognition (IJDAR) **18**(4), 303–316 (2015)
152. Sharma, M.K., Dhaka, V.P.: Pixel plot and trace based segmentation method for bilingual handwritten scripts using feedforward neural network. Neural Comput. Appl. **27**(7), 1817–1829 (2016)
153. Sharma, M.K., Dhaka, V.P.: Segmentation of english offline handwritten cursive scripts using a feedforward neural network. Neural Comput. Appl. **27**(5), 1369–1379 (2016)
154. Sharma, N., Chanda, S., Pal, U., Blumenstein, M.: Word-wise script identification from video frames. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 867–871 (2013). https://doi.org/10.1109/ICDAR.2013.177
155. Sharma, N., Mandal, R., Sharma, R., Pal, U., Blumenstein, M.: Bag-of-visual words for word-wise video script identification: A study. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2015)
156. Sharma, N., Pal, U., Blumenstein, M.: A study on word-level multi-script identification from video frames. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1827–1833. IEEE (2014)
157. Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., Tan, C.L.: A new method for word segmentation from arbitrarily-oriented video text lines. In: 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 1–8. IEEE (2012)
158. Shi, B., Bai, X., Yao, C.: Script identification in the wild via discriminative convolutional neural network. Pattern Recogn. **52**, 448–458 (2016)
159. Shi, B., Yao, C., Zhang, C., Guo, X., Huang, F., Bai, X.: Automatic script identification in the wild. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 531–535. IEEE (2015)
160. Shivakumara, P., Sharma, N., Pal, U., Blumenstein, M., Tan, C.L.: Gradient-angular-features for word-wise video script identification. In: 2014 22nd International Conference on Pattern Recognition, pp. 3098–3103. IEEE (2014)
161. Shivakumara, P., Yuan, Z., Zhao, D., Lu, T., Tan, C.L.: New gradient-spatial-structural features for video script identification. Comput. Vis. Image Underst. **130**, 35–53 (2015)
162. Singh, M.P., Dhaka, V.: Handwritten character recognition using modified gradient descent technique of neural networks and representation of conjugate descent for training patterns. International Journal of Engineering pp. 145–158 (2009)
163. Singh, P.K., Chatterjee, I., Sarkar, R.: Page-level handwritten script identification using modified log-gabor filter based features. In: 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 225–230. IEEE (2015)

164. Singh, P.K., Sarkar, R., Nasipuri, M.: Offline script identification from multilingual indic-script documents: a state-of-the-art. Computer Science Review **15**, 1–28 (2015)
165. Singhal, V., Navin, N., Ghosh, D.: 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)
166. Sinha, S., Pal, U., Chaudhuri, B.: Word–wise script identification from indian documents. In: International Workshop on Document Analysis Systems, pp. 310–321. Springer (2004)
167. Sudholt, S., Fink, G.A.: Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 277–282. IEEE (2016)
168. Thadchanamoorthy, S., Kodikara, N., Premaretne, H., Pal, U., Kimura, F.: Tamil handwritten city name database development and recognition for postal automation. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 793–797. IEEE (2013)
169. Tsai, M.J., Tao, Y.H., Yuadi, I.: Deep learning for printed document source identification. Sig. Process. Image Commun. **70**, 184–198 (2019)
170. Ubul, K., Tursun, G., Aysa, A., Impedovo, D., Pirlo, G., Yibulayin, T.: Script identification of multi-script documents: a survey. IEEE Access **5**, 6546–6559 (2017)
171. Ukil, S., Ghosh, S., Obaidullah, S.M., Santosh, K., Roy, K., Das, N.: Deep learning for word-level handwritten indic script identification. arXiv preprint arXiv:1801.01627 (2018)
172. Wang, X.Y., Wang, Q.Y., Yang, H.Y., Bu, J.: Color image segmentation using automatic pixel classification with support vector machine. Neurocomputing **74**(18), 3898–3911 (2011)
173. Xing, L., Qiao, Y.: Deepwriter: A multi-stream deep cnn for text-independent writer identification. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 584–589. IEEE (2016)
174. Zheng, Y., Iwana, B.K., Uchida, S.: Mining the displacement of max-pooling for text recognition. Pattern Recogn. **93**, 558–569 (2019)