Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-019-09332-0

A Study on Recognition of Pre-segmented Handwritten Multi-lingual Characters

Munish Kumar[1] · **Simpel Rani Jindal**[2]

Received: 9 September 2018 / Accepted: 18 February 2019

Abstract
Wide research has been carried out for recognition of handwritten text on various languages that include Assamese, Bangla, English, Gujarati, Hindi, Marathi, Punjabi, Tamil etc. Recognition of multi-lingual text documents is still a challenge in the pattern recognition field. In this paper, a study of various features and classifiers for recognition of pre-segmented multi-lingual characters consisting of English, Hindi and Punjabi has been presented. In feature extraction phase, various techniques, namely, zoning features, diagonal features, horizontal peak extent based features and intersection and open end point based features are considered. In classification phase, three different classifiers, namely, k-NN, Linear-SVM, and MLP are attempted. Different combinations of various features and classifiers have been also performed. For script identification, we have achieved maximum accuracy of 92.89% using a combination of Linear-SVM, k-NN, and MLP classifiers, and for character recognition of English, Hindi and Punjabi, we have achieved a recognition accuracy of 92.18%, 84.67% and 86.79%, respectively.

1 Introduction

The field of pattern recognition contributed up to a great extent in the machine vision applications. Handwriting recognition is a part of the area under pattern recognition community. Handwriting recognition is a technique or ability of a computer to receive and interpret intelligible handwritten text input from source such as paper documents, touch screen, photographs, etc. In general, handwriting recognition is classified into two types, namely, online recognition and offline recognition. Online handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or personal digital assistant (PDA), where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. That kind of data can be regarded as a dynamic representation of handwriting. Offline handwriting recognition involves the automatic conversion of text in an image into letter codes which are useable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. The process of converting textual symbols on a paper to a machine process-able format is known as optical character recognition (OCR) which is the core of the field of document analysis system (DAS). Thus, it plays an important role in transformation of paper based society to paperless electronic information society. The proposed paper presents a system for multi-lingual character recognition consisting of English, Hindi and Punjabi text. Punjabi is an Indo-Aryan language spoken by about 130 million people mainly in West Punjab in Pakistan and in East Punjab in India. Punjabi is one of India's 22 official languages and it is the first official language in East Punjab. In India, Punjabi is written with the Gurmukhi script. The word Gurmukhi has been commonly translated as "from the mouth of Guru", and contains thirty-five letters. Hindi is an Indo-Aryan language used in the northern states of Rajasthan, Delhi, Haryana, Uttarakhand, Uttar Pradesh, Madhya Pradesh, Chhattisgarh, Himachal Pradesh, Jharkhand and Bihar. Hindi, written in Devanagari script, is one of the official languages of India. The Devanagari script is used for over 120 languages and contains forty-seven primary characters, of which fourteen are vowels and thirty-three are consonants. English is the

✉ Munish Kumar
munishcse@gmail.com

Simpel Rani Jindal
simpel_jindal@rediffmail.com

[1] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

[2] Computer Science and Engineering Section, Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India

Published online: 08 March 2019

third most widely used language in the world, behind Mandarin and Spanish. Latin script is used for writing English language. Latin or Roman script is the basis of International Phonetic Alphabet and contains 26 letters. Optical Character Recognition (OCR) provides a large number of applications like OCR can be used by organizations for automated form processing in places where a large number of data is available in printed form. Other uses of OCR include processing utility bills, passport validation, mail sorting, bank cheque reading, signature verification, pen computing, etc. Bag and Harit [1] have presented a survey on optical character recognition for Bangla and Devanagari scripts. In this article, they have reported various works on OCR in two major Indian scripts—Bangla and Devanagari. They have reported the research trends and compared the techniques being used in the modern OCR systems. Govindaraju and Setlur [2] have presented a comprehensive book on the topics of Indic Script OCRs. In this book, they have presented all major research groups working in this area. They discussed about dataset creation for OCR development and describes about OCR systems that cover eight different scripts: Bangla, Devanagari, Gurmukhi, Gujarati, Kannada, Malayalam, Tamil, and Urdu (Perso-Arabic). They have also explored the various challenges of Indic script handwriting recognition and examine the development of handwriting-based text input systems. But, presently no recognition system is available for segmentation free handwritten multi-lingual Text. So, in this paper, we have presented a recognition system for multi-lingual text consisting of English, Hindi and Punjabi text.

2 Motivations

It is not possible to save the historical documents, writer's books for many years in the original format. But, once it is digitized, then it's very easy to use such documents for the generation to generations. Because of the improvement in the technology of the past few decades, the older historical documents are stored in the digitized form. Hence, it can be helpful for the future generation for extracting, modifying and storing the data. There are many local languages and written scripts in the India (22 languages and 12 scripts). Hence, there is a clear need for commercial offline handwriting recognition engines in local scripts such as Bengali, Devanagari, Gurumukhi, Tamil, Telugu, etc. for providing services to a local-language-literate population. This could empower Indian enterprises to start deploying forms-processing technologies with handwriting in local scripts. As per the tri-scripts principle of Indian constitution, every state Government has to produce an official document containing a national script (Devanagari), official script (Roman) and the state script (or regional script). For example, an official document of Punjab state contains Devanagari, Roman and Gurmukhi scripts. Moreover, OCR is useful in helping blind and visually impaired people to read text. So there is a need to develop efficient and reliable system for recognition of multilingual document i.e. a document containing text in many languages.

3 Pros and Cons of OCR

Pros:

- OCR provides fast, automated data capture which can save considerable time and labor costs of organizations.
- Documents can become editable with OCR. We can convert the files to any editable digital formats.
- It increases the efficiency and effectiveness of office work.
- It provides facility of copy and paste tools on the document instead of rewriting everything to correct it.
- When combined with other technologies such as scanning and file compression, the advantages of OCR truly shine.
- Workflow is increased since employees no longer have to waste time on manual labor and can work quicker and more efficiently.
- OCR technology can significantly improve the efficiency of a bank's indexing process.

Cons:

- No OCR software is 100% accurate. The number of errors depends upon the quality and type of document, including the font used.
- Handwritten documents cannot be easily read by OCR software.
- OCR has difficulty differentiating between characters, such as the number zero and a capital "O".
- Even if the scanned image of the original document is high-quality, additional steps must occur to clean up the OCR text.
- Most document formatting is lost during text scanning, except for paragraph marks and tab stops.

4 Challenges

India is a multilingual country. More than one language is used in a document frequently. For example, various forms in banks use three languages i.e. state language, national language and international language. Recognition of more

Table 1 Similar shaped characters in English language

E and F	I and J
U and V	Y and X
I and T	H and N
c and d	c and e
u and v	m and n

Table 2 Similar shaped characters in Punjabi language

ਬ and ਥ	ਤ and ਭ	ਤ and ੜ	ਏ and ਟ
ਠ and ਨ	ਸ and ਮ	ਸ and ਗ	ਖ and ਪ
ਚ and ਜ	ਚ and ਦ	ਫ and ਢ	ਰ and ਹ
ਏ and ਲ	ਦ and ਢ	ਖ and ਥ	ਟ and ਠ

than one language at a time is a great challenge. There are

Table 3 Similar shaped characters in Hindi language

ख and र	ख and श	ख and स
घ and ध	य and थ	व and ब
भ and म	ज and ञ	प and ष
त and ल	ट and ढ	ङ and ड
ढ and ढ	ट and ठ	प and य

Table 4 Characters with similar shapes in Hindi and Punjabi language

ਕ, ख, ਰ, श स	ਖ, ਪ, ਧ, ਥ	ਹ, ਰ, ਗ
ਸ, ਮ, ਯ, ਮ	ਬ, ਕ, ਕ	ਟ, ਠ, ਨ
ਜ, ਦ, ਚ, ਢ	ਧ, ध, ਥ	ੜ, ਤ, ਭ
ਏ, ਟ, ਫ, ਢ	ਛ, ਭ, ङ	ਢ, ਟ

various challenges faced during multilingual text recognition which are listed as below:

i. Some characters have similar shape with the same meaning in multiple languages. For example, ਗ, ਟ, ਠ, ਪ, ਛ. These characters exist in both Hindi and Punjabi language and it is difficult to identify the script of these characters before their recognition from multilingual text document.
ii. Similarity in the shape of different characters makes confusion during the recognition process of multilingual text document, like ਟ and ਠ (Tables 1, 2, 3, 4).
iii. Extraction and selection of appropriate/efficient feature extraction techniques for multi-lingual text recognition.
iv. Low quality of text images can also be the cause of low recognition rate.
v. Offline handwriting recognition is comparatively difficult; as different people have different writing styles.
vi. Recognition strategies heavily depends on the nature of the data to be recognized. In the cursive case, the problem is made complex by the fact that the writing is fundamentally ambiguous as the letters in the word are generally linked together, poorly written and may be missing.
vii. Large number of character set present in a particular language makes it an open problem for researchers.
viii. The artifacts of the complex interactions between hardware equipment and subsequent operations such as scanning and binarization present additional challenges to algorithms for offline handwritten character recognition.

5 Background

In the history of character recognition field, a number of techniques are available for isolated character recognition of the particular script like Bangla, Devanagari, Gurmukhi, Kannada, etc. But research for multilingual text recognition is not so wide. In this section, work related to character recognition has been considered. For example, Pandey et al. [3] have proposed an approach to handwritten script recognition using soft computing, which emphasis on the block-level technique. They have used combined approach of discrete cosines transforms (DCT) and discrete wavelets transform (DWT) for feature extraction and neural network for classification task. The proposed approach obtained 82.70% accuracy on database of 961 handwritten samples in three scripts, namely, Hindi, English and Urdu. Rani et al. [4] have presented a system for script identification at character level that consists of English and Punjabi characters and digits by using SVM classifier with its different kernels. They have experimented with multi-font and multi-sized character using Gabor feature based on directional frequency and gradient feature based on gradient information. They obtained average identification rates of 98.9% and 99.45% with Gabor and Gradient features, respectively. A system for recognition of multi-lingual real-world offline handwritten images has been proposed by Kozielski et al. [5]. They have described methods for image denoising, line removal, deskewing, deslanting and text line segmentation. Features are reduced using PCA. The system has been trained using HMM and LSTM recurrent neural network to recognize text written in French and English. The proposed system outperformed other approaches and scored the first place for Latin script recognition. Kaur and Mahajan [6] have presented an identification system of English and Punjabi script at line level through headline and character density feature. The proposed system achieved script identification rates of 91.8% and 89.7% for English and Punjabi, respectively. Surinta et al. [7] have

used local gradient feature descriptors for recognition of handwritten characters. They evaluated the feature descriptors and two classifiers, kNN, SVM on three different scripts, namely, Thai, Bangla and Latin. The results reveal that SVM in combination with proposed feature descriptors obtained high accuracy. Chakraborty and Pal [8] proposed a baseline detection scheme for unconstrained handwritten text lines of multilingual documents. Experiments are conducted on six scripts, namely, Bengali, Roman, Kannada, Oriya, Devanagari and Persian scripts using SVM as classifier trained with the orientation invariant features. In case of multi-oriented text lines, the proposed method generates 14.8% errors as compared to the method proposed by Morillot et al. [9] which gives 45.3% errors. To the best of authors' knowledge, this is the first work to use machine learning directly for baseline detection. A novel fuzzy approach to segment touching characters has been proposed by Farulla et al. [10] that combines three state-of-the-art features, namely, the peak-to-valley function, the distance from the center of the pattern and the ratio of the second difference of the vertical projection. Experiments are evaluated on two different datasets composed by Latin handwritten characters (dataset A) and printed characters (dataset B). The proposed approach achieved recognition accuracy of 88.9% and 96.1% on datasets A and B respectively. Mandal et al. [11] have proposed Gaussian Mixture Model (GMM) posterior features for online handwriting recognition. Kumar et al. [12] have presented a systematic survey for offline handwritten character recognition of various Indic and Non-Indic scripts. They have also presented various issues and challenges for offline handwritten and numeral recognition. Three databases, namely, UNIPEN English character database, UNIPEN ICROW-03 English word database and locally collected Assamese digit database have been used to evaluate the proposed approach using SVM classifier.

6 Multi-lingual Text Recognition System

In a multilingual country like India, a document may contain text words in more than one language. For a multilingual environment in order to reach a larger cross section of people, it is necessary that a document should be composed of text contents in different languages. But, on the other hand, this causes practical difficulty in optical character recognition such a document, because the language type of the text should be pre-determined, before employing a particular optical character recognition system. It is perhaps impossible to design a single recognizer which can identify a large number of scripts/languages. So, it is essential to identify the language region of the document before feeding the document to the corresponding optical character recognition system. The multi-lingual text recognition system consists of various stages as shown in Fig. 1. These stages are briefly discussed in following subsections. Digitization and pre-processing phases are used for converting the paper based documents into digitized format and thinned image format. After that, features are being extracted. Depending on the features extracted in the previous phase, the script has been identified. Finally, the character has been recognized by using different classification techniques.

6.1 Digitization and Pre-processing

Digitization is an initial phase of a recognition system. In this phase, input data is converted into electronic form. In this process, hard copied document gets converted into soft copied image in the system. After digitization, digitized image is passed to its following pre-processing phase. Pre-processing includes noise removal, skew detection/correction and skeltonization. It helps in detecting and removing all unwanted bit pattern which may lead to reduce the recognition accuracy. After pre-processing of text, various features have been extracted using different techniques for recognition purpose.

6.2 Feature Extraction

Feature extraction is the most important phase of a recognition system. We have used statistical feature extraction techniques, namely, zoning features, diagonal features, horizontal peak extent features and intersection and open end points based features. We have extracted various features for character recognition but we have noticed that these four techniques are performing better than other techniques so in his papers, these four feature extraction techniques are considered. We have also attempted to use the combination of these techniques to improve the recognition accuracy of the proposed system.

- Zoning features

Zoning is a technique for region-based feature extraction. Zoning features are the most popular and efficient statistical features that provide high speed and low complexity of character and word recognition. In zoning features, the image is divided into a number of zones as shown in Fig. 2 and particular features are extracted from each zone. These are calculated from the density of pixels, or pattern characteristics in several zones. The number of foreground pixels in each zone is calculated. These numbers $p_1, p_2, \ldots, p_n$, obtained for all n zones are normalized to [0, 1] resulting into a feature set of n elements. The goal of zoning is to

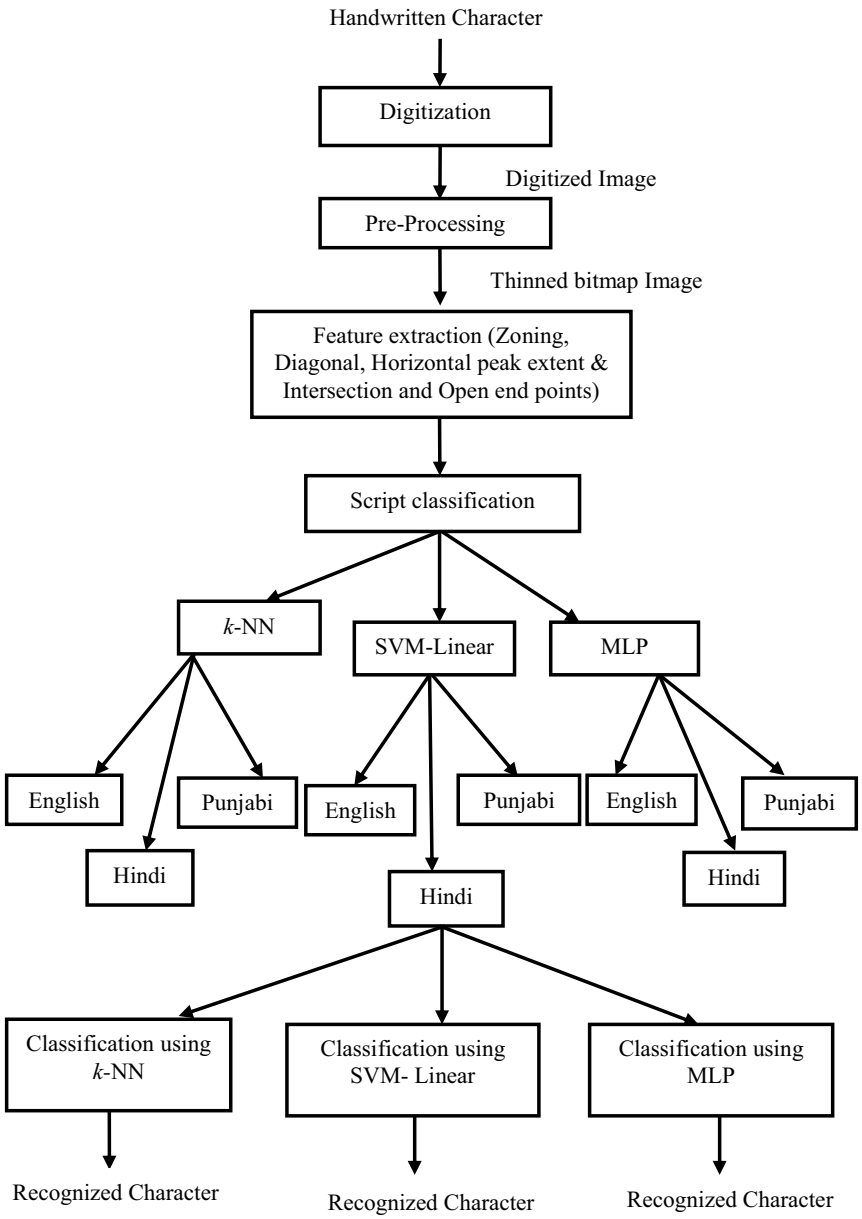

Fig. 1 Block diagram of multi-lingual text recognition system

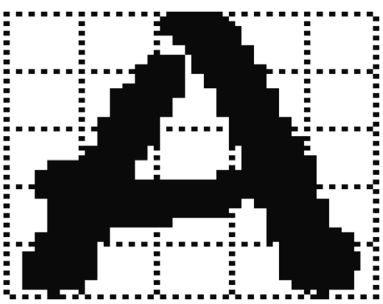

Fig. 2 Zoning of a character image

obtain local characteristics instead of global characteristics. Zoning has been successfully applied to character recognition since it has been found that region-based approaches for feature extraction can absorb variability of handwritten patterns, that derives from different writing styles and changeable writing conditions.

- Diagonal features

In character recognition, diagonal features are very important features to achieve higher recognition accuracy and reducing misclassification. In this method, the image of the character is divided into n number of zones and features of the image are extracted from the pixels by moving

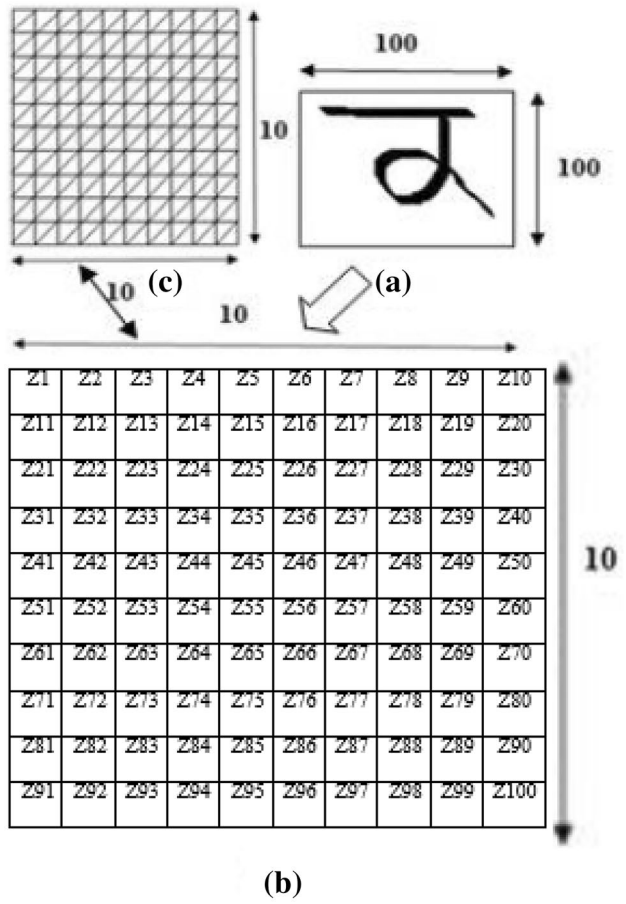

Fig. 3 **a** Normalized character image. **b** Image divided into 100 zones. **c** Diagonal feature extraction in a zone each of size 10×10 pixels

along its diagonal. Each zone consists of (2n−1) diagonals, whose values are averaged to get a single value as a feature value of feature vector for complete image.

As shown in Fig. 3, consider the computation of Diagonal Features for each character image of size 100 × 100 pixels having 10 × 10 zones and thus each zone having 10 × 10-pixel size. Each of these zones are having 19 diagonals. The number of foreground pixels along each diagonal are summed up to get 19 features from each zone, then these features for each zone are averaged to extract a single feature from each zone.

- Horizontal peak extent features

In this technique, sum of successive foreground pixel extents in the horizontal direction in each row are considered. Peak values in each row of the zone are summed up to calculate the corresponding feature value for a zone of an image of a character. Following steps are used to extract the horizontal peak extent based features:

- The input image is divided into *n* number of divisions, each containing $m \times m$ pixels.
- Calculate the sum of successive foreground pixel extents in horizontal direction in each row of division.

Then check the largest value in each row and swap it with each foreground pixel in row.

- Intersection and open end point features

The Intersection point is a point that contains more than one pixel in its neighbour and an open end point that has only one point in its neighbour. Following steps have been used to extract these features:

- Divide the image into n number of zones.
- Calculate the number of intersection and open end point for each zone of a character image.

6.3 Script Identification

Script identification is a process in multi-lingual text recognition that helps in identifying the particular language from the multi-lingual text document. It is the process of classifying the script, so that the document can be recognized easily. Script identification process has been shown in Fig. 4.

6.4 Classification

Classification is a decision making phase of a recognition system. Classification phase uses the features, extracted in the feature extraction phase, which is used for classification purpose. It is the process of grouping the similar objects in a single class. In this work, we have considered *k*-NN and SVM classification techniques.

- k-NN classifier

k-NN is considered a lazy learning algorithm that classifies data sets based on their similarity with neighbours. Here k stands for the number of data set items that are considered for the classification. A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour. Usually Euclidean distance is used for calculating

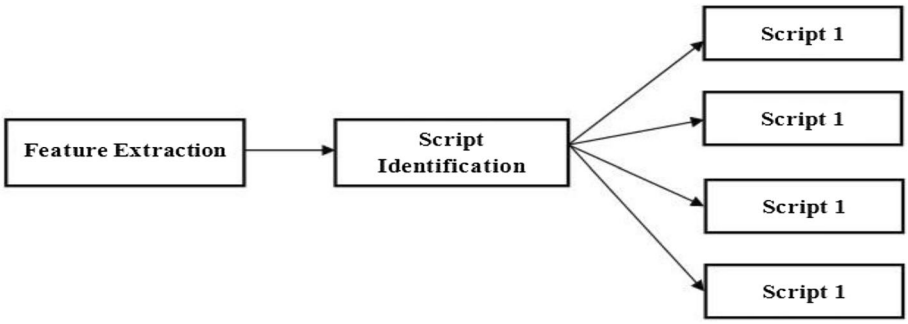

Fig. 4 Script identification

the distance between stored feature vector and candidate feature vector in k-Nearest Neighbour algorithm. Distance between new vector and stored vector can be calculated as:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Here n is the number of feature values in the feature vector, x_i is the library stored feature vector and y_i is the candidate feature vector. For the given attributes A = {X1, X2, … , XD} Where D is the dimension of the data, we need to predict the corresponding classification group G = {Y1, Y2, … , Yn} using the proximity metric over K items in D dimension that defines the closeness of association such that X € R^D and Yp € G. Proximity metric also termed as "Similarity Measure" quantifies the association among different items. Choosing the optimal value for K is best done by first inspecting the data. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. In this work, we have considered $k = 1$ in k-NN classifier.

- SVM classifier

SVM is the most widely used supervised learning classifier in the pattern recognition field. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships as shown in Fig. 5. As shown in figure, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object falling to the right is labeled, i.e., classified, as GREEN (or classified as RED if it falls to the left of the separating line).

The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective

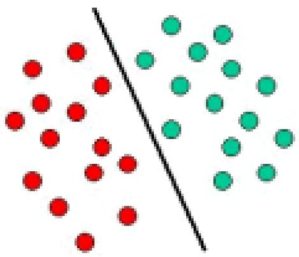

Fig. 5 Linear-SVM classification. (Color figure online)

groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). In such cases, a full separation of the GREEN and RED objects would require a curve which is more complex than a line. Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks. SVM is available with its four types of kernels, namely, linear kernel, polynomial kernel, RBF kernel and sigmoid kernel. In this paper, we have considered linear-SVM classifier for multi-lingual text recognition using C-SVC lib-SVM tool.

- MLP classifier

Multilayer perceptron classifier (MLPC) is a classifier based on the feedforward artificial neural network. MLPC consists of multiple layers of nodes. Each layer is fully connected to the next layer in the network. Nodes in the input layer represent the input data. All other nodes maps inputs to the outputs by performing linear combination of the inputs with the node's weights w and bias b and applying an activation function. It can be written in matrix form for MLPC with K + 1 layers as follows:

$$y(x) = f_K\left(\ldots f_2\left(w_2^T f_1\left(w_1^T x + b_1\right) + b_2\right) \cdots + b_K\right)$$

Nodes in intermediate layers use sigmoid (logistic) function:

$$f(zi) = 1/\left(1 + e^{-zi}\right)$$

Nodes in the output layer use softmax function:

$$f(zi) = e^{zi}/\left(\sum_{k=1}^{N} e^{zk}\right)$$

The number of nodes N in the output layer corresponds to the number of classes. MLPC employs backpropagation for learning the model. Optimization of the network configuration Pruning describes a set of techniques to trim network size (by nodes not layers) to improve computational performance and sometimes resolution performance. The gist of these techniques is removing nodes from the network during training by identifying those nodes which, if removed from the network, would not noticeably affect network performance (i.e., resolution of the data). Obviously, if you use a pruning algorithm during training then begin with a network configuration that is more likely to have excess (i.e., 'prunable') nodes. By applying a pruning algorithm to your network during training, you can approach optimal network configuration; whether you can do that in a single "up-front" such as a genetic-algorithm-based algorithm. In the present work, Back Propagation (BP) learning algorithm bookmark 37 with learning rate (y) = 0.3 and momentum term (a) = 0.2 is used for the training of these MLPs based classifiers.

7 Database Collection

In Fig. 6, a real life example for multi-lingual text document has been shown. It consists of numeric data, Punjabi text and English alphabets. This document needs to be classified in order to identify the language of text before its recognition. A number of other data sources are available for public where multi-lingual data exists. For recognition of such data, we need multi-lingual text recognition system, which is the main goal of this paper. In this work, we have considered three different languages, namely, English, Hindi and Punjabi. We have collected a dataset from 40 different writers. Each writer is requested to write English, Hindi and Punjabi characters separately. We have considered lower case and upper case characters of the English language as different classes. Therefore, we consist of 52-class problem

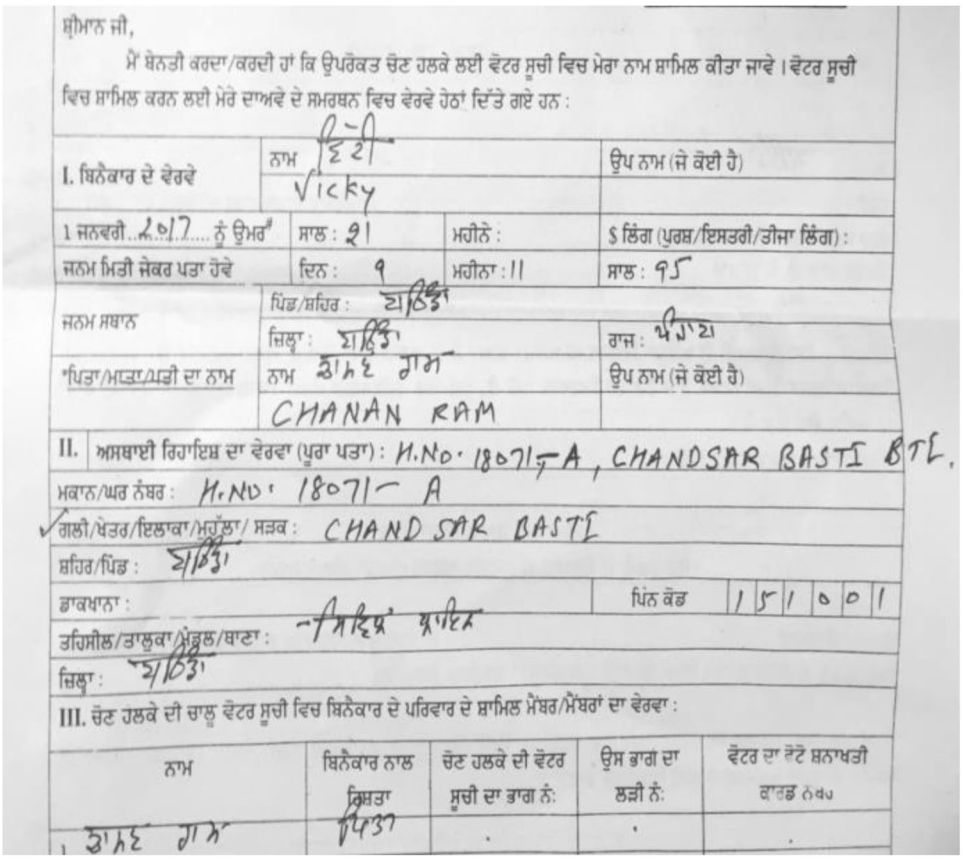
ਸ਼ੁਮਾਨ ਜੀ,
ਮੈਂ ਬੇਨਤੀ ਕਰਦਾ/ਕਰਦੀ ਹਾਂ ਕਿ ਉਪਰੋਕਤ ਚੋਣ ਹਲਕੇ ਲਈ ਵੋਟਰ ਸੂਚੀ ਵਿਚ ਮੇਰਾ ਨਾਮ ਸ਼ਾਮਿਲ ਕੀਤਾ ਜਾਵੇ। ਵੋਟਰ ਸੂਚੀ ਵਿਚ ਸ਼ਾਮਿਲ ਕਰਨ ਲਈ ਮੇਰੇ ਦਾਅਵੇ ਦੇ ਸਮਰਥਨ ਵਿਚ ਵੇਰਵੇ ਹੇਠਾਂ ਦਿੱਤੇ ਗਏ ਹਨ :
I. ਬਿਨੈਕਾਰ ਦੇ ਵੇਰਵੇ — ਨਾਮ: Vicky — ਉਪ ਨਾਮ (ਜੇ ਕੋਈ ਹੈ)
1 ਜਨਵਰੀ 2017 ਨੂੰ ਉਮਰ — ਸਾਲ : 21 — ਮਹੀਨੇ : — ਲਿੰਗ (ਪੁਰਸ਼/ਇਸਤਰੀ/ਤੀਜਾ ਲਿੰਗ) :
ਜਨਮ ਮਿਤੀ ਜੇਕਰ ਪਤਾ ਹੋਵੇ — ਦਿਨ : 9 — ਮਹੀਨਾ : 11 — ਸਾਲ : 95
ਜਨਮ ਸਥਾਨ — ਪਿੰਡ/ਸ਼ਹਿਰ : ਫਿਰੋਜ਼ਪੁਰ — ਜ਼ਿਲ੍ਹਾ : ਫਿਰੋਜ਼ਪੁਰ — ਰਾਜ : ਪੰਜਾਬ
*ਪਿਤਾ/ਮਹਤਾ/ਪਤੀ ਦਾ ਨਾਮ — ਨਾਮ ਚਾਨਣ ਰਾਮ — ਉਪ ਨਾਮ (ਜੇ ਕੋਈ ਹੈ)
CHANAN RAM
II. ਅਸਥਾਈ ਰਿਹਾਇਸ਼ ਦਾ ਵੇਰਵਾ (ਪੂਰਾ ਪਤਾ) : H.No. 18071-A, CHANDSAR BASTI BTL.
ਮਕਾਨ/ਘਰ ਨੰਬਰ : H.No. 18071-A
ਗਲੀ/ਖੇਤਰ/ਇਲਾਕਾ/ਮਹੁੱਲਾ/ ਸੜਕ : CHAND SAR BASTI
ਸ਼ਹਿਰ/ਪਿੰਡ : ਫਿਰੋਜ਼ਪੁਰ
ਡਾਕਖਾਨਾ : — ਪਿੰਨ ਕੋਡ 152001
ਤਹਿਸੀਲ/ਤਾਲੁਕਾ/ਮੰਡਲ/ਥਾਣਾ : ਫਿਰੋਜ਼ਪੁਰ ਸ਼ਹਿਰ
ਜ਼ਿਲ੍ਹਾ : ਫਿਰੋਜ਼ਪੁਰ
III. ਚੋਣ ਹਲਕੇ ਦੀ ਚਾਲੂ ਵੋਟਰ ਸੂਚੀ ਵਿਚ ਬਿਨੈਕਾਰ ਦੇ ਪਰਿਵਾਰ ਦੇ ਸ਼ਾਮਿਲ ਮੈਂਬਰ/ਮੈਂਬਰਾਂ ਦਾ ਵੇਰਵਾ :
ਨਾਮ	ਬਿਨੈਕਾਰ ਨਾਲ ਰਿਸ਼ਤਾ	ਚੋਣ ਹਲਕੇ ਦੀ ਵੋਟਰ ਸੂਚੀ ਦਾ ਭਾਗ ਨੰ:	ਉਸ ਭਾਗ ਦਾ ਲੜੀ ਨੰ:	ਵੋਟਰ ਦਾ ਵੋਟਰ ਸ਼ਨਾਖਤੀ ਕਾਰਡ ਨੰ
1 ਚਾਨਣ ਰਾਮ	ਪਿਤਾ			

Fig. 6 A sample of multi-lingual text document

in English language (26 for lower case and 26 for upper case) whereas Hindi and Punjabi contains 36-class and 35 class problems, respectively. We have 2080 English characters (1040 lower case and 1040 upper case), 1440 Hindi characters and 1400 Punjabi characters. Total 4920 samples of pre-segmented characters written in English, Hindi and Punjabi are collected.

8 Experimental Results

Various experimental results for script identification and character recognition have been presented in this section. We have extracted four types of features, namely, zoning feature, diagonal feature, horizontal peak extent feature and intersection and open end point features and three classifiers as discussed in Sect. 3. We have also used different combinations of these techniques. For computing the experimental results, 4920 samples of isolated multi-lingual characters written by 40 different writers are considered in this work. For script identification, maximum accuracy of 92.89% has been achieved using a combination of Linear-SVM, k-NN, and MLP classifiers as shown in Table 5. Script identification results are graphically depicted in Fig. 7. For developing a training set and a testing set for each of the scripts, employed for this work, the relevant dataset is segmented into a ratio of 70:30. 70% data is taken training dataset and 30% data is taken as testing dataset.

In Table 6, we have depicted recognition results of English text recognition. We have achieved an accuracy of 92.18% for English text by using the combination of zoning, diagonal and intersection and open end point

Table 5 Script identification results using various feature extraction techniques and classifier

Feature extraction techniques	Feature vector length	Classification techniques						
		Linear-SVM (C_1) (%)	k-NN (C_2) (%)	MLP (C_3) (%)	C_1+C_2 (%)	C_1+C_3 (%)	C_2+C_3 (%)	$C_1+C_2+C_3$ (%)
Zoning (F_1)	64	74.79	76.19	85.17	81.29	82.81	92.57	88.36
Diagonal (F_2)	64	74.77	75.97	85.42	80.75	82.57	92.85	87.77
Intersection (F_3)	128	80.79	79.65	85.40	87.82	86.57	92.82	91.46
Peak extent (F_4)	64	73.25	68.55	76.84	79.67	74.51	83.52	86.60
F_1+F_2	128	78.21	79.57	87.52	85.14	86.49	91.28	88.54
F_1+F_3	192	84.53	83.31	89.22	89.92	90.55	89.16	91.48
F_1+F_4	128	79.18	77.76	85.71	87.12	84.52	89.16	90.69
F_2+F_3	192	82.88	80.95	86.68	91.19	87.99	90.22	91.12
F_2+F_4	128	77.62	75.46	85.56	84.37	82.02	89.12	89.10
F_3+F_4	192	84.14	80.24	85.56	90.27	87.22	89.07	92.14
$F_1+F_2+F_3$	256	84.41	83.02	88.86	91.87	90.24	92.59	92.18
$F_2+F_3+F_4$	256	85.18	80.44	87.03	89.92	87.43	92.18	91.89
$F_1+F_3+F_4$	256	86.28	83.04	89.22	91.78	90.26	92.87	92.89
$F_1+F_2+F_3+F_4$	320	86.09	82.96	89.04	91.57	90.17	92.48	92.12

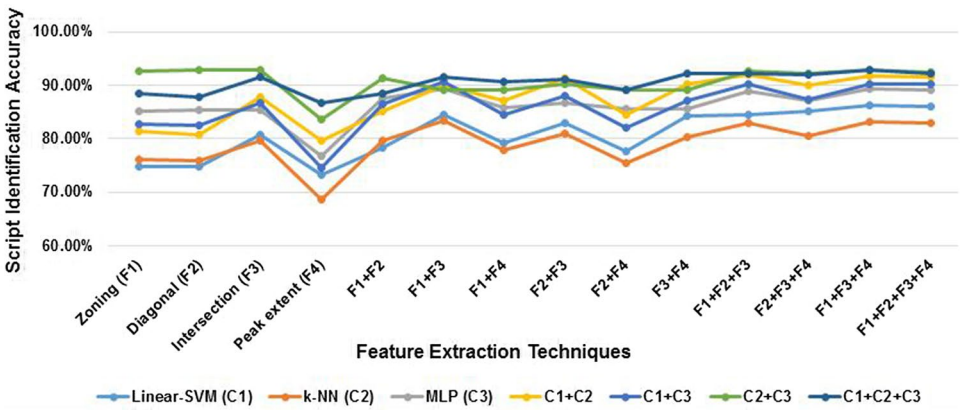

Fig. 7 Script identification results using different combinations of features and classifiers

Table 6 Recognition accuracy achieved for English language

Feature extraction techniques	Feature vector length	Classification techniques						
		Linear-SVM (C_1) (%)	k-NN (C_2) (%)	MLP (C_3) (%)	C_1+C_2 (%)	C_1+C_3 (%)	C_2+C_3 (%)	$C_1+C_2+C_3$ (%)
Zoning (F_1)	64	86.61	80.15	75.07	90.07	83.35	78.07	88.36
Diagonal (F_2)	64	86.61	80.23	75.07	89.97	83.43	78.28	87.77
Intersection (F_3)	128	89.76	82.69	81.00	91.24	85.99	84.24	91.46
Peak extent (F_4)	64	79.61	79.52	65.07	82.79	82.26	67.72	86.60
F_1+F_2	128	86.00	80.15	75.00	89.44	83.35	78.00	88.54
F_1+F_3	192	90.00	82.30	81.46	91.60	85.59	84.71	91.48
F_1+F_4	128	85.84	71.30	76.00	89.27	74.15	79.04	90.69
F_2+F_3	192	90.23	82.23	81.46	91.82	85.51	84.81	91.12
F_2+F_4	128	85.76	71.30	76.00	89.19	74.15	79.04	89.10
F_3+F_4	192	89.46	79.46	82.15	91.10	82.63	85.43	92.14
$F_1+F_2+F_3$	256	90.30	82.15	80.30	91.91	85.43	83.51	92.18
$F_2+F_3+F_4$	256	89.61	80.92	82.38	91.19	86.78	85.67	91.89
$F_1+F_3+F_4$	256	89.69	80.92	82.46	91.27	84.15	85.75	90.29
$F_1+F_2+F_3+F_4$	320	89.76	80.92	81.69	91.35	82.28	84.79	92.12

features and a combination of Linear-SVM, k-NN, and MLP classifiers.

Table 7, depicts the recognition results of Hindi text recognition. We have achieved an accuracy of 84.67% for Hindi text with a combination of zoning, diagonal, and intersection and open end points based features and linear-SVM, k-NN, and MLP classifiers as shown in Table 7. Table 8, depicts the recognition results of Punjabi text recognition. We have achieved an accuracy of 86.79% for Punjabi text recognition with zoning, intersection and open end points and peak extent based features based features and a combination of Linear-SVM, k-NN, and MLP classifiers as shown in Table 8. These results are graphically depicted in Figs. 8, 9 and 10. Figure 8 depicts the recognition results of English language, Fig. 9 depicts the recognition results

Table 7 Recognition accuracy achieved for Hindi language

Feature extraction techniques	Feature vector length	Classification techniques						
		Linear-SVM (C_1) (%)	k-NN (C_2) (%)	MLP (C_3) (%)	C_1+C_2 (%)	C_1+C_3 (%)	C_2+C_3 (%)	$C_1+C_2+C_3$ (%)
Zoning (F_1)	64	76.77	61.34	58.48	65.02	61.98	70.03	81.37
Diagonal (F_2)	64	76.65	61.53	58.88	65.22	62.41	70.52	81.24
Intersection (F_3)	128	74.90	68.68	65.45	72.80	69.37	78.38	79.39
Peak extent (F_4)	64	55.80	62.28	52.24	66.01	55.37	62.56	59.14
F_1+F_2	128	76.58	61.53	58.92	65.22	62.45	70.56	81.17
F_1+F_3	192	79.26	68.81	66.08	72.93	70.04	79.14	84.01
F_1+F_4	128	75.84	62.16	55.56	65.99	58.89	66.54	80.39
F_2+F_3	192	78.96	68.88	66.32	73.01	70.29	79.42	83.69
F_2+F_4	128	75.84	65.29	55.56	69.20	62.17	70.25	80.39
F_3+F_4	192	76.65	63.45	63.89	67.52	67.72	76.52	81.24
$F_1+F_2+F_3$	256	79.88	68.32	66.44	72.41	72.78	82.24	84.67
$F_2+F_3+F_4$	256	79.13	64.29	64.26	68.14	68.29	77.16	83.87
$F_1+F_3+F_4$	256	79.01	65.84	64.14	69.79	67.19	75.92	83.75
$F_1+F_2+F_3+F_4$	320	80.20	66.28	64.47	70.25	69.35	78.36	83.01

Table 8 Recognition accuracy achieved for Punjabi language

Feature extraction techniques	Feature vector length	Classification techniques						
		Linear-SVM (C_1) (%)	k-NN (C_2) (%)	MLP (C_3) (%)	C_1+C_2 (%)	C_1+C_3 (%)	C_2+C_3 (%)	$C_1+C_2+C_3$ (%)
Zoning (F_1)	64	77.28	65.07	61.64	80.37	67.67	64.10	70.69
Diagonal (F_2)	64	77.28	64.85	61.78	81.12	67.44	64.25	81.20
Intersection (F_3)	128	78.14	68.71	67.00	82.26	71.45	69.68	79.95
Peak extent (F_4)	64	62.57	58.35	53.21	66.07	72.85	52.28	80.69
F_1+F_2	128	71.94	61.37	54.05	75.89	63.82	56.23	82.15
F_1+F_3	192	74.00	64.57	58.05	74.72	67.15	60.37	79.61
F_1+F_4	128	71.42	58.74	52.91	75.42	52.68	56.97	79.73
F_2+F_3	192	73.82	64.68	58.11	76.78	67.26	61.56	83.04
F_2+F_4	128	71.20	58.97	52.42	74.41	52.92	57.38	82.20
F_3+F_4	192	71.65	58.34	54.80	74.52	60.67	57.92	82.34
$F_1+F_2+F_3$	256	74.62	64.68	58.62	77.61	67.26	61.98	82.91
$F_2+F_3+F_4$	256	73.88	61.08	57.60	76.83	63.52	59.89	82.99
$F_1+F_3+F_4$	256	74.00	61.08	57.42	76.96	63.52	59.71	86.79
$F_1+F_2+F_3+F_4$	320	74.51	61.72	57.42	77.49	64.18	62.18	81.01

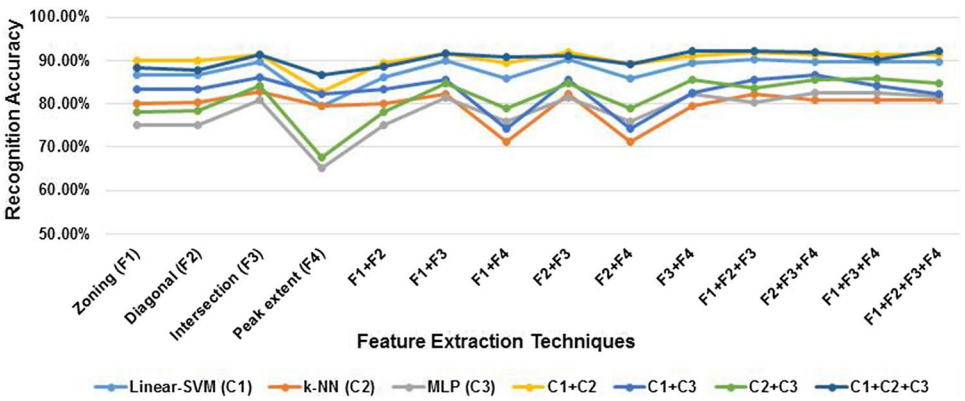

Fig. 8 Recognition results for characters written in English language

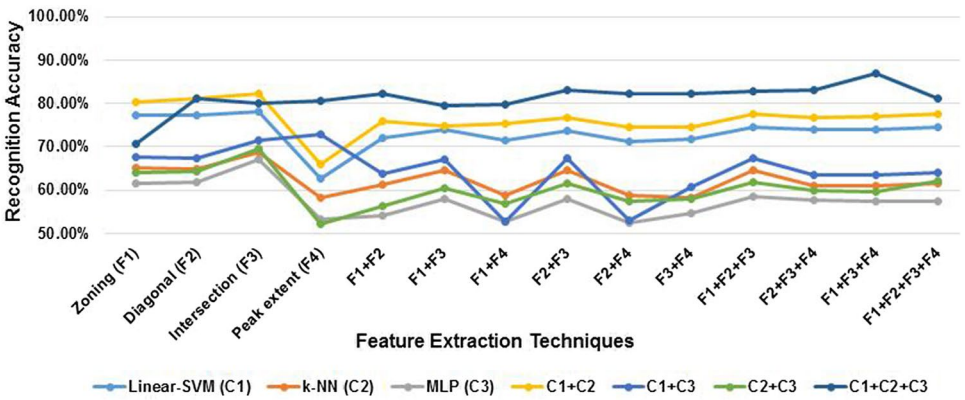

Fig. 9 Recognition results for characters written in Hindi language

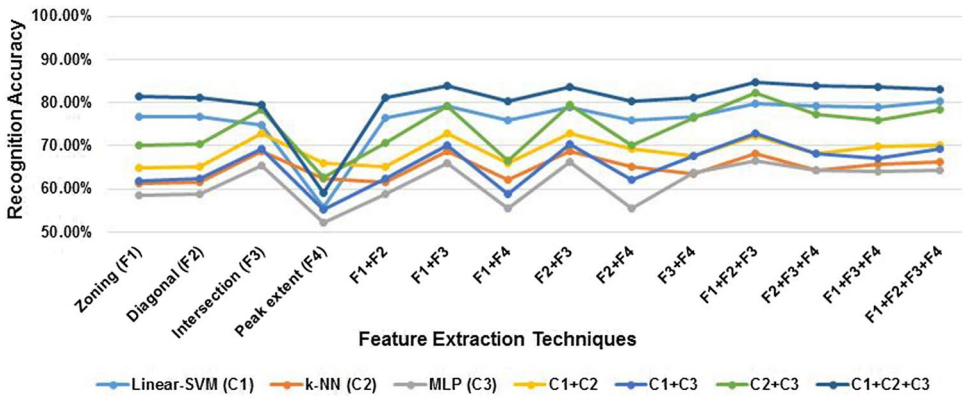

Fig. 10 Recognition results for characters written in Punjabi language

of Hindi language and Fig. 10 depicts the recognition of Punjabi language.

9 State-of-the-Art Work

In this section, we have discussed about the state-of-the-art work related to the present study. A depicted in Table 9, few researchers have reported work for multi-lingual character recognition. But, they promising results are not achieved by them. In presented paper, authors have presented a multi-lingual character recognition system for English, Hindi and Punjabi character recognition. They have achieved a recognition accuracy of 92.18%, 84.67% and 86.79% for English, Hindi and Punjabi, respectively.

10 Conclusion

This paper presents the recognition system for handwritten multi-lingual text document comprising English, Hindi and Punjabi language. Before stating the recognition system, description about the challenges faced during recognition of multi-lingual documents, their application areas and some state-of-the-art work in this area has been given. For script identification, four features, namely, zoning features, diagonal features, horizontal peak extent based features, intersection and open end point based features and three classifiers, namely, k-NN, Linear-SVM and MLP are considered. Before generating the results, a detailed description about features and classifiers considered in this work has been given. Experimental results are conducted on a database of 4920 samples of pre-segmented characters written in English, Hindi and Punjab language and achieved a recognition rate of 92.18%, 84.67% and 86.79% for English, Hindi

Table 9 Comparative study of state-of-the-art-works versus proposed work

Authors	Dataset	Feature extraction	Classification technique	Results
Pandey et al. [3]	Hindi, English and Urdu scripts	DCT and DWT	Neural network	82.70% accuracy
Rani et al. [4]	English and Punjabi characters	Gabor and Gradient features	SVM	Identification rates of 98.9% and 99.45% with Gabor and Gradient features, respectively
Kozielski et al. [5]	French and English text	Feature reduction using PCA	HMM and LSTM recurrent neural network	Scored the first place for Latin script recognition
Kaur and Mahajan [6]	English and Punjabi script	Headline and character density feature	Matches the extracted features with stored features of character	Script identification rates of 91.8% and 89.7% for English and Punjabi, respectively
Surinta et al. [7]	Thai, Bangla and Latin scripts	Local gradient feature descriptors	kNN and SVM	SVM obtained high accuracy
Chakraborty and Pal [8]	Bengali, Roman, Kannada, Oriya, Devanagari and Persian scripts	Orientation invariant features	SVM	14.8% errors as compared to the method proposed by Morillot et al. [10] which gives 45.3% errors
Mandal et al. [11]	UNIPEN English character, UNIPEN ICROW-03 English word, and locally collected Assamese digit databases	GMM posterior features	SVM	Recognition results are promising over the reported work employing point-based features
Proposed work	English, Hindi, and Punjabi	Zoning, Diagonal, Intersection and open end points, peak extents	Linear-SVM, k-NN and MLP	92.18%, 84.67% and 86.79% for English, Hindi and Punjabi

and Punjabi, respectively. For script identification, maximum accuracy of 92.89% has been achieved using a combination of considered classifiers.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bag S, Harit G (2013) A survey on optical character recognition for Bangla and Devanagari scripts. SADHANA 38(1):133–168
2. Govindaraju V, Setlur SR (2010) Guide to OCR for Indic scripts. Document recognition and retrieval series title advances in computer vision and pattern recognition. https://doi.org/10.1007/978-1-84800-330-9
3. Pandey A, Singh S, Kumar R, Tiwari A (2012) Handwritten script recognition using soft computing. Int J Adv Res Comput Sci Electron Eng 1(6):6–11
4. Rani R, Dhir R, Lehal GS (2013) Script identification of pre-segmented multi-font characters and digits. In: International conference on document analysis and recognition, pp 1150–1154
5. Kozielski M, Doetsch P, Hamdani M, Ney H (2014) Multilingual off-line handwriting recognition in real-world images. In: 11th IAPR international workshop on document analysis systems, pp 121–125
6. Kaur I, Mahajan S (2015) Bilingual script identification of printed text image. Int Res J Eng Technol 2(3):768–773
7. Surinta O, Karaaba MF, Schomaker LRB, Wiering MA (2015) Recognition of handwritten characters using local gradient feature descriptors. Eng Appl Artif Intell 45:405–414
8. Chakraborty D, Pal U (2016) Baseline detection of multi-lingual unconstrained handwritten text lines. Pattern Recognit Lett 74:74–81
9. Morillot O, Likforman-Sulem L, Grosicki E (2013) New baseline correction algorithm for text-line recognition with bidirectional recurrent neural networks. J Electron Imaging 22(2):1–11
10. Farulla GA, Murru N, Rossini R (2017) A fuzzy approach to segment touching characters. Expert Syst Appl 88:1–13
11. Mandal S, Prasanna SRM, Sundaram S (2018) GMM posterior features for improving online handwriting recognition. Expert Syst Appl 97:421–433
12. Kumar M, Jindal MK, Sharma RK, Jindal SR (2018) Character and numeral recognition for Non-Indic and Indic scripts: a survey. Artif Intell Rev, pp 1–27. https://doi.org/10.1007/s10462-017-9607-x