Artificial Intelligence Review (2020) 53:5517–5558
https://doi.org/10.1007/s10462-020-09827-4

Ancient text recognition: a review

Sonika Rani Narang[1] · **M. K. Jindal**[2] · **Munish Kumar**[3]

Published online: 10 April 2020

Abstract
Optical character recognition (OCR) is an important research area in the field of pattern recognition. A lot of research has been done on OCR in the last 60 years. There is a large volume of paper-based data in various libraries and offices. Also, there is a wealth of knowledge in the form of ancient text documents. It is a challenge to maintain and search from this paper-based data. At many places, efforts are being done to digitize this data. Paper based documents are scanned to digitize data but scanned data is in pictorial form. It cannot be recognized by computers because computers can understand standard alphanumeric characters as ASCII or some other codes. Therefore, alphanumeric information must be retrieved from scanned images. Optical character recognition system allows us to convert a document into electronic text, which can be used for edit, search, etc. operations. OCR system is the machine replication of human reading and has been the subject of intensive research for more than six decades. This paper presents a comprehensive survey of the work done in the various phases of an OCR with special focus on the OCR for ancient text documents. This paper will help the novice researchers by providing a comprehensive study of the various phases, namely, segmentation, feature extraction and classification techniques required for an OCR system especially for ancient documents. It has been observed that there is a limited work is done for the recognition of ancient documents especially for Devanagari script. This article also presents future directions for the upcoming researchers in the field of ancient text recognition.

Keywords OCR · Feature extraction · Classification · Devanagari · Ancient

✉ Munish Kumar
munishcse@gmail.com

[1] Department of Computer Science, D.A.V. College, Abohar, Punjab, India

[2] Department of Computer Science and Applications, Panjab University Regional Centre, Muktsar, Punjab, India

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

1 Introduction

In this digital day and age, it has become obligatory to have all the available information in a digital form recognized by machines. In the digitization phase of any document, the printed or handwritten text is converted into digital form either by scanning the given document or by using some digital camera or writing with a digitizer connected with a LCD. Character recognition is a process to identify characters from these sources. There are many issues which make the development of an optical character recognition (OCR) system very complex. Some of these issues are discussed below:

- Unique writing styles make development of an OCR system for handwritten documents very difficult.
- Noisy and degraded documents make pre-processing a complex task.
- Touching and overlapping characters make segmentation difficult.
- Thick and uneven characters make segmentation and recognition very difficult.
- Faded documents make recognition very complex.
- Collection of the database is a tedious and manual task.
- Feature extraction and classification techniques must be chosen with care and after thorough experimentation.

After studying the various aspects as well as behavior of the present system, the obtained knowledge will help to develop an OCR for Devanagari ancient text documents. For any research work, literature review helps the researcher to gain knowledge about the current developments in the related field and to get motivation for a new concept. This paper presents a comprehensive survey on the work for different stages of a typical OCR system with special reference to ancient manuscripts. The main focus of this paper is to present various approaches and techniques for segmentation, feature extraction and classification stages of an OCR process. This study aims at helping novice researchers to be acquainted with various existing approaches and techniques. This study provides the existing approaches and techniques with their probable application areas in a nutshell. This paper shows the best accuracies obtained by using different feature extraction and classification techniques. It makes it easy for the researchers to compare these techniques. Tabulation has been done for the existing work. It makes easy and quick analysis of the existing work. From this study, very less work has been done for the recognition of ancient documents. So, this paper presents a guideline for selecting new research areas.

1.1 Block diagram of OCR system

Block diagram of a typical OCR system is depicted in Fig. 1. The main phases of OCR include image acquisition, pre-processing, segmentation, feature extraction, classification, and post-processing. This paper presents a survey on various phases of an OCR system with special focus on a survey of ancient documents.

1.2 Application areas of an OCR system

In this section, authors have listed some important applications of OCR system:-

- Postal address recognition

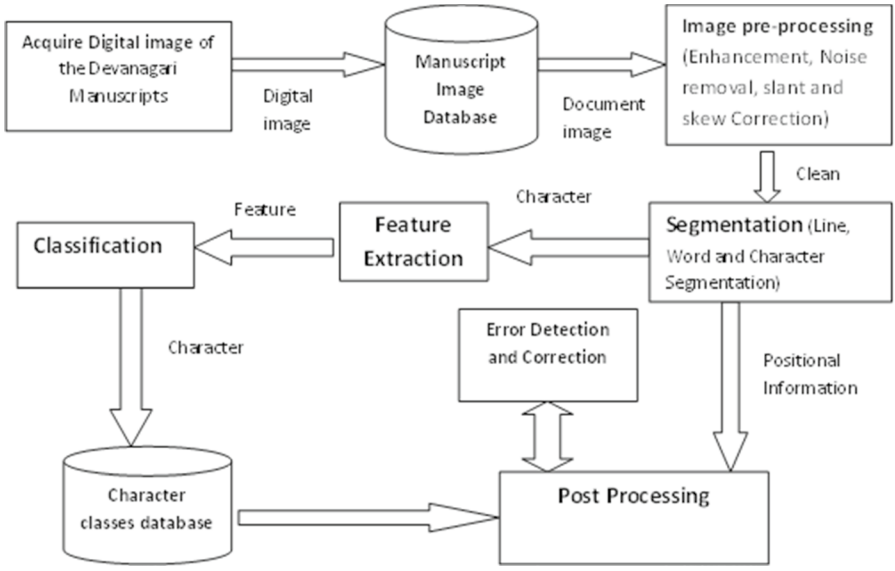

Fig. 1 Block diagram of a typical OCR system

- Cheque processing
- Reading devices for blind persons
- Digitization of libraries
- Form processing
- Automatic number plate recognition
- Contact lists from business card information.

1.3 History of OCR

Depending on the versatility, robustness, and efficiency, commercial OCR systems may be divided into the following four generations (Pal and Chaudhuri 2004).

1.3.1 First generation OCR systems

Character recognition originated as early as 1870 when Carey invented the retina scanner, which is an image transmission system using photocells. In 1900, Russian scientist Tyurin used it as an aid to the visually handicapped. However, the first generation machines appeared in the beginning of the 1960s with the development of the digital computers. It is the first time OCR was realized as a data processing application to the business world (Mantas 1986). In first generation machines, only some specially designed characters could be recognized by machines. The first commercialized OCR of this generation was IBM 1418, which was designed to read a special IBM font, 407. It used template matching for recognition. Template matching compares the character image with a library of prototype images for each character.

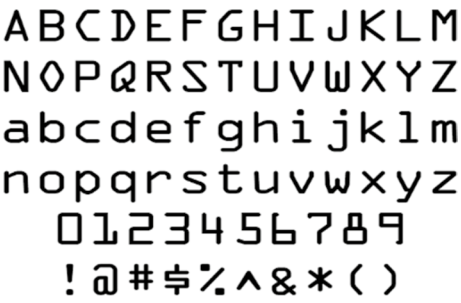

Fig. 2 OCR-A font

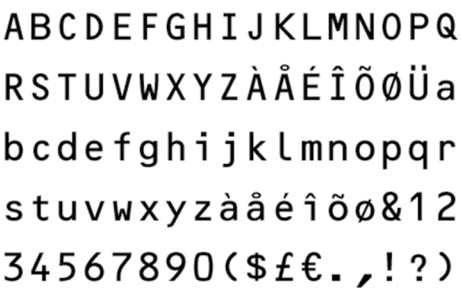

Fig. 3 OCR-B font

1.3.2 Second generation OCR systems

Next generation OCR machines appeared between 1960 and 1970. These machines could recognize numerals and a few letters and symbols of machine printed text. The character set was limited to numerals and a few letters and symbols. Structured analysis approach was used for character recognition. Some standard fonts like OCR-A (Fig. 2) and OCR-B (Fig. 3) were designed to facilitate OCR.

1.3.3 Third generation OCR systems

Third generation systems appeared during 1975 to 1985. Now, OCR systems for large printed and handwritten character sets could be recognized. For third generation OCR systems, documents of poor quality, low cost, and high performance were the main challenges. Commercial OCR systems were developed with the above said capabilities.

1.3.4 Fourth generation OCR systems (OCRs today)

The fourth generation OCR systems are able to deal with complex documents which include text, graphics, tables, color documents, mathematical symbols, unconstrained handwritten characters, and noisy documents. A large number of research papers claim accuracy rate up to 99%. But, this accuracy is for finely printed text documents. A lot of work has to be done for handwritten documents and degraded documents (documents with noisy, broken, touching or overlapping characters, skewed or slanted text, multilingual text etc). Not much work has been done for ancient documents, especially in Devanagari script.

In fact, at various research laboratories, the challenge is to develop robust methods that remove as much as possible of the typographical and noise restrictions while maintaining rates similar to those provided by limited-font commercial machines (Belaid 1997).

2 Reported work

A lot of literature is available for various stages of an OCR system. A lot of research has been done for segmentation, feature extraction, classification and post processing of the text document. Next subsection describes literature review for the various phases of a standard OCR system.

2.1 Segmentation

Segmentation is the process of extracting important objects by partitioning an image into foreground and background pixels. The segmentation stage separates the manuscript image into various logical parts. It is the most important phase of the OCR system. The result of segmentation phase determines the accuracy of the next phases. Accuracy of next phases like feature extraction and classification will be affected adversely if the segmentation is not done properly. If logical parts are not segmented correctly, it will result in a wrong identification. So, much emphasis is laid on this phase. Rani (2015) has presented a systematic review of segmentation phase of OCR. Narang and Jindal (2018) have presented various problems in the segmentation of Devanagari ancient text documents. Segmentation stage is further divided into following sub-phases:

- Page segmentation
- Line segmentation
- Word segmentation
- Character segmentation.

2.1.1 Page segmentation

There may be text areas as well as non-text areas in a document image. Page segmentation is used to segment text areas from non-text areas, identification of the number of columns in text areas, identification of text areas spanned among multiple columns, etc. (Rani 2015).

2.1.2 Line segmentation

In this phase, boundaries of various lines are identified in the image. To identify lines is easy if the lines are clearly isolated from each other. Line identification becomes complex if the document image has touching lines, overlapping lines, curved lines, unevenly spaces between lines, skewed lines, variable-sized characters or noise.

2.1.2.1 Issues in line segmentation For handwritten documents in general and ancient documents in particular, line segmentation is not an easy task due to the following issues:

- Documents may contain touching components i.e. characters in one line may be touching with some character of the adjacent line. In such case, it is very difficult to find the line boundary.
- Documents may contain overlapping lines i.e. characters of one line are extended beyond their line boundary. Y-coordinate of a part of characters in the previous line may be more than the y-coordinate of a part of a character in the next line. In other words, y-coordinates of the characters of two adjacent lines overlap. It makes segmentation difficult.
- Lines in a document may be skewed. It is difficult to segment skewed lines. Lines segmentation is even more difficult if the line has multiple skews.
- Curvilinear lines make line segmentation complex.
- Ascenders and descenders in a script like Devanagari and Gurumukhi complicate line segmentation.
- Line segmentation is very difficult in documents which are degraded due to noise, broken characters, holes, etc.

2.1.2.2 Approaches and techniques for line segmentation Rani (2015) has done a systematic review on line segmentation. She has categorized line segmentation as depicted in Fig. 4. Line segmentation uses three main approaches:

- Top-down approach
- Bottom-up approach
- Hybrid approach.

In top-down approach, text line segmentation methods start from larger text regions, recursively partition the document image into smaller regions on the basis of some property and try to find separating white space between the text lines. In bottom-up approach, text line segmentation methods identify and progressively combine smaller units to get larger units and finally text lines using several similarity and distance measures to find group aggregation of foreground pixels that belong to the same text line. Hybrid approach is a combination of both the top-down and the bottom-up approach to achieve better results.

A. Top-down approach

Top-down approach of line segmentation includes following methods:

- Projection based methods
- Level set method
- Smearing methods
- Dynamic programming.

I. Projection based segmentation

In this method, a binary image is used and the count of black pixels in each row is calculated. Line boundary is identified where count of black pixels is less than or equal to a certain threshold value (typically zero). This method works well with printed text or with clean handwritten text but it does not give good results for text with short lines, narrow lines and overlapped or touching lines or characters. Shapiro (1993) enhanced projection

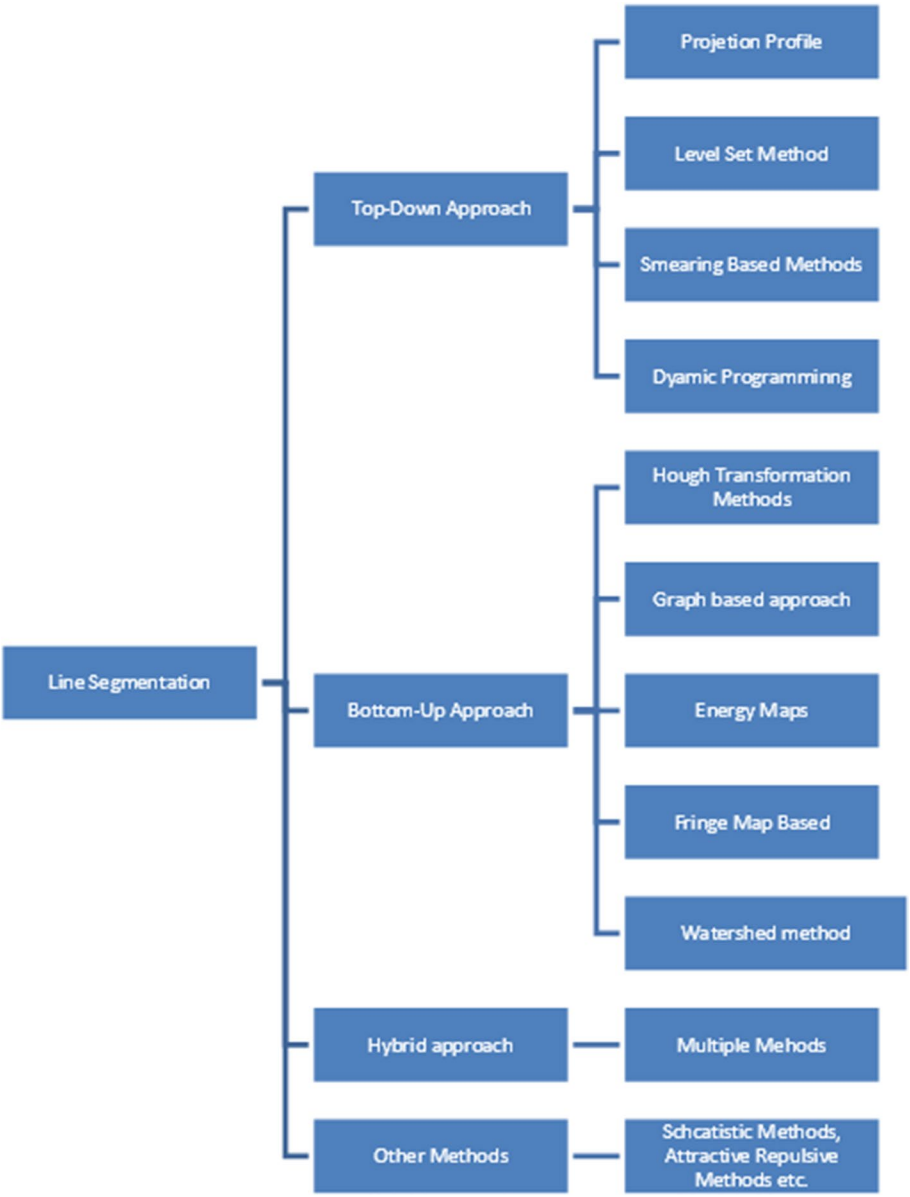

Fig. 4 Categories of line segmentation

profile for skewed text. They used a projection at an angle according to the slope of the text lines. Further, Hough transform is used to determine that angle. This eliminates the problem of handwriting slope. Lehal and Dhir (1999) presents a range free skew detection technique for machine printed Gurmukhi text documents. This method is based on projection profiles. In images, if there are skewed lines or some moderately fluctuating lines, documents need to be divided in a number of vertical stripes and black pixels are counted

in each stripe. This method is known as piecewise projection (Zahour et al. 2001; Pal and Datta 2003; Jindal and Lehal 2012). Zahour et al. (2001) considered border line as a collection of horizontal line segments. They used a partial contour following with the projection method to detect the separating lines from the handwritten Arabic texts. The technique used in this work divides the text image into columns. Partial projection is performed on each column, in this approach. Afterwards, a partial contour following method is used for the detection of separating lines, in the direction of writing and also for the opposite direction. Tripathy and Pal (2004) combined projection based method with water reservoir technique to improve the segmentation results. They divided the document into vertical stripes. By analyzing the height of the water reservoirs obtained from different components of the document, the width of a stripe is calculated. Stripe-wise horizontal histograms are then figured out and the relationship of the peak-valley points in the histograms is then used for the process of line segmentation.

Jindal et al. (2010) proposed a solution for the segmentation of horizontally overlapping lines and also provided a solution for the problem of other stripes to eight most widely used printed Indian scripts. Firstly, the entire document has been divided into stripes and then, the proposed algorithm has been applied for segmentation of horizontally coinciding lines and associating small stripes to their respective lines. The approximate accuracy achieved by the algorithm is 96.45–99.79% which is dependent upon the script sample. The segmentation process to segment horizontal overlapping has also been tried on the script containing different sized text, i.e. the newspaper articles in which large heading lines get overlapped with normal sized text lines. Alaei et al. (2011) used painting technique which works piece-wise for the segmentation of lines in unconstrained handwritten text. Arivazhagan et al. (2007) divided the whole document in vertical stripes and then calculated projection profiles of each stripe with the projection profiles method. Firstly, the candidate lines were assessed from the first 25% projection profile of the document. The lines traverse around any obstructing handwritten connected component by associating it to the lines present above or below. This conclusion was made either (i) by the probability obtained from a distance metric or (ii) by modeling the text lines as bivariate Gaussian densities and evaluating the probability of the component for each Gaussian. The algorithm can also handle the documents having skew and overlapping text lines. The authors obtained 97.31% line segmentation accuracy. Ptak et al. (2017) presented projection based line segmentation with a variable threshold on unconstrained handwritten Polish documents. They developed an algorithm based on projection profile. It works on thresholding, but the threshold value is allowed to change so as to determine low or overlapping peaks in the graph. The algorithm is robust in text-line detection with respect to different text-line lengths. Weliwitage et al. (2005) described a modified version of projection-based method which employs an optimization technique called Cut Text Minimization (CTM) method. They used a modified projection method to obtain starting line estimates. Then, the CTM method finds a cut line (path) in between the text lines in such a way that it minimizes the text-line pixels cut by the segmentation line. The method tries to track around the ascenders or descenders to avoid cutting them. If the deviation is very large, the segmentation line aborts and continues its forward path. This method can handle distorted documents. Din et al. (2016) presented a system based on projection profile and a set of heuristics for segmentation of printed Urdu text images into lines and ligatures. In this method, initially, the lines are segmented using projection profiles. The height of each resulting zone is then computed and the median zone height in the document is determined. They dilated the document image with a square structuring element to join neighboring ligatures and words as well as the secondary ligatures with the primary components. Median zone height is then

used as the threshold for finding local peaks in a dilated version of the document image which in turn is used for finding the valley index (that is the probable text-line boundary) between two consecutive local peaks in the image.

II. Level set method

Level-set methods (LSM) are a conceptual framework which uses the tool of level sets for the numerical analysis of different shapes. Level-set methods are used in the numerical computations of curves and surfaces that can be performed on a fixed Cartesian grid without parameterizing these objects. Level-set method is a computational technique for tracking the propagating interface and can also be considered as a numerical method which can follow the evolution of interfaces. While using the level-set method in image segmentation, a closed surface is evolved by expanding from a point, fitting the closed surface of the region it is released (Phillips, 1999). The idea behind is to evolve a curve that will ultimately stop along the edges of an object in the given image (Manjusha et al. 2012). Level set based method is proved to be an effective top-down methodology for unconstrained handwritten documents. Ding et al. (2012) proposed an approach based on the level set method and density estimation. From an input document image, a probability map was estimated, in which each element represents the probability that the underlying pixel belongs to a text line. The level set method is then exploited to evolve an initial estimate which can further determine the boundary of neighboring text lines. The knowledge used by this algorithm is not the script-specific knowledge. The proposed algorithm is robust to change of scale, noise, and rotation. The main disadvantage of this algorithm is its high computation complexity. A combination of the level-set method and diffusion techniques is used by Majusha et al. (2012) for improving the segmentation accuracy in document processing task for Hindi language.

III. Smearing methods

For printed documents, smearing methods, such as run-length smoothing or fuzzy run-length matrix, can be applied. The technique involves smearing of consecutive black pixels along the horizontal direction. If their distance is within a predefined threshold, the white spaces between them are filled with black pixels. In the smeared image, the bounding boxes of the connected components are considered as text lines (Razak et al. 2008). Wong et al. (1982) employed Run Length Smearing Algorithm (RLSA) for line segmentation in printed documents. If the distance between the black pixels representing foreground in the binary image, is below a predefined threshold, they are linked together along the horizontal direction. The bounding boxes of the connected components in the smeared image enclose text lines. Shi et al. (2005) used an adaptive Local Connectivity Map (ALCM), where the value of each pixel is calculated by summing up of all pixels in the original image within a specified horizontal distance. After thresholding the smeared image, the connected components then represent probable regions of text lines. Kennard and Barrett (2006) used a method with an additional approach to deal with free form handwritten historical documents. A variant of this method adopted to gray level images is described by Lebourgeois (1997). Also there are slight modifications of the RLSA used for recognition of handwritten documents (Sarkar et al. 2011). They applied a modified version of Run Length Smoothing Algorithm (RLSA), named Spiral Run Length Smearing Algorithm (SRLSA), to extract the words from the text lines of the images of unconstrained documents written in Bangla. Shi and Govindaraju (2004) built fuzzy run length matrix for detection of lines. At each pixel, the fuzzy run length was the highest amount of the background along

the horizontal direction. If the number of foreground pixels does not exceed a predefined value, they were skipped. This matrix was considered to be the threshold matrix for making several pieces of text lines that appear without descenders and ascenders. The authors implemented their algorithms on different historical manuscripts.

Gomathi et al. (2012) described top-down approach of line segmentation for handwritten text documents. Run-length is then used for obtaining structural knowledge which classifies components into upper and lower text-lines. Distance metrics and connected components are used recursively to segment short and skewed lines. The accuracy achieved by the system is 91.92%. Nikolaou et al. (2010) proposed a novel Adaptive Run Length Smoothing Algorithm (ARLSA) methodology for segmentation of lines, words and characters in historical and degraded machine-printed text documents. To segment the lines, noise was removed and punctuation marks were eliminated first. Adaptive Run Length Smoothing Algorithm (ARLSA) was used to smear the resulting image that helped grouping together identical text regions. In the next phase, obstacles were spotted and used in order to separate different text lines and text columns. The proposed technique also performed successfully in cases of documents having different sized text, or when text and non-text areas are lying close to each other and when there exists distorted, non-straight and overlapping text lines. This technique has got success when used in handwritten, printed, and historical documents. This technique can also be used for line segmentation, word segmentation and for documents containing only small skew, but is not useful for documents having touching and overlapping components.

IV. Dynamic programming

Liwicki et al. (2007) presented an innovative approach using dynamic programming to detect text lines in on-line handwritten documents. Firstly, starting point of each textual line was estimated using projection profile method. After skew removal, dynamic programming procedure was applied to identify the text lines in the document. The proposed technique applied a cost function to search an optimal path. The cost function applied consisted of a combination of a number of auxiliary functions. It prevented the optimal path from crossing a text line. Line segmentation accuracy of 99.79% on 100 documents of the IAM-OnDB database was claimed by the authors. Fragkou et al. (2004) introduced a dynamic programming algorithm which accomplishes linear text segmentation by global minimization of a segmentation cost function which integrates two factors: (a) word similarity within the segment and (b) segment length information. For segmentation of Choi's text collection, the algorithm attained the best segmentation accuracy value so far reported in the literature.

B. Bottom-up approach

Top-down approach partitions the document image recursively into text regions, lines, words and characters whereas, Bottom-up approach groups small units of the document image (connected components, words, characters, pixel etc.) into text lines and then into text regions. Bottom up grouping work as a clustering process which aggregates the close components of the image and does not depend on the assumption of straight lines (Yin and Liu 2009).

Bottom-up approach for line segmentation involves following methods:

- Grouping methods
- Hough transform based method
- Graph based method
- Energy maps
- Fringe map based method
- Watershed method

I. Grouping methods

This method involves building alignments by aggregating units in a bottom-up approach. Grouping methods such as nearest neighbor joining scheme is used for segmentation. In this method, the connected components of black pixels are grouped. These connected components can then be joined to form lines. It is possible to choose between different units that can be joined with the same neighborhood by using some quality criteria. The joining scheme depends upon local as well as global criteria such as similarity, continuity and proximity. This approach is more used for documents analysis (Hussain et al. 2015). The Docstrum method of O'Gorman (1993) is a bottom-up grouping method. It performs well on fairly printed documents as well as on handwritten documents having slightly curved characters. It merges adjacent components using some rules which are based on the geometric relationship between K-nearest neighbor units. Likforman-Sulem et al. (1995) developed an iterative method which was based on perceptual grouping on the basis of Gestalt criteria of similarity, direction continuity and proximity to connected components in the group.

II. Hough-based transformations

Hough transformations can be applied to find the skew angle. In this method, a set of points of the initial image comprises the input while the lines that fit best in this set of points are calculated (Louloudis et al. 2009). The set of points considered in the voting procedure of the Hough transform is usually the gravity centers of the CCs. Louloudis et al. (2009) presented a text line detection method for unconstrained handwritten documents based on three distinct steps. The first step comprises preprocessing for image enhancement, connected component extraction and average character height estimation. In the second step, a block-based Hough transform is applied to the detection of potential text lines while a third step is used to correct possible false alarms. A grouping method of the remaining connected components which uses the gravity centers of the corresponding blocks is applied. Almazan et al. (2017) proposed a probabilistic algorithm that merged perceptual grouping in the image domain or global accumulation in the Hough domain. In the first stage, lines were detected using a global probabilistic Hough approach. In the second stage, each detected line was analyzed in the image domain to localize the line segments that generated the peak in the Hough map. A probabilistically optimal labeling can be computed exactly using a standard dynamic programming algorithm, in linear time

III. Graph based method

This method makes use of graphs to segment lines. One graph based method is the use of a minimum spanning tree. This method assumes that the distance between words is less than the distance between lines. In this method, a graph of the main strokes of a document is constructed and then the minimum spanning tree is found. Yin and Liu (2009) proposed a method based on the Minimum Spanning Tree (MST) clustering with distance metric learning for handwritten text line segmentation. This bottom-up method can extract

multi-skewed, curved and slightly overlapping text lines. The algorithm is free of artificial parameter, and supervised learning of distance metric improves the accuracy of text line detection significantly. Kumar et al. (2006) presented a graph cut based framework using a swap algorithm to segment document images. The text block is first segmented into lines using the projection profile approach. The framework enables learning of the spatial distribution of the components of a specific script and can adapt to a specific document collection, such as a book. Moreover, they can use both corrections made by the user as well as any segmentation quality metric to improve the segmentation quality.

IV. Energy maps

Different pixels from an image carry different information content. A document can be viewed as a group of low information pixels representing the space between lines of text and respectively a group of high information pixels representing the actual lines of text. Each pixel in the energy map has a value associated with it that represents the amount of information that the given pixel stores in the image. If a high energy pixel is removed from the image, the resulting image has a significant drop in detail, whereas removing a low energy pixel results in a negligible information loss (Boiangiu 2014). Saabni et al. (2014) proposed an approach which computes an energy map of the input text block image and determines the seams that pass across and between text lines. They developed two algorithms, one for binary images and the other for gray scale images. The first algorithm works on binary document images and assumes it is possible to extract the components along text lines. The second algorithm works directly on gray-scale document images. It constructs distance transforms on gray-scale images and computes two types of seams: medial seams and separating seams. The medial seam determines a text line and separating seams define the upper and lower boundaries of the text line. Boiangiu et al. (2014) described an improvement to text line segmentation for handwritten documents based on seam carving using energy map. They proposed dynamic allocation of weights based on the local text direction information. This way, pixels from the same line have a higher probability of selection when calculating the minimum values in energy cost map. Main limitation of this method is the computational time overhead. Arvanitopoulos and Süsstrunk (2014) proposed an algorithm for automatic text line extraction on color and grayscale manuscript pages without prior binarization. Their algorithm was based on seam carving to compute separating seams between text lines. Two types of seams were found: medial seam and separating seams. Seam carving is likely to produce seams that move through gaps between neighboring lines, if no information about the text geometry is incorporated into the problem.

V. Fringe map based methods

Koppula and Negi (2011) proposed text line segmentation of Telugu script documents based on fringe maps. This method assigns a fringe number with each white pixel. Fringe numbers are assigned on the basis of distance from the nearest black pixel. After that, Peak Fringe Number (PFN) has been extracted in some particular direction. PFNs between text lines are identified using a filtering operation. Clustering of PFNs between adjacent lines has been done by considering a broad region. Finally, a segmenting path between lines has been generated by joining the PFNs of the region. Jetley et al. (2012) proposed an improved binarization approach, which after a Pre Segmentation Binarization uses fringe

map based text line segmentation algorithm and then post segmentation binarisation. Binarization is performed twice - once before and once after text line segmentation. The first one is incorporated to carry out quick binarization with the result being just sufficiently suitable for effective text line segmentation. The second one, in contrast, targets a higher accuracy for the recognition stage ahead and capitalizes on parallel processing for obtaining higher speed. This method improves speed and accuracy.

VI. Watershed method

This method involves segmenting an image into the catchment and basin by flooding the morphological surface at the local minimum and then constructing ridges at the places where different components meet (Vincent and Soille 1991). As watershed associates each region with a local minimum, it can lead to serious over-segmentation. To mitigate this problem, some methods allow the user to provide some initial seed positions which help improve the result (Beare 2006). Souhar et al. (2017) presented a watershed based method for line segmentation. They found a vector of local linear regions for each text component that participates to the same line. A recursive function is used on this vector in order to find all components that will be linked together from their centroids in a line. Then the watershed transform is again applied to the new image to estimate the location of the lines. Basu et al. (2007) proposed “water flow" method for text line segmentation. It works on an assumption of hypothetical water which flows from both the sides of the image frame. The characters work as obstruction of this water. The areas left unwetted on the image are stripped and labeled for the extraction of text lines. Brodic (2012, 2015) further extended and improved this algorithm by extracting the connected-components by bounding boxes over text. By this approach, connected components get mutually separated from each other. Hence, the water flow angle is adaptively determined and it defines the unwetted areas on the images. By appropriate water flow angle selection, the unwetted areas get lengthened which leads to the better text line segmentation results.

C. Hybrid approach

Hybrid approach combines the features of top-down and bottom-up approaches to achieve enhanced results. Adiguzel et al. (2012) presented text line segmentation approach, which combines projection based information and connected component based information. The proposed system finds out the baselines of each connected component. Lines are then detected by grouping the baselines of connected components belonging to each line by projection information. On the basis of distance metrics with respect to size of the components, they are assigned to lines. Authors claimed 92.0% accuracy with this approach. Clausner et al. (2012) proposed a line segmentation method that uses a combination of projection profile analysis (top-down) and rule based grouping of connected components (bottom-up). Firstly analysis of the connected component is performed. Next, grouping of connected components to text line candidates is done on the basis of rules. Splitting of large components in under segmented lines is done using local projection profile. Then, small line candidates are merged to their nearest neighbour. Messaoud et al. (2012) proposed a technique which is a combination of three methods of line segmentation viz., detection of nearest neighbor, projection profiles and grouping of connected components. First method

was applied to regions of problems to solve the problem of touching and overlapping components whereas the last two methods were applied on the full image.

D. Other approaches

Apart from above discussed approaches, some other methods are also used for line segmentation as given below:

I. Repulsive-attractive network method

Repulsive-Attractive network works directly on grey-level images. In this method, one by one construction of baselines is performed from the top of the image to the bottom. Pixels in the image are the driving force for baselines and prior extracted baselines act as repulsive forces. The length of lines in this method must be similar. We get many pseudo-baselines which pass through word bodies (Sulem et al. 2006).

II. Stochastic methods

Stochastic methods can be used to find non-linear paths between overlapping text lines (Tseng and Lee 1999). They are based on a probabilistic algorithm. Hidden Markov Modeling is applied for extraction of lines. Depending upon stroke width, image is divided into small cells where each cell corresponds to a different state of Hidden Markov Model. The best segmentation paths are searched from left to right. The segmentation path is a path which does not cross many black pixels and is as straight as possible. In the case of the components which touch each other, the path of maximum probability will cross the touching component at points with as less black pixels as possible. However, the method may fail when that contact point contains a lot of black pixels.

E. Observations for line segmentation from literature review

From the literature review, it was found that OCR research is old in the field of pattern recognition. It was viewed as a problem that could be solved easily but the case was not so. After some initial easy progress it was a very difficult problem. The main steps in any OCR process are: pre-processing, segmentation, feature extraction, classification and post-processing. The phase of pre-processing involves binarization, skew and slant correction, noise cleaning, etc. Segmentation involves different processes for line, word and character segmentation. We can encounter various problems during segmentation like: overlapping line, touching characters, little or uneven space between characters, words and lines, no space between different words etc. Various techniques are given for segmentation. Most popular techniques for segmentation are based on projection profiles. We have horizontal projection profiles for line segmentation and vertical projection profiles for word and character segmentation. This method works well with printed text or with clean handwritten text but it does not give good results for text with short lines, narrow lines and overlapped or touching lines or characters. Hough transform can be applied to find the skew angle. Then projection can be achieved along this angle. For printed documents, smearing methods such as run-length smoothing or fuzzy run-length matrix can be applied. Grouping methods such as nearest neighbor joining scheme is used for segmentation. During grouping we can find connected components by this method. It is possible to choose between different units that can be joined with the same neighborhood by using some quality criteria. This method can work with ancient documents also. Repulsion-attraction network method works directly on grey level images and finds y positions of baselines passing through the word bodies. It works on lines with similar lengths and

can work with ancient documents. Stochastic methods can be used to find non-linear paths between overlapping text lines. Hidden Markov Modeling is used for line extraction. We can conclude that projection, smearing and Hough based methods are easier to implement. These work with straight lines. Local considerations like piece-wise projections, moving windows etc. are used along with these methods to solve the overlapped components problem or touching lines or close lines. Stochastic method is more robust but it requires great care while implementation. It can find non-linear paths to separate overlapping or touching characters. The repulsion attraction method is recurrent method. It may result in false or bad line extraction. Grouping methods don't work well when the lines are closer. An erroneous assessment in an early stage may result in wrong or incomplete alignment. A brief summary of related work for line segmentation is depicted in Table 1.

2.1.3 Word segmentation

This phase involves the separation of words in a line. Words are a collection of characters. Words are identified by inter-word space. Most of the time, words are separated using column histogram. For the construction of column histogram, the count of black pixels in each column is calculated. Zero frequency of black pixel indicates the boundary between two words. Two different techniques are used for word segmentation: distance based approach and recognition based approach (Rani 2015). In distance based approach, the connected components of a line are identified. Afterwards, the distances between these connected components are calculated using various metrics such as Euclidean distance (Louloudis et al. 2009), average run length distance, the convex hull metric (Mahadevan and Nagabushnam 1995), the bounding box distance (Saha et al. 2010; Kim et al. 2004). Finally, the calculated distances are classified as either inter-word or inter-characters gaps based on a predefined threshold. On the other hand, the recognition based approach is used to find the word boundaries based upon classification (Manmatha and Rothfeder 2005). Angadi and Kodabagi (2014) presented a segmentation technique for segmentation of textual lines, words and characters from Kannada text present in the form of low resolution display board images. Vertical profile features were used for extraction of character images from text lines. K-means clustering was used to group inter character gaps into character and word cluster space, which were then used to compute thresholds for words extraction process. The method took into account, the variations in word and character gaps. They achieved 97.54% accuracy for word segmentation. The proposed method is tolerant to variations in font, absence of free segmentation path due to consonant and vowel modifiers, space variations between characters and words, noise and other degradation. Saha et al. (2010) used Hough transform on the line segments to generate its words-level Hough image. For this purpose, various parameters of the Hough transform, like deltaTheta, deltaRo, startTheta, endTheta, pixels Count and connect Distance are initialized or adjusted in such a way that the words can be extracted as a set of connected characters. To analyze the performance of the system with a new Hough image is stored as bmp file. Manmatha and Rothfeder (2005) described a word segmentation technique for handwritten historical documents. They used a gray-level projection profile algorithm to find lines in the images. Each line image is then filtered using an anisotropic Laplacian at a number of scales. This procedure produces blobs corresponding to the portions of characters at small scales and in words at larger scales. Finding the optimum scale for words is the crucial task and is performed by finding the maximum over the scale of the extent or area of the blobs. The recovery of words is

Table 1 Brief summary of work done for text line segmentation

Line segmentation approach	Line segmentation technique	Type of material of text or writing type	Citations
Top-down approach	Projection based segmentation	Handwritten text	Pal and Datta (2003), Jindal and Lehal (2012)
		Skewed handwritten text	Shapiro (1993)
		Printed Gurmukhi documents	Lehal and Dhir (1999)
		Handwritten Arabic text	Zahour et al. (2001)
		Unconstrained Oriya text	Tripathy and Pal (2004)
		Handwritten text	Arivazhagan et al. (2007)
		Printed Indian scripts handwritten Hindi text	Jindal et al. (2010)
		Unconstrained handwritten text	Alaei et al. (2011)
		Distorted documents	Weliwitage et al. (2005)
		Printed Urdu documents	Din et al. (2016)
		Unconstrained handwritten Polish documents	Ptak et al. (2017)
	Level set method	Curves and surfaces	Phillips (1999)
		Hindi language	Manjusha et al. (2012)
		No script specific knowledge	Ding et al. (2012)
	Smearing Methods	Printed documents	Wong et al. (1982)
		Gray scale historical documents	Shi et al. (2005)
		Freeform handwritten historical documents	Kennard and Barrett (2006)
		Gray level images	Lebourgeois (1997)
		Unconstrained handwritten Bangla document images	Sarkar et al. (2011)
		Historical manuscript	Shi and Govindaraju (2004)
		Handwritten text	Gomathi et. al. (2012)
		Historical and degraded machine-printed documents	Nikolaou et al. (2010)
	Dynamic programming	On-line handwritten documents	Liwicki et al. (2007)
		Choi's text collection	Fragkou et al. (2004)

Table 1 (continued)

Line segmentation approach	Line segmentation technique	Type of material of text or writing type	Citations
Bottom-up approach	Grouping methods	Handwritten Assamese characters	Hussain et al. (2015)
		Slightly curved handwritten documents and printed documents	O'Gorman (1993)
		Handwritten documents	Likforman-Sulem et al. (1995)
	Hough transform based method	Unconstrained handwritten documents	Louloudis et al. (2009)
		YorkUrbanDB dataset	Almazan et al. (2017)
	Graph based method	Handwritten text	Yin and Liu (2009)
		Document images	Kumar et al. (2006)
	Energy maps	Handwritten documents based on seam carving using energy map	Boiangiu et al. (2014)
		Binary document images Gray-scale document images	Saabni et al. (2014)
		Color and grayscale manuscript	Arvanitopoulos and Süsstrunk (2014)
	Fringe map based method	Telugu script documents	Koppula and Negi (2011)
		Printed/handwritten documents	Jetley et al. (2012)
	Watershed method	Grayscale images	Vincent and Soille (1991)
		Medical images	Beare (2006)
		Arabic handwritten text	Souhar et. al. (2017)
		Handwritten Bengali and English documents	Basu et al. (2007)
		Handwritten English text	Brodic (2012, 2015)

Table 1 (continued)

Line segmentation approach	Line segmentation technique	Type of material of text or writing type	Citations
Hybrid approach	Grouping base lines of connected components and projection profile	Handwritten documents	Adiguzel et al. (2012)
	Grouping of connected components (bottom-up) and projection profile analysis (top-down)	Historical documents	Clausner et al. (2012)
	Projection profiles, grouping of connected components and detection of nearest neighbor	Handwritten historical documents	Messaoud et al. (2012)
Other approaches	Repulsive-attractive network method	Gray level images	Sulem et al. (2006)
	Stochastic methods	Chinese handwritten document	Tseng and Lee (1999)

Table 2 Summary of work done for word recognition

Word segmentation approach	Word segmentation technique	Type of material of text or writing type	Citations
Distance based approach	Euclidean distance	Handwritten documents	Louloudis et al. (2009)
	Bounding box distance	Bengali, English and mixed script documents	Saha et al. (2010)
	Average run length distance	Korean handwritten text	Kim et al. (2004)
	Convex hull metric	Handwritten English text	Mahadevan and Nagabushnam (1995)
Recognition based approach	Scale space algorithm	handwritten historical documents	Manmatha and Rothfeder (2005)
	k-means clustering	Kannada text	Angadi and Kodabagi (2014)
	Hough transform	Bengali, English and mixed script documents	Saha et al. (2010)

done using the blobs recovered at the optimum scale which is then bounded with a rectangular box. The elimination of boxes of unusual size which are not likely to correspond to words is done in the post processing filtering step. Authors claimed that the technique beats a state-of-the-art gap metrics word-segmentation algorithm on their data. Table 2 gives the summary of literature review for word segmentation.

2.1.4 Character segmentation

In this phase, characters of a word are separated from the already identified words. This is the last step of the segmentation phase. If characters are segmented properly, only then these can be recognized. This is the most difficult step of the segmentation phase if the document contains degraded characters, touching characters, overlapping characters, heavily printed characters etc. Casey and Lecolinet (1996) published a paper illustrating various techniques of character segmentation. Character segmentation techniques were divided by the authors into four categories: holistic strategies, recognition-based segmentation, dissection techniques and mixed strategies (over-segmentation). Survey papers are available for segmentation of characters from the document image (Dunn and Wang, 1992). Dongre and Mankar (2010) gave a review of research on Devanagari Character Recognition and discussed various Devanagari optical character recognition techniques. Segmentation discussed in the paper was straight segmentation which is the technique of identifying members of a character set without identifying their specific classification. It is quite useful for printed character set but less effective for cursive text. Several approaches to segmentation were discussed in the paper. Fujisawa et al. (1992) used projection profiles for touching component segmentation with 95% accuracy. They presented a pattern-oriented segmentation method for optical character recognition that leads to document structure analysis. They extracted connected pattern components and measured spatial interrelations between components grouped into meaningful character patterns. But, this method failed when the text is strongly skewed or overlapping. Kim et al. (2000) have used contour-tracing features for segmentation. From the contour of a touching component, valley and mountain points were estimated and the cutting path was found to segment the characters. The contour tracing algorithm does not work well if the two numerals touch in a straight-line fashion. They obtained a recognition rate of 91.8%. Chen and Wang (2000) used thinning-based methods for segmenting touching handwritten numeral strings. In this paper, background and foreground analysis was used to segment handwritten numeral strings which gets touched. Thinning of both the background and foreground regions was performed. Several possible segmentation paths were constructed. As a final point, the parameters of geometric properties of every possible segmentation path were found and the best segmentation path was then decided. Though, this algorithm can get 96% of the correct rate with a rejection rate of about 7.8%. But, it is time-consuming and also generates protrusions due to which the algorithm sometimes gives wrong results.

Oliveira et al. (2000) used a novel segmentation approach based on contour and profile features. Firstly, local minima of contour and profile features were defined as a basic point (BP). Secondly, an intersection point (IP) was defined as a point which has more than two pixels in its neighborhood. After that, the Euclidean distance system was applied to determine proximity between the basic point and the intersection point. This approach does not successfully solve the problem of multiple touching. Bansal et al. (2002) used the structural properties of the script like presence and relative location

of a vertical line, number of positions of the vertex points, horizontal zero crossings, moments and the nature of constituent pure consonant form in the conjunct for segmentation of touching and fused Devanagari characters. Pal et al. (2003) used water reservoir based method and morphological, structural features to segment touching numerals. In this sense, if we pour water from the top or bottom, the space will then be filled with water. The difference can be performed based on the number and size of the water reservoir. Technique of bounding box (BB) was applied on touching component to find the touching position. Features such as reservoir height, closed loops and distance from center of the component were considered by the authors for determining segmentation points. They achieved 94.8% accuracy. Several drawbacks reported in this work include: big ratio of two segmented digits, long cutting length, nonexistence of the best reservoir, and boundary of the reservoir contains a break point. Nevertheless, the water reservoir approach might have failed while dealing with broken character.

Sharma and Lehal (2006) suggested an algorithm for segmentation of the characters in an iterative manner by focusing on the aspect ratio of the characters, headline and projection profiles. This work performed segmentation in three phases. First phase performs basic segmentation, second phase segments under-segmented words and over-segmentation was handled in the third phase. Tripathy and Pal (2006) proposed water reservoir based method to segment characters of unconstrained Oriya text. First of all, identification of touching and isolated characters in a word was performed. Then reservoir base area points and structural features of the component were used to segment touching characters of the word.

Jindal et al. (2007) presented a study of touching characters found in degraded Gurumukhi text. They divided touching characters in various categories and the proposed new algorithm to segment touching characters in the middle zone of the machine printed script. Jindal et al. (2009) studied segmentation problems in the printed degraded Gurmukhi script and proposed solution for segmenting touching characters in the upper zone of machine printed Gurmukhi script. Structural properties of the Gurmukhi script characters were the basis of the technique. Convexity and concavity of the characters were studied and the touching characters in upper zone were segmented using the top profile projections. They achieved a 91% recognition rate while applying segmentation to the touching characters in the upper zone. Alam and Kashem (2010) gave an Optical Character Recognition (OCR) system for printed characters in Bangla. They used the headline and baseline to segment characters. Reddy et al. (2010) used topological properties in terms of zones, component combinations, and behavioral aspects of syllables in the segmentation process for Telugu script. They proposed split profile algorithm while handling touching components. Bag and Krishna (2015) proposed segmentation of characters in handwritten Hindi words based on structural patterns. The proposed method can work well with high variations in writing style and when skewed header lines are given as input. The average success rate achieved is 96.93%. Sridevi and Subashini (2012) have used computational intelligence techniques for text line and character segmentation of Tamil ancient documents. In this work, two methods were proposed for line segmentation and character segmentation. Method for line segmentation used projection profile and PSO for line segmentation. The method used to segment the characters, used a combination of connected components along with the nearest neighborhood methods. Table 3 gives a summary of literature review for character segmentation.

Table 3 Brief summary of work presented for character recognition

Character segmentation approach	Accuracy (%)	Type of material of text or writing type	Citations
Projection profiles	95.0	Touching component	Fujisawa et al. (1992)
Holistic strategies	91.8	Contour-tracing features	Kim et al. (2000)
Dissection techniques	96.0	Touching handwritten numeral strings	Chen and Wang (2000)
Euclidean distance scheme	97.84	Warped English text	Oliveira et al. (2000)
Mixed strategies (over-segmentation)	–	Touching and fused Devanagari characters	Bansal et al. (2002)
Water reservoir based method	94.8	Unconstrained Bangla text	Pal et al. (2003)
Presence of headline, aspect ratio of the characters and projection profiles	96.22	Handwritten Gurmukhi text	Sharma and Lehal (2006)
Water reservoir based method	95.1	Unconstrained Oriya text	Tripathy and Pal (2006)
Structural properties	91	Printed degraded Gurmukhi script	Jindal et al. (2009)
Use of headline and baseline	97	Printed Bangla characters	Alam and Kashem (2010)
Split profile algorithm	99.98	Telugu script	Reddy et al. (2010)
Structural patterns	96.93	Handwritten Hindi words	Bag and Krishna (2015)
Computational intelligence techniques	–	Tamil ancient documents	Sridevi and Subashini (2012)

2.2 Feature extraction

Feature extraction is a process of extracting meaningful information from an image so that it can be used effectively for classification. Devijver and Kittler (1985) defined feature extraction as the process of taking out information from the given data which is more significant for the purpose of classification, by the way of minimizing pattern variability within the class while enhancing pattern variability between the class. These features can be different for different types of problems. Many kinds of features have been defined in literature. Features are generally classified into three main categories, namely, statistical features, structural features and global transformation and moments based features.

2.2.1 Statistical features

Statistical features are the ones which are obtained from the statistical distribution of pixels in the character image. Statistical features being the topological features are not much sensitive to distortions and local noise. The most commonly used statistical features are zoning, projections, profiles, crossings, etc.

2.2.2 Structural features

Structural features are those features which are based on the geometrical and topological properties of the character. The most commonly used statistical features are aspect ratio, loops, branch points, cross points, strokes and their directions, curves etc.

2.2.3 Global transformations and moments

To reduce the dimensionality of the feature vector, global transformations are used and then the extracted features are made invariant to global transformations like rotation and translation. The most common features based on transformation are Fourier transform, Hough transform, Gabor Transform and Discrete Cosine transform. Moments are scalar quantities which are used to capture the significant features of a function. Moments are purely considered to be the statistical measure of pixel distribution around the center of gravity of the image. Moments also allow capturing of information regarding global shapes (Gonzalez and Woods 1992). If viewed as a mathematical point, moments are considered as projections of a function onto a polynomial basis. In computer vision and image processing techniques, an image moment is calculated by computing the weighted average (moment) intensities of the image pixels or a function of such moments, usually chosen on the basis of some interpretation or property (Singh et al. 2016). Singh et al. (2016) discussed six types of moments, namely, Legendre moment, Zernike moment, geometric moment, moment invariant, affine moment invariant, and complex moment. Hu moments and Zernike moments are among the most commonly used moment for OCR.

2.3 Classification

Classification determines the unknown pattern region of feature space (Singh and Budhiraja 2011). It is used to assign a class to a pattern based on some features. Liu and Han (2007) has summarized the classification methods in 4 categories, namely, statistical methods, kernel methods, artificial neural networks, and multiple classifier combinations.

Statistical classifiers are based on Bayes decision rule, and can be divided into the categories of parametric and non-parametric classifiers (Fukunaga 1990). Feed forward neural networks, including radial basis function (RBF) network, multilayer perceptron (MLP), higher-order neural network (HONN), CNN etc., have been extensively used for the purpose of pattern recognition. Kernel methods include kernel principal component analysis (KPCA), kernel Fisher discriminant analysis (KFDA) and support vector machines (SVMs) etc. Various classifiers are usually combined through ensemble methods to improve the recognition accuracy. Ensemble methods can be divided into sequential ensemble methods and parallel ensemble methods. In sequential ensemble methods, there is sequential generation of the base classifiers. Sequential ensemble methods finish up the dependency between the base classifiers. The overall performance can be improved by considering previously mislabeled examples with higher weight. One example of such methods is AdaBoost. Parallel ensemble methods are the ones where the base classifiers are generated in parallel. Parallel methods end up the independence between the base learners since averaging is done to reduce the error dramatically. One example of such methods is bootstrap aggregating.

Many survey papers for feature extraction are available in literature (Sharma et al. 2013; Trier et al. 1996; Bhopi and Singh 2018; Kumar and Gupta 2018). Kumar et al. (2018) have presented a survey for character and numeral recognition of various non-Indic and Indic scripts. They have also presented a few future directions in the field of character and numeral recognition and discussed various challenges for character and numeral recognition of different scripts. Bag and Harit (2013) endeavored to provide a broad survey of OCR work on Devanagari and Bangla scripts published in the year 2000 and afterwards. This survey contributes work related to the recognition of printed characters, handwritten characters, numerals, handwritten numerals, compound (also known as 'conjunct') characters and a mixture of printed and handwritten characters. It also provides a comparison of all the reported methods. Shah and Badgujar (2013) gave a review on Devanagari Handwritten Character Recognition (DHCR) for Ancient Documents. They described various methods used for OCR for ancient Devanagari documents. Dongre and Mankar (2010) studied and reviewed the available Devanagari Optical Character Recognition techniques. Shahi et al. (2012) presented a brief literature survey on the OCR systems for Handwritten Hindi Curve Script methodologies and also listed various important contributions in this area using Artificial Neural Network. Hussain et al. (2015) used the technique of zoning. In this technique, image was divided into non-overlapping zones, and then the percentage of black pixels in each zone was computed. This percentage worked as feature for that zone. Kimura and Shridhar (1991) have used the static zoning topology on contour representation of character. Depending upon the orientation the contour line segments have been grouped into four subgroups, namely: horizontal, vertical, and diagonal (two orientations: 45 and 135). The number of contour segments belonging to each group was used as the extracted feature for each zone. Singh and Budhiraja (2011) presented an outline of several OCR systems for handwritten Gurmukhi text. It described various feature extraction methods such as projection histogram features, zoning, distance profile features, background directional distribution features. It also described various methods of classification such as k-NN, Support Vector Machine and Probabilistic Neural Network. According to this work, Zoning and k-NN gives 72.5% accuracy, Zoning and SVM gives 73.02% accuracy, Zoning Density and background directional distribution features and SVM with RBF kernel give 95.04% accuracy.

Verma and Ali (2012) described many feature extraction and classification techniques. Statistical feature techniques used by them includes zoning, moments, projection

histograms, n-tuples, crossing and distances. Convexities, concavities, number of end points, number of holes etc. have been used as structural features. For classification various methods described are: statistical methods, syntactic methods, Artificial Neural Networks, template matching and kernel methods. Kumar et al. (2014) used hierarchical zoning with four features, namely, horizontal peak extent, the vertical peak extent, centroid and diagonal features. They proposed a feature set of 105 elements for recognition of offline handwritten Gurmukhi characters. They have achieved a recognition accuracy of 91.80% with SVM classifier. Kumar et al. (2012) have used combinations of various features and classifiers for the recognition of offline handwritten Gurumukhi characters. To find the feature set for a given character, the features such as zoning, directional feature, diagonal feature, intersection and open end points feature, parabola curve fitting based feature, transition feature, and power curve fitting based feature extraction technique have been used. They have used k-NN and SVM with Linear, Polynomial and RBF kernels for classification. The proposed system achieved an accuracy of 94.8% for recognition without applying PCA and accuracy of 97.7% for recognition was achieved by applying PCA with the database of 7000 isolated Gurmukhi characters. In another paper by Kumar et al. (2013), the authors have used two feature extraction techniques, namely, power curve fitting based features and parabola curve fitting based feature methods for the recognition of offline handwritten Gurmukhi characters. Bansal and Sinha (2001) have presented a complete Devanagari text recognition system. They considered real printed text written in Devanagari which consisted of a noisy environment and character fusions. The features used by them include a vertical bar feature, coverage of the region of the core strip, horizontal zero crossings, moments, number of positions of the vertex points and structural descriptors of the characters. For classification, they used decision tree classifier and achieved an accuracy of about 93.0%. Lehal and Singh (1999) worked for feature extraction and classification for OCR of Gurmukhi script. They developed two sets of features: Primary feature set and Secondary feature set. They used many structural features like: Presence of sidebar, Number of junctions with the headline, Presence of a loop, Number of endpoints and their location, Number and location of junctions, No Loop formed with headline etc. They used some other features also. They used nearest neighbours and binary tree classifier for classification and achieved 91.6% accuracy.

Jindal et al. (2008) presented work for feature extraction of degraded documents in Gurmukhi Script. They have discussed various structural features for recognition of degraded printed Gurmukhi characters and achieved 83.6% accuracy. Arora et al. (2007) presented a two stage classification approach for handwritten Devanagari characters. Structural properties like shirorekha, spine in character are used in the initial stage whereas some intersection features of characters got exploited in the second stage. This approach used feed forward neural network for classification. On the basis of differential distance, this work found a nearly straight line for spine and shirorekha. The approach gave 89.12% success. Sinha and Mahabala (1979) designed a syntactic pattern analysis system for Devanagari script recognition which worked by storing structural descriptors for every symbol in the script. The accuracy of 90% was achieved by that system. Lehal (2009) proposed a robust a font independent Gurmukhi OCR system. This work used four classifiers. First two classifiers are Binary Tree classifier and k-NN classifier. These two classifiers worked in serial mode and use structural features for feature extraction. Third classifier Support Vector Machine used Gabor filters with feature vector size 189. Fourth classifier was also Support Vector Machine and it operated on some structural and statistical features and achieved 98.18% accuracy. Holambe et al. (2011) presented an outline of feature extraction and selection methods for recognition of numerals & characters of the Devanagari script. They used

Zernike moment for feature extraction and Euclidian distance-based k-NN classifier. Yadav et al. (2013) proposed an OCR for printed Hindi recognition, using Artificial Neural Network (ANN). They used projection profiles for segmentation and histograms of projection based on mean distance, histogram of projection based on pixel value, vertical zero crossing for feature extraction and back-propagation neural network with two hidden layers for classification. They achieved a recognition accuracy of 90.0%.

Shao et al. (2014) presented restoration method for the character images which was applied for the recognition of unconstrained handwritten Chinese characters. For this work, the character image was exhibited as the blend of the ideal character image with two types of noise images: the added stroke noise image and the omitted stroke noise image. For preserving the original gradient features, restoration was done with the gradient features. Those features were then used to discriminate like characters. Katiyar and Mehfuz (2016) presented a hybridized recognition system for off-line handwritten characters by combining multiple features extracted by the use of seven different approaches. They used a Genetic Algorithm to optimize the total features along with adaptive Multi-Layer Perceptron classifier. A standard database of CEDAR (Centre of Excellence for Document Analysis and Recognition) on the English character set was used by them. They achieved accuracy of 94.65% for capital alphabets and 91.11% for small alphabets. DCT is another feature which has been applied widely in pattern recognition problems (Quacimy et al. 2014). Liu and Han (2007) proposed a fusion scheme on DCT filtered trace transformed face images. They calculated Trace transforms and DCT features. After extraction of the features, Support Vector Machine was used for testing and training purposes. Experimental results on benchmark face database revealed that the facial features proposed by them are robust and effective. They achieved 98.5% accuracy. Lawgali et al. (2011) have compared DCT to discrete wavelet transformation (DWT) on Arabic isolated character recognition, and the results predicted that DCT features lead to a better recognition rate. They achieved 79.87% accuracy with DCT features. Khodadad et al. (2011) used DCT features for online Arabic or Persian character recognition and achieved 95.69% accuracy. DCT is also used for video text detection (Ngo and Chan 2005), car-plate recognition (Parisi et al. 1998) and iris recognition (Monro et al. 2007).

Dalal and Triggs (2005) made a paradigm shift by introducing Histogram of Oriented Gradient (HOG) features instead of Eigenfaces. HOG features have many features like orientation binning, local contrast normalization and fine scale gradient which are necessary for good results. They achieved 89.0% accuracy with HOG features. Kobayashi et al. (2007) employed HOG features as candidate extracted from the locations presented in a grid on the image and have applied Principal Component Analysis to get the vectors. Linear SVM was used for the classification of pedestrian/non-pedestrian and accuracy of 99.3% was achieved. Aggarwal et al. (2012) used the gradient features for basic Devanagari character recognition collected from various writers. Their sample contained 7200 characters. They normalized sample images to 90×90 pixel sizes. Support Vector Machines (SVM) was used for classification. They obtained accuracy of 94%. A good number of other feature extraction techniques have been proposed for document analysis and recognition. Gabor filters have been used in different applications of pattern recognition. Dennis Gabor proposed Gabor function in 1946. Later, it was realized that Gabor filters give good results in pattern recognition applications. In late 1980s, Daugman (1980) introduced the use of the 2D Gabor filter in computer vision. Deng et al. (1994) used Gabor for Chinese handwritten character recognition. They anticipated a system by applying Gabor spatial filters with different spatial frequencies and directions to the character images. They used self-organizing map for clustering feature codes and a multi-staged linear vector quantization with a fuzzy

judgment for classification and achieved 94.2% accuracy. Singh et al. (2012) used Gabor filters for isolated handwritten Gurmukhi character recognition. They achieved 94.29% accuracy using 5-fold cross validation of the whole database with RBF-SVM classifier.

Ramanathan et al. (2009) proposed a novel and robust technique for feature extraction using Gabor filters. They used 2D Gabor filters. This technique extracted fifty features based on global texture analysis and achieved 92.5% accuracy. Rani et al. (2014) have used Gabor filter banks with k-NN, SVM and P-NN classifiers to identify the scripts at line level of tri-lingual documents. They achieved accuracy of 99.85% for script identification of tri-lingual documents. Scale Invariant Feature Transform (SIFT) is another type of features which are used for character recognition. Lowe (2004) proposed an algorithm for extracting features which can be used for reliable matching between different images. These features are scale and rotation invariant. He used these features for object recognition and achieved 98% accuracy. Diem and Sablatnig (2009) used binarization-free approach for segmentation of the characters from ancient Slavonic manuscripts using some local descriptors. SVM classifier was used for the purpose of classification, through which Scale Invariant Feature Transform (SIFT) features were extracted. Khanale and Chitnis (2011) presented a Devanagari character recognition system using artificial Neural Networks which used two layer feed forward network with 10 neurons for each layer. Transfer function was log-sigmoid. Training was done with back propagation. Performance function sum squared error was used. Results were obtained with up to 96% accuracy for some characters. For classification, ensemble methods are proposed in literature. Ensemble methods make use of more than one classifiers for classification. These classifiers can be used sequentially or in parallel. Singh and Lehal (2014) presented a comparative performance analysis of the feature(s)-classifier combination for Devanagari OCR. They extracted zoning, transition, Gabor filters, directional distance distribution, DCT, gradient and profile direction codes features. For classification, they used three classifiers namely support vector machines, artificial neural networks and k-nearest neighbors. They evaluated the performance by using the combination of these feature extraction approaches. They found that SVM trained with Gradient features provide the classification correctness of 99.43%.

Singh et al. (2015) presented a system for the recognition of handwritten Devanagari characters. They use gradient features with radial basis function neural network (RBF) neural network. RBF network with one input and one output layer has been performed to consider the recognition accuracy. Shelke and Apte (2015) presented a new approach for the classification of unconstrained handwritten Devanagari characters. They used two stages for classification, the first stage is based on fuzzy inference system and the second stage is based on structural parameters. The fuzzy system improves the classification over crisp classification. The classified characters are communicated to the feature extraction stage. The final stage implements feed forward neural network for character recognition. Acharya et al. (2015) published a new image dataset for Devanagari script, Devanagari Handwritten Character Dataset (DHCD). The dataset consisted of 92000 images of 46 unique classes of characters of Devanagari script segmented from handwritten files. Also, they proposed a deep learning architecture (CNN) for recognition of these characters. The proposed system scored the accuracy of 98.47% on DHCD dataset. Ghosh and Roy (2015) presented two zone-based feature extraction approaches for online handwritten character recognition of Bengali and Devanagari scripts. In the first approach, named zone wise structural and directional features (ZSD), structural and directional features are extracted for each stroke in each of these local zones. In the second approach, named zone wise slopes of dominant points (ZSDP), the dominant points are detected first from each stroke and next the slope angles between consecutive dominant points are calculated and features are extracted

in these local zones. SVM classifier is used for the recognition of these characters. They achieved maximum accuracy of 92.48% and 90.63% with ZSDP for Bengali and Devanagari scripts, respectively. Purkaystha et al. (2017) presented a CNN based OCR for Bengali Handwritten Character Recognition. Roohi and Alizadehashrafi developed a Handwritten Persian Character Recognition using Convolutional Neural Network. Jangid and Srivastava (2016) presented Handwritten Devnagari Character Recognition system. They used Deep Convolutional Neural Networks and Adaptive Gradient Methods for classification. Training a CNN requires a lot of computation and a large dataset. Until recently, training a CNN was not so effective. But, with the introduction of unsupervised pre-training phase, CNNs have become very effective. The role of unsupervised training is investigated by Erhan et al. (2010).

Khanduja et al. (2016) proposed a hybrid mechanism that combined the structural features of Devanagari character images and a mathematical model of curve fitting to get the better features. They achieved 93.4% accuracy. Kumar (2016) used multi-layer perceptron (MLP) network for handwritten Devanagari OCR. For characters belonging to the center region, every individual character is passed to a three phase classifier to know the correct class of character. For upper area and lower area characters, just single stage classifier has been trained. Dutta et al. (2018) released a new handwritten word dataset for Devanagari, IIIT-HW-Dev. They bench marked this dataset using a CNN-RNN hybrid architecture. They empirically showed that use of synthetic data and cross lingual transfer learning helps alleviate the issue of lack of training data. We used the proposed pipeline on a public dataset, RoyDB and achieved state of the art results.

Kim et al. (2017) proposed Adaboost method for classification. In general, Adaboosting algorithm initially assign same weight to each training data. Authors proposed to assign different weights to train datum based on some statistics. They used soft fuzzy decision with classification. Quo and Boukir (2017) exploited marginal theory for the selection of training data for bagging. They used iterative guided bagging algorithm for image classification exploiting low margin instances and obtained better results. Quo and Boukir (2014) used marginal theory for the design of better ensemble classifiers. They derived a new ensemble diversity method that revealed sources of diversity in image data. Ameta (2017) proposed EUSBoost, ensemble based classifier for the early stage detection of breast cancer. Image processing is used for this. EUSBoost classifier takes the benefits of both-boosting algorithm with random under sampling techniques. This method solves the problem of classification imbalance in which instances of one class outwork the instances of the other class. Earlier, many feature extraction techniques have been suggested in literature for image processing. Table 4 gives a summary of literature review for feature extraction and classification techniques.

3 Ancient text documents recognition

A huge amount of important information is stored in historical documents of almost each script all over the world. Digital preservation of historical document collections residing in libraries, museums and archives is very much required to provide access to massive community (Rani 2015). These documents are decaying due to age. These need to be preserved and made available to a mass community. Research is going on for the recognition of ancient documents. Following subsection gives the work done for the recognition of ancient documents.

Table 4 Brief summary of feature extraction and classification techniques

Citation	Feature extraction technique	Classification technique	Input pattern	Accuracy
Hussain et al (2015)	Zoning	ANN and SVM	Handwritten Assamese characters	–
Kimura and Shridhar (1991)	Static zoning	Combination of statistical and structural classifiers	Handwritten Numeric data	Very low error rates (0.2% or less) and rejection rates below 4%
Singh and Budhiraja (2011)	Zoning	kNN	Handwritten Gurmukhi text	72.5%
	Zoning	SVM	Handwritten Gurmukhi text	73.02%
	Zoning Density and background directional distribution features	SVM with RBF kernel	Handwritten Gurmukhi text	95.04%
Kumar et al. (2014)	Hierarchical zoning with horizontal peak extent, vertical peak extent, centroid and diagonal features	SVM classifier	Offline handwritten Gurmukhi characters	91.80%
Bansal and Sinha (2001)	Coverage of the region of the core strip, a vertical bar feature, horizontal zero crossings, number of positions of the vertex points, moments, structural descriptors of the characters	Decision tree classifier	Devanagari printed text	93%
Lehal and Singh (1999)	Number of junctions with the headline, Presence of sidebar, Presence of a loop, No Loop formed with headline, Number of endpoints and their location, Number of junctions and their location etc.	Binary tree classifier and nearest neighbours	Gurmukhi script text	91.6%
Jindal et al. (2008)	Structural features	kNN and SVM	Degraded documents in Gurmukhi Script	83.60%
Arora et al. (2007)	Structural and intersection features	Feed forward neural network	Handwritten Devanagari characters	89.12%
Sinha and Mahabala (1979)	Structural descriptors (syntactic pattern analysis)	Tree	Devanagari script recognition	90%

Table 4 (continued)

Citation	Feature extraction technique	Classification technique	Input pattern	Accuracy
Lehal (2009)	Structural features, statistical features, Gabor filters	Binary Tree classifier and kNN classifier and Support Vector machine	Gurmukhi script	98.18%
Holambe et al. (2011)	Zernike moment	Euclidian distance-based k-NN	Devanagari script	–
Yadav et al. (2013)	Histograms of projection based on mean distance, histogram of projection based on pixel value, vertical zero crossing	Back-propagation neural network with two hidden layers	Printed Hindi recognition	90.0%
Shao et al. (2014)	Gradient features	Extended MQDF (modified quadratic discriminant functions)	Unconstrained handwritten Chinese character	93.5%
Katiyar and Mehfuz (2016)	Multiple features with genetic algorithm	Multi-layer perceptron	Handwritten English characters	94.65% for capital alphabet and 91.11% for small alphabet
Lawgali et al. (2011)	DCT and DWT	ANN	Arabic isolated character	79.87%
Khodadad et al. (2011)	DCT features	Neural networks	Online Arabic/Persian character recognition	95.69%
Dalal and Triggs (2005)	HOG features	SVM	pedestrian database	89%
Kobayashi et al. 2007	HOG features	SVM	Pedestrian/non-pedestrian images	99.3%
Aggarwal et al. (2012)	Gradient	SVM	Devanagari character	94%
Deng et al. (1994)	Gabor filters	Self-organizing maps and multi-staged linear vector quantization (LVQ) with a fuzzy judgement	Chinese handwritten character recognition	94.2%
Singh et al. (2012)	Gabor filters	RBF-SVM classifier	Isolated handwritten Gurmukhi character recognition	94.29%
Ramanathan et al. (2009)	2D Gabor filters	SVM	Tamil text	92.5%
Rani et al. (2014)	Gabor filter	k-NN, SVM and P-NN	Tri-lingual documents	99.85%
Lowe (2004)	SIFT features	Nearest neighbour followed by Hough transform for clustering	Different images	98%
Diem and Sablatnig (2009)	SIFT and other local descriptors	SVM	Ancient Slavonic manuscripts	–

Table 4 (continued)

Citation	Feature extraction technique	Classification technique	Input pattern	Accuracy
Khanale and Chitnis (2011)	–	Artificial neural networks (two layer Feed forward network)	Devanagari Characters	96%
Singh and Lehal (2014)	Zoning, transition, gabor filters, directional distance distribution, DCT, Gradient, profile direction codes	SVM, ANN, kNN	Devanagari Characters	99.429%
Acharya et al. (2015)	CNN	CNN	Devanagari Characters	98.47%
Gosh and Roy (2015)	Zone wise structural and directional features, Zone wise slopes of dominant points	SVM	Devanagari Characters	92.48%
Khanduja et al. (2016)	Structural features and mathematical model of curve fitting		Devanagari Characters	93.4%
Jangid and Srivastava (2016)	Gradient features		Devanagari Characters	95.94%

3.1 Line segmentation in ancient text documents

One of the major challenges of ancient manuscripts recognition is text line segmentation. Because of many distinct features of ancient documents (uneven space between neighbouring lines, touching or overlapping lines), text line segmentation is very difficult. Text line segmentation methods for contemporary documents involve printing of text along straight baselines whereas in the case of historical documents baselines are frequently changed in random fashion (Clausner et al. 2012). Sulem et al. (2006) presented a study of line segmentation methods presently used for historical documents. Six line segmentation techniques presented by them are: Projection profiles, Repulsive-attractive network, Grouping, Smearing methods, Hough based methods, and Stochastic methods. Sulem et al. (2006) presented a survey on text line segmentation of historical documents. They discussed various techniques used for pre-processing and line segmentation of historical documents. Feldbach and Tonnies (2001) proposed a method for detection and segmentation of lines in historical registers of the church. The basis of this method involved the detection of local minima of connected components. It was then applied to represent chain code of the connected components. Gradual construction of line segments was carried out until a unique text line was formed. This algorithm worked well to segment text lines which were close to each other, the lines that touched each other and also for fluctuating text lines. Jindal and Lehal (2012) discussed method for segmenting lines for Gurmukhi handwritten ancient manuscripts. They divided the document into non-overlapping vertical stripes. For each stripe, they projected all black pixels on the Y-axis and selected positions whose number of accumulative pixels is minimal. Further the text blocks have been divided into three categories, Small Text Blocks (STB) containing upper zone or lower zone characters or some part of middle zone, Average Text Blocks (ATB) containing middle zone along with upper and/or lower zone, Large Text Blocks (LTB) containing overlapping lines. Then they segmented large text blocks into lines using average size. Panichkriangkrai et al. (2013) have proposed a system for line and text extraction for Japanese historical woodblock printed books. Vertical projection was used on binarized images for separating text lines. Adaptive binarization on the gray scale images was applied to extract the connected components. The connected components were then split up or merged by applying Rule-based integration.

Chen et al. (2016) have presented a Conditional Random Field (CRF) model to segment handwritten historical document images into different regions. Page segmentation was considered as a pixel-labeling problem. Features were learnt from pixel intensity values with stacked convolutional auto-encoders in an unsupervised manner. Then a CRF model was introduced to improve the segmentation. Rao et al. (2015) have implemented a hybrid model that involves segmentation in noisy images, which was then followed by binarization for segmentation of ancient Telugu documents. First phase involved the convolution of horizontal profile pattern using a Gaussian kernel. An extensive analysis of the geometrical patterns of meaningful units was carried out to explore the statistical properties of the meaningful units. Second phase involved cleaning of noisy documents using a modified IGT algorithm. Then segmentation was performed using the conventional profile mechanism. They obtained maximum accuracy of 95.59% for the cleaned story books. Kleber et al. (2008) presented a method to extrapolate missing parts of a degraded ancient document based on a priori knowledge. The paper introduced an algorithm for ruling estimation of glagolitic texts based on text line extraction and is suitable for the manuscripts present in degraded form. For line extraction, connected component based approach was

used. Bar-Yosef et al. (2009) proposed a novel approach based on adaptive local projection techniques for segmenting text lines in degraded documents having large skew in text lines. Local algorithm was applied in an incremental manner which was adaptable to the skewed lines in the text. They have achieved accurate results on a sample of degraded documents with lines having different curvatures and skew angles. Gatos et al. (2014) presented a novel segmentation module for text zone as well as text line detection for handwritten ancient documents to handle several challenging cases such as horizontal and vertical rule lines overlapping with the text, two column documents and characters of different text lines touching vertically. Table 5 gives a summary of line segmentation review for ancient documents.

3.2 Feature extraction and classification for ancient text documents recognition

Kim et al. (2004) presented a dedicated OCR system for Hanja Historical documents. Korean name given to Chinese characters is Hanza. These characters are incorporated into the Korean language with Korean pronunciation. They set 2568 classes for character recognition out of the 5599 classes which were found in vol. 29 of the Seungjungwon Diary. Sousa et al. (2005) proposed a system based on fuzzy logic for optical character recognition for ancient documents in the printed form. The proposed OCR system built fuzzy membership functions from oriented features extracted using Gabor filter banks. This work gave a success rate of 88%. Diem and Sablatnig (2009) presented a work to recognize degraded characters using local features. This work was done on an ancient manuscript where character were washed out (partially visible) due to age. Due to washed out characters, it was not suitable for binarization. So an approach based on local descriptors was developed, the approach remained free from segmentation. Support Vector Machine was used to classify Local descriptors and then identified by a voting scheme of neighboring local descriptors. Phan et al. (2016) presented Nom historical document recognition system. In this work, recursive X–Y cut and Voronoi diagrams for segmentation was used by the authors. They extracted gradient features. They used generalized learning vector quantization and k–d tree for classification. Further, for more fine classification, the modified quadratic discriminant function (MQDF) was used. Garz et al. (2011) proposed a robust method to separate text area from decorative area in ancient documents using scale Invariant feature transforms (SIFT) and local descriptors. Cecotti and Belaid (2005) have presented a hybrid approach which was accompanied by some dedicated ICR for ancient documents. A model was further proposed which combined several OCRs, and some specific intelligent character recognition based on convolutional neural network. Sumetphong and Tangwongsan (2012) proposed a novel solution to recognize broken characters in Thai Historical documents based on set-partitions. Daubechies Wavelets were used for feature extraction and further to reduce the large feature space, standard Principal Component Analysis (PCA) was applied. They trained an ensemble of Neural Nets, one for characters on each zone of the Thai language for classification. Kavitha et al. (2013) proposed two approaches for historical document classification. First approach (SA) is based on skewness for Indus document classification from English and South Indian scripts. Nearest Neighbour based Approach (NNA) was then applied to classify English from South Indian scripts. The SA approach compared the skewness between the components in the Indus document image with respect to x-axis with the skewness between the components in English and South Indian documents and found the former was greater than the later. The Neighbour based

Table 5 A few work presented for line segmentation of ancient text documents

Line segmentation technique	Text type	Citations
Local minima detection of connected components	Touching text lines and fluctuating text lines	Feldbach and Tonnies (2001)
Piecewise projection profile	Gurmukhi handwritten ancient manuscripts	Jindal and Lehal (2012)
Vertical projection	Japanese historical woodblock printed books	Panichkriangkrai et al. (2013)
Conditional random field (crf)	Handwritten historical document images into different regions	Chen et al. (2016)
Gaussian kernel and IGT algorithm	Ancient Telugu documents	Rao et al. (2015)
Connected component based approach	Degraded ancient document	Kleber et al. (2008)
A novel approach	Set of degraded documents	Bar-Yosef et al. (2009)

Approach to identify the existence or nonexistence of the modifiers which are common in South Indian document images and are not present in English document images.

Soumya and Kumar (2015) proposed the acknowledgment of text in old Kannada script of Ashoka and Hoysala era. Segmentation of characters was performed using Nearest Neighbor clustering algorithm. After the segmentation process, statistical features such as Homogeneousness, Standard Deviation, Mean, Correlation, Variance, Kurtosis, Skewness, Contrast, Energy, and Coarseness were extracted. Mamdani Fuzzy Classifier was used in the classification of characters. In the final phase, classified characters of early time were exhibited in new Kannada form. Recognition rate for Brahmi script was found to be appreciable in comparison to Hoysala script. Avadesh and Goyal (2018) proposed a CNN based OCR for the recognition of printed ancient Sanskrit manuscripts. They calculated pixel intensities to identify letters in the image. They considered typical compound characters (half letter combinations) as separate classes. They achieved 93.32% accuracy. Narang et al. (2018) used DCT and HOG features with Naïve Bayes, decision tree and SVM classifiers. They obtained maximum accuracy of 90.70% accuracy with DCT features and SVM classifier on the same database. Narang et al. (2019) used adaptive boosting and bootstrap aggregating methodologies for improvement in recognition of Devanagari ancient manuscripts. They used discrete cosine transform (DCT) zigzag for feature extraction. Decision tree, Naïve Bayes and support vector machine classifiers were used for the recognition of basic characters segmented from Devanagari ancient manuscripts. Maximum accuracy of 91.70% was achieved for adaptive boosting (AdaBoost) with RBF-SVM.

4 Inferences and future directions

OCR is the process of identifying characters from printed or handwritten documents. In this article, authors have presented a comprehensive survey of the reported work, with the segmentation techniques used, feature extraction methods used, classifiers and the accuracy achieved with special emphasis on ancient manuscripts. This provides the abstract view of the reader towards the nature of documents, various segmentation techniques, features extraction methods and the classification methods. From the survey, authors try to establish that research for OCR of ancient manuscripts is still in an infant stage. The paper presents a collection of good findings for different stages of the OCR process especially in ancient manuscripts. The focus of this paper is to present various approaches and techniques for segmentation, feature extraction and classification stages of an OCR process. This study aims at helping novice researchers to be acquainted with various existing approaches and techniques. This study provides the existing approaches and techniques with their probable application areas in a nutshell. This paper shows the best accuracies obtained by using different feature extraction and classification techniques. The lack of a standard database on ancient manuscripts is a major concern in this respect. Synthesis analysis and the comparative study of the bases of the reported work are also mentioned after the reported work. Through paper, we also present the summarized list of segmentation, features and classifiers. Development of new algorithms for segmentation of ancient documents can be contemplated as a future direction. Also, novel approaches for feature extraction and classification need to be developed for achieving the best accuracy results.

Compliance with ethical standards

Conflict of interest Authors have declared that they have no conflict on interest in this work.

References

Acharya S, Pant AK, Gyawali PK (2015) Deep learning based large scale handwritten Devanagari character recognition. In: Proceedings of the 9th international conference on software, knowledge, information management and applications (SKIMA), pp 1–6

Adiguzel H, Sahin E, Dugulu P (2012) A hybrid approach for line segmentation in handwritten documents. In: Proceedings of the international conference on frontiers in handwriting recognition (ICFHR), pp 503–508

Aggarwal A, Rani R, Dhir R (2012) Handwritten Devanagari character recognition using gradient features. Int J Adv Res Comput Sci Softw Eng 2(5):85–90

Alaei A, Nagabhushan P, Pal U (2011) Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents. Pattern Anal Appl 14(4):381–394

Alam M, Kashem AM (2010) A complete Bangla OCR system for printed characters. J Cases Inf Technol 1(1):30–35

Alizadehashraf B, Roohi S (2017) Persian handwritten character recognition using convolutional neural network. In: Proceedings of the 10th Iranian conference on machine vision and image processing, pp 247–251

Almazan EJ, Tal R, Qian Y, Elder JH (2017) MCMLSD: a dynamic programming approach to line segment detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5854–5862

Ameta D (2017) Ensemble classifier approach in breast cancer detection and malignancy grading—a review. Int J Manag Public Sector Inf Commun Technol (IJMPICT) 8(1):17–26

Angadi SA, Kodabagi MM (2014) A robust segmentation technique for line, word and character extraction from Kannada text in low resolution display board images. In: Proceedings of the fifth international conference on signals and image processing, pp 42–49

Arivazhagan M, Srinivasan H, Srihari S (2007) A statistical approach to line segmentation in handwritten documents. In: Proceedings of SPIE document recognition and retrieval, pp 1–11

Arora S, Bhatcharjee D, Nasipuri M, Malik L (2007) A two stage classification approach for handwritten Devnagari characters. In: Proceedings of the international conference on computational intelligence and multimedia applications (ICCIMA 2007), Sivakasi, Tamil Nadu, pp 399–403

Arvanitopoulos N, Süsstrunk S (2014) Seam carving for text line extraction on color and grayscale historical manuscripts. In: Proceedings of the 14th international conference on frontiers in handwriting recognition (ICFHR-14), pp 721–726

Avadesh M, Goyal N (2018) Optical character recognition for sanskrit using convolution neural networks. In: Proceedings of the 13th IAPR international workshop on document analysis systems (DAS), Vienna, pp 447–452

Bag S, Harit G (2013) A survey on optical character recognition for Bangla and Devnagari scripts. Sadhana 38(1):133–168

Bag S, Krishna A (2015) Character segmentation of hindi unconstrained handwritten words. Proc 17th Int Workshop Comb Image. Anal 9448:247–260

Bansal V, Sinha RMK (2001) A complete OCR for printed Hindi text in Devanagari script. In: Proceedings of the 6th international conference on document analysis and recognition, pp 800–804

Bansal V, Sinha R, Kumar MK (2002) Segmentation of touching and fused Devanagari characters. Pattern Recognit 35(4):875–893

Bar-Yosef A, Hagbi N, Kedem K, Dinstein I (2009) Line segmentation for degraded handwritten historical documents. In: Proceedings of the 10th international conference on document analysis and recognition, Barcelona, pp 1161–1165

Basu S, Chaudhuri C, Kundu M, Nasipuri M, Basu DK (2007) Text line extraction from multi-skewed handwritten documents. Pattern Recognit 40(6):1825–1839

Beare R (2006) A locally constrained watershed transform. IEEE Trans Pattern Anal Mach Intell 28(7):1063–1074

Belaid A (1997) OCR print - an overview. In: Survey of the state of the art in human language technology, pp 71–74

Bhopi SA, Singh MP (2018) Review on optical character recognition of Devanagari script using neural network. Int J Future Revolut Comput Sci Commun Eng 4(3):415–420
Boiangiu CA, Tanase MC, Ioanitescu R (2014) Handwritten documents text line segmentation based on information energy. Int J Comput Commun Control 9(1):8–15
Brodic D (2012) Extended approach to water flow algorithm for text line segmentation. J Comput Sci Technol 27(1):187–194
Brodic D (2015) Text line segmentation with water flow algorithm based on power function. J Electr Eng 66(3):132–141
Casey RG, Lecolinet E (1996) A survey of methods and strategies in character segmentation. IEEE Trans Pattern Anal Mach Intell 18(7):690–706
Cecotti H, Belaid A(2005) Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. In: Proceedings of the 8th international conference on document analysis and recognition, vol 2, pp 1045–1049
Chen YK, Wang JF (2000) Segmentation of single or multiple-touching handwritten numeral string using background and foreground analysis. IEEE Trans Pattern Anal Mach Intell 22(11):1304–1317
Chen K, Seuret M, Liwicki M, Hennebert J, Liu C, Ingold R (2016) Page segmentation for historical handwritten document images using conditional random fields. In: Proceedings of the 15th international conference on frontiers in handwriting recognition (ICFHR), Shenzhen, pp 90–95
Clausner C, Antonacopoulos A, Pletschacher S (2012) A robust hybrid approach for text line segmentation in historical documents. In: Proceedings of the 21st international conference on pattern recognition (ICPR), pp 335–338
Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05), vol 1, pp 886–893
Daugman JG (1980) Two-dimensional spectral analysis of cortical receptive field profiles. Vision Res 20(10):847–856
Deng D, Chan KP, Yu Y (1994) Handwritten Chinese character recognition using spatial Gabor filters and self-organizing feature maps. In: Proceedings of the international conference on image processing, vol 3, pp 940–944
Devijver PA, Kittler J (1985) Pattern recognition: a statistical approach. Image Vis Comput 3(2):87–88
Diem M, Sablatnig R (2009) Recognition of degraded handwritten characters using local features. In: Proceedings of the 10th international conference on document analysis and recognition, pp 221–225
Diem M, Sablatnig R (2010) Recognizing characters of ancient manuscripts. In: Proceedings of the international conference on computer image analysis in the study of art, pp 753106–753112
Din I, Malik Z, Siddiqi I, Khalid S (2016) Line and ligature segmentation in printed Urdu document images. J Appl Environ Biol Sci 6(3S):114–120
Ding X, Li Y, Belatreche A, Maguire L (2012) Constructing minimum volume surfaces using level set methods for novelty detection. In: Proceedings of the 2012 international joint conference on neural networks (IJCNN), Brisbane, QLD, pp 1–6
Dongre VJ, Mankar VH (2010) A review of research on Devanagari character recognition. Int J Comput Appl 12(2):8–15
Dunn CE, Wang PSP (1992) Character segmentation techniques for handwritten text-a survey. In: Proceedings of the 11th international conference on recognition methodology and systems, vol 2, pp 577–580
Dutta K, Krishnan P, Mathew M, Jawahar CV (2018) Offline handwriting recognition on Devanagari using a new benchmark dataset. In: Proceedings of the 13th IAPR international workshop on document analysis systems (DAS), Vienna, pp 25–30
Erhan D, Bengio Y, Courville A, Manzagol P, Vincent P, Bengio S (2010) Why does unsupervised pretraining help deep learning? J Mach Learn Res 11:625–660
Feldbach M, Tonnies KD (2001) Line detection and segmentation in historical church registers. In: Proceedings of the 6th international conference on document analysis and recognition, pp 743–747
Fragkou P, Petridis V, Kehagias A (2004) A dynamic programming algorithm for linear text segmentation. J Intell Inf Syst 23(2):179–197
Fujisawa H, Nakano Y, Kurino K (1992) Segmentation methods for character recognition: from segmentation to document structure analysis. Proc IEEE 80(7):1079–1092
Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, New York
Garz A, Sablatnig R, Diem M (2011) Using local features for efficient layout analysis of ancient manuscripts. In: Proceedings of the 19th European Signal Processing Conference (EUSIPCO) Barcelona, Spain, pp 1259–1263

Gatos B, Louloudis G, Stamatopoulos N (2014) Segmentation of historical handwritten documents into text zones and text lines. In: Proceedings of the international conference on frontiers in handwriting recognition, ICFHR, pp 464–469. https://doi.org/10.1109/ICFHR.2014.84

Ghosh R, Roy PP (2015) Study of two zone-based features for online Bengali and Devanagari character recognition. In: Proceedings of the 13th international conference on document analysis and recognition (ICDAR), pp 401–405

Gomathi R, Uma RS, Mohanval S (2012) Segmentation of touching, overlapping, skewed and short handwritten text lines. Int J Comput Appl 49:24–27

Gonzalez RC, Woods RE (1992) Digital image processing. Prentice-Hall, NewDelhi

Holambe N, Thool RC, Jagade SM (2011) A brief review and survey of feature extraction methods for Devnagari OCR. In: Proceedings of the 9th international conference on ICT and knowledge engineering, pp 99–104

Hussain E, Hannan A, Kashyap K (2015) A zoning based feature extraction method for recognition of handwritten assamese characters. Int J Comput Sci Technol 6(2):226–228

Jangid M, Srivastava S (2016) Accuracy enhancement of devanagari character recognition by gray level normalization. In: Proceedings of the 7th international conference on computing communication and networking technologies ACM, p 25

Jetley S, Belhe S, Koppula VK, Negi A (2012) Two-stage hybrid binarization around fringe map based text line segmentation for document images. In: Proceedings of the international conference on pattern recognition, pp 343–346

Jindal S, Lehal GS (2012) Line segmentation of handwritten Gurmukhi manuscripts. In: Proceedings of the document analysis and recognition, Mumbai, IN, India Copyright 2012 ACM, pp 74–78

Jindal MK, Sharma RK, Lehal GS (2007) Segmentation of horizontally overlapping lines in printed Indian scripts. Int J Comput Intell Res 3(4):277–286

Jindal MK, Sharma DV, Lehal GS (2008) Structural features for recognising degraded printed Gurmukhi script. In: Proceedings of the 5th international conference on information technology, pp 668–673

Jindal M, Lehal G, Sharma RK (2009) On segmentation of touching characters and overlapping lines in degraded printed gurmukhi script. Int J Image Graph 9:321–353

Jindal MK, Garg NK, Kaur L (2010) Segmentation of handwritten Hindi text. Int J Comput Appl 1:19–23

Katiyar G, Mehfuz S (2016) A hybrid recognition system for off-line handwritten characters. SpringerPlus 5:1–18

Kavitha AS, Shivakumara P, Hemantha G (2013) Skewness and nearest neighbour based approach for historical document classification. In: Proceedings of the international conference on communication systems and network technologies, pp 602–606

Kennard DJ, Barrett WA (2006) Separating lines of text in free-form handwritten historical documents. In: Proceedings of the 2nd international conference on document image analysis for libraries (DIAL-06), pp 12–23

Khanale PB, Chitnis SD (2011) Handwritten Devanagari character recognition using artificial neural network. J Artif Intell 4(1):55–62

Khanduja D, Nain N, Panwar S (2016) A hybrid feature extraction algorithm for Devanagari script. ACM Trans Asian Low-Resour Lang Inf Process 15(1):2

Khodadad M, Sid-Ahmed E, Raheem A (2011) Online Arabic/Persian character recognition using neural network classifier and DCT features. In: Proceedings of the 54th international Midwest symposium on circuits and systems, pp 1–4

Kim KK, Kim JH, Suen CY (2000) Recognition of unconstrained handwritten numeral strings by composite segmentation method. In: Proceedings of the 15th international conference on pattern recognition, pp 594–597

Kim MS, Jang MD, Choi HL, Rhee TH, Kim JH, Kwag HK (2004) Digitalizing scheme of handwritten Hanja historical documents. In: Proceedings of the first international workshop on document image analysis for libraries, pp 321–327

Kim K, Choi H, Oh K (2017) Object Detection using ensemble of linear classifiers with fuzzy adaptive boosting. EURASIP J Image Video Process 17:40

Kimura F, Shridhar M (1991) Handwritten numerical recognition based on multiple algorithms. Pattern Recognit 24(10):969–983

Kleber F, Sablatnig R, Gau M, Miklas H (2008) Ancient document analysis based on text line extraction. In: Proceedings of the 19th international conference on pattern recognition, pp 1–4

Kobayashi T, Hidaka A, Kurita T (2007) Selection of histograms of oriented gradients features for pedestrian detection. In: Proceedings of the international conference on neural information processing, pp 598–607

Koppula VK, Negi A (2011) Fringe map based text line segmentation of printed Telugu document images. In: Proceedings of the international conference on document analysis and recognition (ICDAR-11), pp 1294–1298
Kumar S (2016) A study for handwritten Devanagari word recognition. In: Proceedings of the international conference on communication and signal processing (ICCSP), pp 1009–1014
Kumar D, Gupta D (2018) Review on optical character recognition for off-line Devanagari handwritten characters & challenges. Int J Sci Res Comput Sci Eng Inf Technol 3(3):1364–1367
Kumar KSS, Namboodiri AM, Jawahar CV (2006) Learning segmentation of documents with complex scripts. In: Fifth Indian conference on computer vision, graphics and image processing, Madurai, India, pp 749–760
Kumar M, Jindal MK, Sharma RK (2012) Offline handwritten Gurmukhi character recognition: study of different features and classifiers combinations. In: Proceedings of the international workshop on document analysis and recognition, IIT Bombay, pp 94–99
Kumar M, Sharma RK, Jindal MK (2013) A novel feature extraction technique for offline handwritten Gurmukhi character recognition. IETE J Res 59(6):687–692
Kumar M, Jindal MK, Sharma RK (2014) A novel hierarchical techniques for offline handwritten Gurmukhi character recognition. Natl Acad Sci Lett 37(6):567–572
Kumar M, Sharma RK, Jindal MK (2018) Character and numeral recognition for non-Indic and Indic scripts: a survey. Artif Intell Rev 52:2235–2261
Lawgali A, Bouridane A, Angelova M, Ghassemlooy Z (2011) Handwritten Arabic character recognition: which feature extraction method. Int J Adv Sci Technol 34:1–8
Lebourgeois F (1997) Robust multi-font OCR system from gray level images. In: Proceedings of the international conference on document analysis and recognition, vol 1, pp 1–5
Lehal GS (2009) Optical character recognition of Gurmukhi script using multiple classifiers. In: Proceedings of the international workshop on multilingual OCR, p 7
Lehal GS, Dhir R (1999) A range free skew detection technique for digitized Gurmukhi script documents. In: Proceedings of the fifth international conference on document analysis and recognition, pp 147–152
Lehal GS, Singh C (1999) Feature extraction and classification for OCR of Gurmukhi script. Vivek 12(2):2–12
Likforman-Sulem L, Hanimyan A, Faure C (1995) A Hough based algorithm for extracting text lines in handwritten documents. In: Proceedings of the 3rd international conference on document analysis and recognition, Montreal, Canada, vol 2, pp 774–777
Liu N, Han W (2007) Recognition of human faces using discrete cosine transform filtered trace feature. In: Proceedings of the 6th international conference on information, communications & signal processing (ICICS), pp 1–5
Liwicki M, Indermuhle E, Bunke H (2007) On-line handwritten text line detection using dynamic programming. In: Proceedings of the 9th international conference on document analysis and recognition (ICDAR 07), vol 1, pp 447–451
Louloudis G, Gatos B, Pratikakis I, Halatsis C (2009) Text line and word segmentation of handwritten documents. Pattern Recognit 42(12):3169–3183
Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
Mahadevan U, Nagabushnam RC (1995) Gap metrics for word separation in handwritten lines. In: Proceedings of the 3rd international conference on document analysis and recognition (ICDAR-95), pp 124–127
Manjusha k, Kumar S, Rajendran J, Soman KP (2012) Hindi character segmentation in document images using level set methods and non-linear diffusion. Int J Comput Appl 44(16):42–47
Manmatha R, Rothfeder JL (2005) A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Trans Pattern Anal Mach Intell 27(8):1212–1225
Mantas J (1986) An overview of character recognition methodologies. Pattern Recognit 19(6):425–430
Messaoud B, El-Abed H, Amiri H, Margner V (2012) A multilevel textline segmentation framework for handwritten historical documents. In: Proceedings of the international conference on frontiers of handwriting recognition (ICFHR-12), pp 513–518
Monro DM, Rakshit S, Zhang D (2007) DCT-based iris recognition. IEEE Trans Pattern Anal Mach Intell 29(4):586–595
Narang SR, Jindal MK (2018) Issues in Devanagari ancient character recognition: a study. J Adv Sch Res Allied Educ 15(10):6–11

Narang SR, Jindal MK, Sharma P (2018) Devanagari ancient character recognition using HOG and DCT features. In: Proceedings of the 5th IEEE international conference on parallel, distributed and grid computing (PDGC-2018), Solan, India
Narang SR, Jindal MK, Kumar M (2019) Devanagari ancient character recognition using DCT features with adaptive boosting and bootstrap aggregating. Soft Comput 23:13603–13614
Ngo W, Chan CK (2005) Video text detection and segmentation for optical character recognition. Multimed Syst 10(3):261–272
Nikolaou N, Makridis M, Gatos B, Stamatopoulos N, Papamarkos N (2010) Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths. Image Vis Comput 28(4):590–604
O'Gorman L (1993) The document spectrum for page layout analysis. IEEE Trans Pattern Anal Mach Intell 15(11):1162–1173
Oliveira LS, Lethelier E, Bortolozzi F, Sabourin R (2000) A new segmentation approach for handwritten digits. In: Proceedings of the 15th international conference on pattern recognition, vol 2, pp 323–326
Pal U, Chaudhuri BB (2004) Indian script character recognition: a survey. Pattern Recognit 37(9):1887–1899
Pal U, Datta S (2003) Segmentation of Bangla unconstrained handwritten text. In: Proceedings of the international conference on document analysis and recognition, pp 1128–1132
Pal U, Belaïd A, Choisy C (2003) Touching numeral segmentation using water reservoir concept. Pattern Recognit Lett 24:261–272
Panichkriangkrai C, Li L, Hachimura K (2013) Character segmentation and retrieval for learning support system of Japanese historical books. In: Proceedings of the ACM international conference Proceeding series, pp 118–122. https://doi.org/10.1145/2501115.2501129
Parisi R, Claudi ED, Lucarelli G, Orlandi G (1998) Car plate recognition by neural networks and image processing. In: Proceedings of the IEEE international symposium on circuits and systems, vol 3, pp 195–198
Phan TV, Nguyen KC, Nakagawa M (2016) A Nom historical document recognition system for digital archiving. Int J Doc Anal Recognit 19(1):49–64
Phillips CL (1999) The level set method. MIT Undergrad J Math 1:155–164
Ptak R, Żygadło B, Unold O (2017) Projection-based text line segmentation with a variable threshold. Int J Appl Math Comput Sci 27(1):195–206
Purkaystha B, Datta T, Islam MS (2017) Bengali handwritten character recognition using deep convolutional neural network. In: Proceedings of the 20th international conference of computer and information technology (ICCIT), pp 1–5
Quacimy E, Kerroum MA, Hammouch A (2014) Feature extraction based on DCT for handwritten digit recognition. Int J Comput Sci Issues 11(6):27–33
Quo L, Boukir S (2014) Ensemble margin framework for image classification. In: Proceedings of the IEEE international conference on image processing, France, pp 4231–4235
Quo L, Boukir S (2017) Building an ensemble classifier using ensemble margin. Application to image classification. In: Proceedings of the 2017 IEEE international conference on image processing, Beijing, pp 4492–4496
Ramanathan R, Ponmathavan S, Thaneshwaran L, Nair AS, Valliappan N, Soman KP (2009) Tamil font recognition using gabor filters and support vector machines. In: Proceedings of the international conference on advances in computing, control, and telecommunication technologies, Trivandrum, Kerala, pp 613–615
Rani S (2015) Recognition of Gurmukhi handwritten manuscripts. Ph.D. thesis, Punjabi University, Patiala, India
Rani R, Dhir R, Lehal GS (2014) Gabor features based script identification of lines within a bilingual/trilingual document. Int J Adv Sci Technol 66:1–12
Rao VN, Sastry ASCS, Chakravarthy A, SrinivasaRao AV (2015) Analysis of canonical character segmentation technique for ancient Telugu text documents. J Theor Appl Inf Technol 82(2):311–320
Razak Z, Zulkiflee K, Idris MYI, Tamil EM, Noorzaily M (2008) Off-line handwriting text line segmentation: a review. Int J Comput Sci Netw Secur 8(7):12–20
Reddy LP, Babu TR, Rao, NV &Babu BR (2010) Touching syllable segmentation using split profile algorithm. Int J Comput Sci 7(3):1–10
Saabni R, Asi A, El-Sana J (2014) Text line extraction for historical document images. Pattern Recognit Lett 35:23–33
Saha S, Basu S, Nasipuri M, Basu DK (2010) A Hough transform based technique for text segmentation. J Comput 2(2):134–141

Sarkar R, Moulik S, Das N, Basu S, Nasipuri M, Kundu M (2011) Suppression of non-text components in handwritten document images. In: Proceedings of the international conference on image and information processing, pp 1–7
Sesh Kumar KS, Namboodiri AM, Jawahar CV (2006) Learning segmentation of documents with complex scripts. In: Proceedings of the fifth Indian conference on computer vision, graphics and image processing, Madurai, India, pp 749–760
Shah KR, Badgujar DD (2013) Devnagari handwritten character recognition (DHCR) for ancient documents: a review. In: Proceedings of IEEE conference on information and communication technology, pp 656–660
Shahi M, Ahlawat A, Pandey BN (2012) Literature survey on offline recognition of handwritten Hindi curve script using ANN approach. Int J Sci Res Publ 2(5):1–6
Shao Y, Wang C, Xiao B (2014) A character image restoration method for unconstrained handwritten Chinese character recognition. Int J Doc Anal Recognit 18(1):73–86
Shapiro VA (1993) From Radon to Hough transform of gray-scale images via digital halftoning. In: Proceedings of the 8th Scandinavian conference on image analysis, pp 665–672
Sharma DV, Lehal GS (2006) An iterative algorithm for segmentation of isolated handwritten words in Gurmukhi script. In: Proceedings of the 18th international conference on pattern recognition (ICPR'06), pp 1022–1025
Sharma N, Patnaik T, Kumar B (2013) Recognition for handwritten English letters: a review. Int J Eng Innov Technol 2(7):318–321
Shelke S, Apte S (2015) A fuzzy-based classification scheme for unconstrained handwritten Devanagari character recognition. In: Proceedings of the international conference on communication, information & computing technology (ICCICT), pp 1–6
Shi Z, Govindaraju V (2004) Line separation for complex document images using fuzzy runlength. In: Proceedings of the international workshop on document image analysis for libraries, p 306
Shi Z, Setlur S, Govindaraju V (2005) Text extraction from gray scale historical document images using adaptive local connectivity map. In: Proceedings of the international conference on document analysis and recognition (ICDAR), vol 2, pp 794–798
Singh P, Budhiraja S (2011) Feature extraction and classification techniques in OCR systems for handwritten Gurmukhi script- a survey. Int J Eng Res Appl 1(4):1736–1739
Singh J, Lehal GS (2014) Comparative performance analysis of feature (S)-classifier combination for Devanagari optical character recognition system. Int J Adv Comput Sci Appl 5(6):37–42
Singh S, Aggarwal A, Dhir R (2012) Use of Gabor filters for recognition of handwritten Gurmukhi character. Int J Adv Res Comput Sci Softw Eng 2(5):234–240
Singh D, Saini JP, Chauhan DS (2015) Hindi character recognition using RBF neural network and directional group feature extraction technique. In: Proceedings of the international conference on cognitive computing and information processing (CCIP), pp 1–4
Singh PK, Sarkar R, Nasipuri M (2016) A study of moment based features on handwritten digit recognition. Appl Comput Intell Soft Comput, Article ID 2796863
Sinha RMK, Mahabala HN (1979) Machine recognition of Devanagari script. IEEE Trans Syst Man Cybern 9(8):435–441
Souhar A, Boulid Y, Ameur EB, Ouagague MM (2017) Watershed transform for text lines extraction on binary Arabic handwritten documents. In: Proceedings of the 2nd international conference on big data, cloud and applications (BDCA'17). ACM, New York. https://doi.org/10.1145/3090354.3090444
Soumya A, Kumar HG (2015) Feature extraction and recognition of ancient Kannada epigraphs. Smart Innov Syst Technol 33:469–478
Sousa JMC, Pinto JRC, Ribeiro CS, Gil JM (2005) Ancient document recognition using fuzzy methods. In: Proceedings of the IEEE international conference on fuzzy systems, pp 833–836
Sridevi N, Subashini P (2012) Segmentation of text lines and characters in ancient tamil script documents using computational intelligence techniques. Int J Comput Appl 52(14):7–12
Sulem LL, Zahour A, Taconet B (2006) Text line segmentation of historical documents: a survey. Int J Doc Anal Recognit 9:123–138
Sumetphong C, Tangwongsan S (2012) An optimal approach towards recognizing broken Thai characters in OCR systems. In: Proceedings of the international conference on digital image computing techniques and applications (DICTA), pp 1–5
Trier OD, Jain AK, Taxt T (1996) Feature extraction methods for character recognition – a survey. Pattern Recognit 29(4):641–642
Tripathy N, Pal U (2004) Handwriting segmentation of unconstrained Oriya text. In: Proceedings of the international workshop on frontiers in handwriting recognition, pp 306–311
Tripathy N, Pal U (2006) Handwriting segmentation of unconstrained Oriya text. Sadhana 31:755–769

Tseng YH, Lee HJ (1999) Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recognit Lett 20(8):791–806
Verma R, Ali Z (2012) A survey of feature extraction and classification techniques in OCR systems. Int J Comput Appl Inf Technol 1(3):1–3
Vincent L, Soille P (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans Pattern Anal Mach Intell 13(6):583–598
Weliwitage C, Harvey AL, Jennings AB (2005) Handwritten document offline text line segmentation. In: Proceedings of the digital image computing: techniques and applications, pp 184–187
Wong K, Casey R, Wahl F (1982) Document analysis systems. IBM J Res Dev 26(6):647–656
Yadav D, Sánchez-Cuadrado S, Morato J (2013) OCR for Hindi language using a neural network approach. J Inf Process Syst 9(1):117–140
Yin F, Liu CL (2009) Handwritten Chinese text line segmentation by clustering with distance metric learning. Pattern Recognit 42(12):3146–3157
Zahour A, Taconet B, Mercy P, Ramdane S (2001) Arabic hand-written text-line extraction. In: Proceedings of the sixth international conference on document analysis and recognition, pp 281–285