An Experimental Technique for OCR Line and Word Segmentation using Probability Distribution Estimation

Rajan Goyal, Rajesh Kumar Narula, Manish Kumar

Abstract*:Segmentation is always an important step in designing an Optical Character Recognition (OCR) of any script. In this paper, we focus on the line and word segmentation in typewritten Gurmukhi script documents. In order to perform this task, we consider OCR based methodology where several processing steps are implemented. The typewritten documents suffer from several issues such as noise, skew, and quality of the document. In this work, we present a combined pre-processing scheme where document thresholding and skew detection and correction schemes are implemented where image thresholding is obtained using Niblack's method and skew correction is carried out using gradient histogram algorithm and uniform orientation is obtained. Later, line segmentation scheme is applied where probability density function is applied to generate the text distribution in the probability map. Here, identifying the relation of the text to the exact line is a challenging task hence, we present a 2D-Gaussian modelling which helps to identify the text boundaries in the x and y direction. The proposed methodology is applied for typewritten Gurmukhi documents and an experimental study is carried out to show that the proposed approach achieves better performance when compared with the existing techniques.

***Keywords*:OCR, Typewritten, Line segmentation, Character segmentation, Probability density function.**

I. INTRODUCTION

OCR is an active research area in the field of pattern recognition, machine learning and computer vision applications [1]. OCR alludes tothe interpretation of images of typewritten, handwritten and printed documents in the form of editable text. This technique of OCR is widely adopted in various real-time applications and implemented for several languages such as Arabic [2], [3] Chinese [4], [5] and various Indian languages such as Devanagari [6], [7]Bangla [8], Gujarati [9]and Gurmukhi [10] etc. These techniques are successfully implemented for machine printed and handwritten documents but scanned document analysis and character recognition still considered as a tedious task for researchers.

Pattern recognition in computer vision systems contains three main stages which are known as pattern observation, pattern segmentation and classification. According to the segmentation process, theobserved patterns are categorized in their corresponding pattern which contains the content of the document in the form of lines and characters. Pattern recognition performs the task of identification of these characters with the help of prior information about the document. Segmentation of these patterns plays an important role in optical character recognition systems. In any automated OCR system, the document is processed through computer vision-based application where text document images are converted into editable [11] which can be used in several applications. Generally, OCR systems perform three primary tasks which are the localization of text in the document,line&character segmentation, and recognition. According to OCR systems, text recognition is considered an important aspect. As discussed before that several OCR systems have been introduced for various languages which consider handwritten and printed document analysis but achieving the eminent performance for typewritten documents is considered as a challenging task. In various government organizations information is stored in the typewritten documents. Moreover, the complexity increases when the scanned typewritten documents are in the cursive script such as Urdu and Gurmukhi [12].

Gurmukhi is the 14th most widely spoken language in the world. This script is derived from a Punjabi term *"Guramukhi"* which means that *"from the mouth of the Guru"*. Several government organizations use Gurmukhi language for official tasks where handwritten and typewritten documents are used for information collection and storage. In order to develop an OCR system for cursive typescript, text [13], line and character segmentation playsanimportant role. In this field, line and character segmentation are considered as a challenging task for cursive typewritten languages. Several techniques have been introduced for line and character segmentation in different automated OCR systems. Desai [14] recently discussed the OCR system for Gujarati handwritten numbers using a neural network. Moreover, pre-processing schemes are also applied such as thinning and skew-correction. In the pre-processing stage, image thresholding is important which helps to identify the structure of the document text lines and characters. Based on these assumptions, Lázaro et al. [15] introduced a novel thresholding approach for image binarization

Revised Manuscript Received on July 22, 2019.

Rajan Goyal, Research Scholar,I.K.Gujral Punjab Technical University Kapurthala, Punjab, India. Email-id: er.rajangoyal@gmail.com

Dr. Rajesh Kumar Narula,Assistant Professor, Department of Mathematics Science, I.K.Gujral Punjab Technical University Kapurthala.Punjab, India Email-id: dr.rknarula@gmail.com

Dr. Manish Kumar Jindal, Professor, Punjab University regional centre, Muktsar, Punjab, India. Email-id: manishphd@rediffmail.com

application by finding the histogram of the input document image where histogram derivative is computed, and the new smoothed histogram is formulated and later neural network model is implemented to obtain the threshold. Similarly, Paiet al. [16] also introduced an adaptive approach for image thresholding or binarization scheme where the block-based approach is implemented with the help of local and global methods. In any OCR system, text recognition is the most desired objective along with the text line and character segmentation. Several techniques have been introduced for text line and character segmentation for handwritten, scanned and typewritten documents. In this field of text line segmentation from handwritten documents, gaps between lines and skewed or curled text-lines are considered some challenging issues. Along with this, overlapping of text and touching text lines proliferate the segmentation complexities. Based on these complexity circumstances, Alaei et al. [17] introduced a novel technique for the segmentation of text lines. This process is carried out by applying various steps such as image painting which helps to divide the background and foreground regions of the image. Later, image dilation and thinning approaches are applied, line separators are formulated and line segmentation is performed. Moreover, text overlapping issue is addressed by constructing a contour trace and then upward and downward tracing is applied to obtain the top and bottom limits. These techniques of text line segmentation are also applied to the Indian languages. Aradhya et al. [18] discussed the challenging issue in handwritten documents which are caused due to skew and curves in the text lines. Authors focused on these issues and presented a novel approach for cursive handwritten Kannada script. This complete process is divided into two main stages where first of all, character component gaps are identified and bridged with the help of morphological techniques. In the next phase, the component extension scheme is implemented for extracting the text lines. Similarly, Koo et al. [19] suggested that text-line segmentation techniques suffer from several issues such as text orientation variation, character scale variation and the interference between the consecutive text lines. In this process, text lines are evaluated and the curvilinearity is identified to identify the line spacing and the construction. These schemes of text-line segmentation can be applied to the historical machine printed documents for extraction of historic information. However, these documents suffer from various issues as discussed before which are known as skew, low-quality image,and ink quality. Moreover, these documents are complex and dense in nature and during the scanning process, punctuation and noise identification becomes a complex task. In order to deal with these issues, Nikolaou et al. [20] presented Adaptive Run Length Smoothing Algorithm *(ARLSA)* scheme which attempts to reduce the segmentation error. Recently, Jindal et al. [21] presented an OCR framework for Gurmukhi handwritten document images.

A. Issues and challenges in OCR systems

OCR technology is widely adopted in the various real-time application and several techniques have been presented during the last decade which urges to improve the performance of the character recognition. However, several challenges are present in this field of character recognition. Some of the critical aspects are discussed in this section which degrades the performance of OCR. In order to achieve the optimal performance of character recognition, high quality and good resolution images are demanded. The image acquisition process is also considered an important factor in the performance enhancement of OCR. Generally, OCR systems are capable of achieving better performance for scanned documents or images when compared with the images captured by the cameras. In general, several issues are present in the various OCR systems which suffer from different types of issues such as scene complexity [22], uneven lighting conditions [23], document text skewness [24] and multilingual environment [25] resulting in poor performance of OCR applications.

B. Contribution of the work

In this work, we focused on the Gurmukhi language-based OCR system and presented a novel approach for text-line and word and character segmentation in typewritten Gurmukhi documents. According to this process, first of all, we apply a skew correction technique for the given typewritten documents. In the next phase, we apply a novel scheme of document binarization which helps to identify the connected components and remove the unwanted pixel positions. Later, we focused on the text-line localization and segmentation where we used density function computation for analyzing the text density and then probability density function is applied which provides the top, bottom, left, and right coordinates for the box for line and finally, word and character segmentation schemes are implemented.

The rest of the manuscript is organized as follows: Section 2 deals with the recent studies in this field of OCR where we have discussed the various techniques of document binarization, pre-processing, text-line segmentation and classification. Section 3 presents the proposed approach for the segmentation of text-lines & characters. Section 4offersa complete experimental study using typewritten Gurmukhi script language and finally, Section 5 gives concluding remarks of the work.

II. LITERATURE SURVEY

Character recognition from ancient documents plays an important role to achieve various sort of information about the civilizations such as cultural diversities and historical knowledge. Chamchong et al. [26] emphasized on Thai character segmentation using Palm leaf documents. Usually, information extraction from ancient documents is a tedious

task due to poor quality, fragility, and document deterioration over the age. Similarly, Thai palm leaf documents are also complex in nature and undergo with various factors such as noise, poor contrast, blurriness, etc. which may affect the information extraction and character segmentation process. Authors in [26] reported a novel approach for Thai character segmentation from ancient documents. First of all, the preprocessing stage is applied to reduce the noise and optimal binarization scheme is also implemented to improve the document quality. In this work, the partial projection profile method is applied for text line segmentation and later contour tracing approach is implemented for character segmentation.

The line and character segmentation are considered the main task of any OCR system.Chen et al. [27] explored the text line segmentation approach using colour and texture of the historical data. Authors adopted pyramidal document analysis approach [28] where the initial document is scaled with a scale factor of $\alpha < 1$ and the document is categorized into four main categories such as out of the page, background, text block, and decoration. This process is based on the machine learning where feature extraction process is accomplished and a real-valued feature vector is formulated using $n \times n$ pixel window. Later color and coordinate features [29] are applied with the help of pixel coordinates and r, g, b values. Moreover, LBP (Local Binary Pattern) and Gabor feature extraction also implemented to formulate a robust feature vector. However, the large feature vector may suffer from computational complexity, hence Fast Correlation-Based Filter (FCBF) algorithm [30] is applied to select the optimal feature set and finally SVM (Support Vector Machine) classifier is used. Text line segmentation can help to achieve information about character alignment in the document, connected components, and spatial information. Later this information is analyzed and the document is divided into text and non-text. According to [31], [32], line segmentation can be categorized in various methods such as Hough transform, projection-based method, grouping method, stochastic method and smearing methods. According to the smearing method, a hypothetical flow of water is analyzed and flow a particular direction represents the text in the image frame. However, this approach doesn't consider flow angle and without water, flow area is neglected. Further, this technique is improved using connected components and bounding boxes which helps to identify the accurate angle of water flow [33].

OCR any language, cursive or connected character segmentation and recognition are known as challenging issues. Several techniques developed for offline handwritten character segmentation such as Gaussian Mixture Model [34], Inertial and Big drop-fall [35] etc. but achieving desired performance remains an unresolved issue due to overlapping of the texts and touching characters. In order to deal with this issue, Lacerda et al. [36] introduced a novel approach for touching handwritten digit segmentation. According to this process, feature point selection is performed using image skeletonization and touching regions are clustered using Self-Organizing Maps. Similar to this, Jindal et al. [21] developed a new approach to line segmentation for Gurmukhi script.However, this work only focused on online segmentation.

Similar to the line segmentation, character segmentation is also an important research area in computer vision-based OCR systems. Handwritten and scanned documents are more complicated for information extraction, line, and character segmentation. Roy et al. [37] focused on the multi-oriented touching character segmentation. According to this study, touching characters generate a cavity in the background region which can be obtained by estimating convex hull. With the help of this information, initial segmentation points are identified in such a way that a single character can be covered in the initial segmentation points. Furthermore, these points are computed in the entire image and merged to formulate an optimal segmentation module. Moreover, this approach identifies the likelihood characters using SVM classifier and objective functions are also formulated which help to formulate the dynamic programming algorithm. Based on the touching character segmentation study, Xu et al. [38] proposed feature extraction-based character segmentation method for Chinese handwritten documents. As discussed in [25], [35] this method also uses contour tracing and skeletonization process which helps to identify the character separation. Later, a filter is designed with the help of the supervised learning scheme which helps to remove the unwanted cuts resulting in improved precision.

III. PROPOSED MODEL

In this work, we focus on the Line, word and character segmentation technique for any typewritten documents of Gurmukhi script. We have considered Gurmukhi script as our research area in OCR field because it is one of the vital scripts used in Indian languages. This section deals with the development of a novel and robust approach for line and word segmentation for Gurmukhi script. This complete section has been divided into three sub-section: (A) Gurmukhi script description, (B) Brief discussion about the flow of process, (C) Image pre-processing phase, (D) Skew correction technique,(E) Line segmentation where we introduce a new algorithm for line segmentation from a typewritten scanned document and (F) Word segmentation where segmented lines processed consecutively and words are segmented.

A. Properties of Gurmukhi script

Gurumukhi script commonly used in Punjabi language and it has been ranked 14th most widely spoken language in the world which is the combination of various symbols such as consonants and modifiers. Gurmukhi script has three vowel bearers, thirty-two consonants, six additional consonants, nine vowel modifiers, three auxiliary signs and three half characters. Gurmukhi script is written in the top to bottom and

left to right style. Sample characters of Gurmukhi script depicted in Fig. 1.

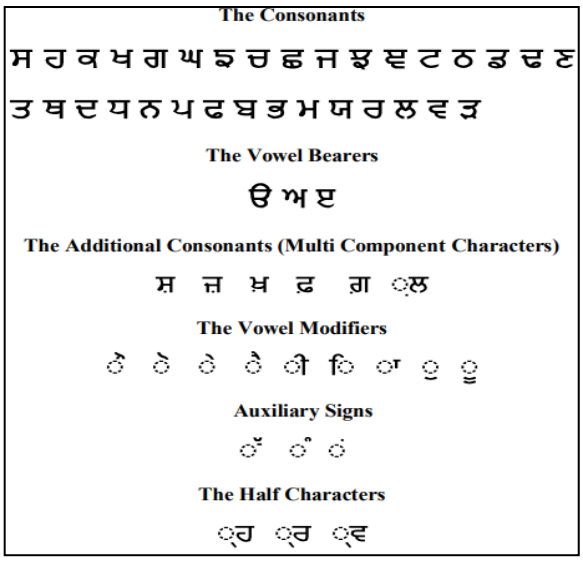


Fig. 1. Gurmukhi script character set.

Gurmukhi script words can be categorized into three main categories based on their zones such as the upper zone, a middle zone, and lower zone. A sample representation of Gurmukhi script is depicted in Fig. 2 which shows all zones in Gurmukhi script word. In Gurmukhi script, most of the characters have shirorekha which have a horizontal line at the upper part of the character and all characters are connected with the help of line which is known as a headline.[21]

Fig.2. Gurmukhi script word and zone representation.

B. Brief discussion of the proposed approach

In this section, we present a brief discussion about the proposed approach for offline typewritten Gurmukhi document line and word segmentation. Generally, two-type of techniques have been presented in the field of OCR for segmentationare (i) Bottom-up method where connected components are used for grouping into the lines and later these lines are grouped into the zones. Another method is a top-down method where the complete document is segmented in small zones and later these zones are segmented into lines. However, these techniques fail for various cases in typewritten documents due to irregular page layout, curvilinear text lines and variation in character sizes, etc. In order to overcome this issue, we have introduced a non-parametric method where the image is binarized and the pixels are represented as 0 and 1. During image acquisition, images may suffer from the skew and orientation issues hence we present a skew correction and detection scheme. Later line segmentation method is applied using probability density function-based text distribution generation. More distribution indicates the existence of text at those locations. Further, to identify the relation of the text to the line is performed using the 2D Gaussian function by computing the distribution in x and y direction. Finally, block computation method is presented for word segmentation. This complete process is depicted in Fig. 3.

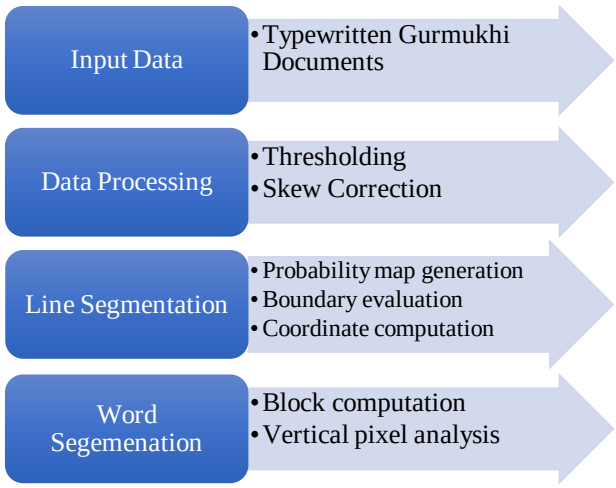


Fig. 3.Flowchart of the proposed approach.

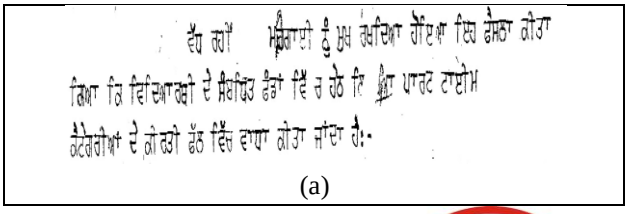
C. Image thresholding approach

In this section, we briefly discuss the image thresholding method using Niblack's[39] where a threshold value is calculated by sliding a rectangular window over the grey-level image. This threshold value (T) is computed with the help of mean m, standard deviation s of all considered pixels in the rectangular window, which can be expressed as:

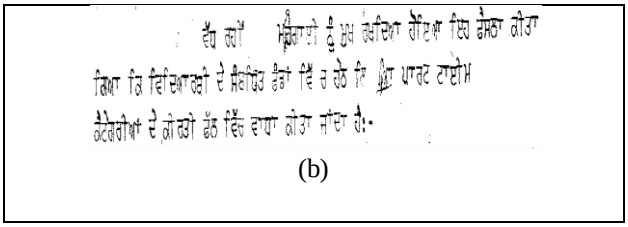
$$T = k * s + m \quad (1)$$

Where k is constant whose value varies between 0 to 1. In this phase, the image binarization quality depends on the value of k and the size of the sliding window. In this work, we have considered the value of k as 0.2 and the size of sliding window is 25x25. However, the selection of k and size of sliding window is a tedious task. Hence, we present a novel computational approach to find the sliding window and k parameters.

According to this approach, first of all, the input image is binarized using a global thresholding method resulting in texture area identification as shown in Fig. 4.



(a)



(b)

Fig. 4.Image Threshold; (a) Input image (b) Binarized image using Niblack's method

D. Skew detection and correction algorithm

In this section, we present a skew detection and correction scheme for Gurmukhi typewritten scanned document images. To achieve this, the gradient histogram algorithm is applied where it is assumed that the gradient orientation is perpendicular to the text line. Once the skew angles are determined, the input image can be rotated according to the identified angle.

Let us consider that input image is given as $I(x, y)$ and the gradient vector is $[m, n]^T$at current position (x, y) can be given as $m = \frac{\partial I(x,y)}{\partial x}$ and $n = \frac{\partial I(x,y)}{\partial y}$ based on this, the orientation of this gradient vector can be estimated which can be used for finding the angle of text-line with correspond to the gradient. This orientation can be computed as:

$$\psi = \arctan\left(\frac{n}{m}\right) \tag{2}$$

Where $\psi = [-\Pi, \Pi]$

At this stage, gradient vector given input can be identified by applying $N \times N$ Sobel filtering technique. A sample representation of Sobel operator for m and n gradient vectors is expressed in Fig. 5.

-1	0	1
-2	0	2
-1	0	1

(a)

1	2	1
0	0	0
-1	-2	-1

(b)

Fig. 5. Sobel Mask; (a) m vector; (b) n vector.

In any skewed document, a number of points can be identified whose gradient is perpendicular to the text line. Moreover, statistical information is also used for skew angle estimation. Orientation histograms $h(\psi)$ are calculated using Eq.(2). With the help of this, the orientation of document can be defined as:

$$\phi = \begin{cases} \psi - \frac{\pi}{2}, if\ \phi = [0, \pi] \\ \frac{\pi}{2} - \psi, \quad if\ \psi = \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \end{cases} \tag{3}$$

Where ψ denotes the maximum value for the given histogram $h(\psi)$. After finding the initial skew angle, the optimal value is obtained by refining these initial values with the help of a cubic polynomial function which is generally given as:

$$y = f(x) = ax^3 + bx^2 + cx + d \tag{4}$$

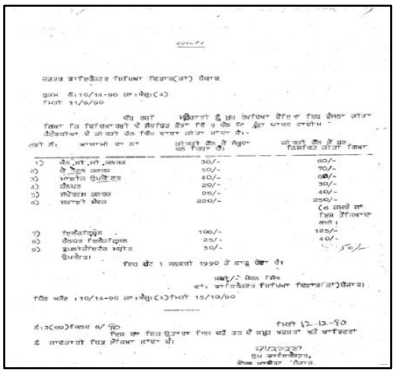
Where a, b, c and d are the constants which need to identify the specific polynomial value. In our experiment, we have considered the values of a, b, c and d as 0.2,0.45,0.3 and 0.6 respectively. With the help of rotation angle, the input image can be rotated to achieve the skew corrected image. During this process of rotation, bilinear interpolation is also applied which helps to reduce the noise. This rotation can be obtained as:

$$\begin{bmatrix} X_{new} \\ Y_{new} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} - \begin{bmatrix} C_x \\ C_y \end{bmatrix} \tag{5}$$

Where (C_x, C_y) denotes the center point of the rotation and γ_{ij} denotes the element of rotation matrix.

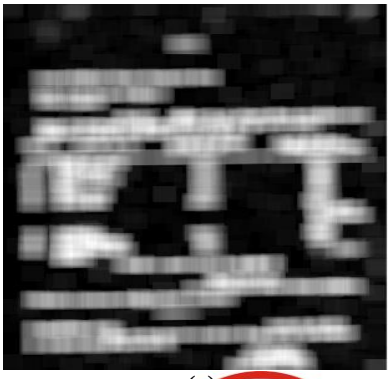
E. Line segmentation

Line segmentation plays an essential role in any OCR system which can be used further for word zone estimation, alignment, and word segmentation. To perform this task, we have used background and foreground pixel representation scheme which helps to identify the compactness of the lines in the given typewritten scanned document. Initially, we assume that background pixels are denoted by 0 and foreground pixels are denoted by 1 in a binary document image.



(a)

(b)



(c)

Fig. 6 Line segmentation process; (a) Input image, (b) Text density analysis, (c) Density function estimation

Generally, text image segmentation is performed on the binary image in the OCR systems. In any typewritten document, some specific properties exist such as character size and distances between lines, etc. However, this type of pattern can be generated by modeling a probability density function which generates a probability map. This probability map helps to identify that whether the current pixel belongs to the text line or not. Figure 6(a) shows a sample input image and a corresponding probability map is depicted in Figure 6(b). At this stage, black pixels can be considered as a random probability density function which represents the text line distribution based on the pixel information. Continuous and significant values of probability density function denote the probability of the presence of a text line whereas small values denote the gap or margin between consecutive lines or words.

Let us consider that $I(x,y)$ is given as a converted binary image where $x = 1, \dots, P$ and $y = 1, \dots, Q$. For this given input binary image, we compute a probability density function as $f_{est}(x,y)$ in a given two-dimension space with the help of Gaussian kernel, this function estimation can be expressed as:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \alpha(u,v) I(x-u, y-v) du dv \qquad (6)$$

Final, a density analysis of text lines isdepicted in Figure 6(c). Peaks and valleys are identified based on the computed probability density function map where peaks(P) represent the text lines and valleys represent the gap between adjacent text lines i.e.

$$PDF_{Map} = \begin{cases} P, \ text\ line \\ V, gap\ between\ text \end{cases}.$$

To obtain better segmentation performance, we find the text line boundary and it is also considered that adjacent text line regions should not overlap. In this work, we present a novel method for text line boundary identification where initial text line estimation is required which is computed

using probability density function. Here, we present a boundary evolution method which can be expressed as:

$$\frac{\partial l}{\partial t} + G_N \nabla f = b_k |\nabla f| \qquad (7)$$

Where l denotes an implicit function, t denotes the time and $\frac{\partial l}{\partial t}$ denotes the variation in the implicit functionbased on the time. This movement is analyzed based on the boundary curvature as b_k and S_N denotes the movement speed in the text line direction. Let us consider that probability density function $f(x,y)$ follows a reasonable speed as G_N. The speed will increase if the black pixels are large and slower in the gap regions. Let us consider that at t point of time, $\mathcal{N}$ number of bounding boxes are obtained in a closed region $S_n^t, n = 1, \dots, \mathcal{N}$. After $t+1$ iteration, the bounding boxes can be given as: $G_n^{t+1}, n = 1, \dots, \mathcal{M}$, $\mathcal{M} \leq \mathcal{N}$. For each consecutive text line, the bounding boxes are denoted as G_j^t and G_k^t. Inorder to overlap the regions, we present a criteria given as:

$$\begin{aligned} L_c(S_k^t) \leq mean(G_j^t) \leq r_c(G_k^t) \\ L_c(S_j^t) \leq mean(G_k^t) \leq r_c(G_j^t) \end{aligned} \qquad (8)$$

WhereL_c denotes the left coordinates, R_c denotes the right coordinates, $mean(G_k^t)$ denotes the x component of center of gravity of G_k^t.Similarly$r_c(G_k^t)$ and $L_c(G_k^t)$are also computed based on the center of gravity. Based on these assumptions boundary of regions R can be expressed as:

$$\begin{aligned} top(R) &= min\left(bottom(G_k^t), top(G_j^t)\right) \\ bottom(R) &= max\left(bottom(G_k^t), top(G_j^t)\right) \\ L_c(R) &= max\left(L_c(G_k^t), L_c(G_j^t)\right) \\ r_c(R) &= min\left(right\,(G_k^t), r_c(G_j^t)\right) \end{aligned} \qquad (9)$$

This process of boundary line detection is applied to avoid the boundary overlap resulting in appropriate segmentation.

Input: Input image data$I(x,y)$, Gaussian kernel function Output: Probability density function $f_{est}(x,y)$, text density and identified boundary coordinates i.e. $top(R)$ and $bottom(R)$, $L_c(R)$ and $r_c(R)$.
Step 1: Compute the probability density function as $(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \alpha(u,v) I(x-u, y-v) du dv$. **Step2:** Represent the probability density function and identify the peaks and valleys. **Step 3:**Represent the peak as text line and the valleys as the gap between neighboring text lines $PDF_{Map} = \begin{cases} P, \ text\ line \\ V, gap\ between\ text \end{cases}$ **Step 4:** Initiate the boundary evolution function using Eq. (7) and compute the boundary curvature. **Step 5:** for $i = 1: total\_pixel$ do **Step 6:** Identify the pixel representation and its speed i.e. speed will increase if the black pixels are large and slower in the gap regions. **Step 7:** Compute the boundary box from the probability density function as left coordinates $L_c(S_k^t) \leq mean(G_j^t) \leq r_c(G_k^t)$ and right coordinates $L_c(S_j^t) \leq mean(G_k^t) \leq r_c(G_j^t)$ **Step 8:** compute the minimum and maximum coordinate values from the left and right coordinates and construct a boundary region using Eq. (9).

F. Word Segmentation

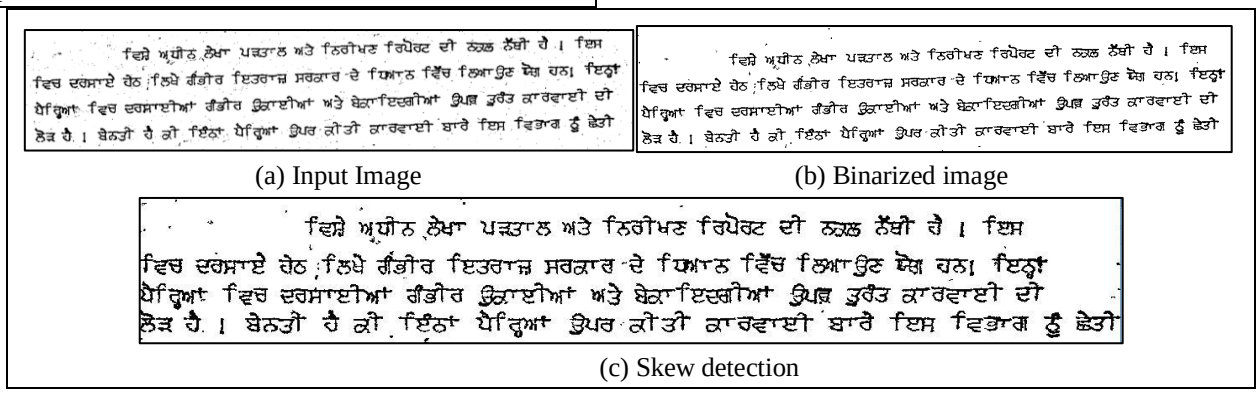
In the phase, the segmented lines are used for the further process of word segmentation. Now, each segmented line is scanned vertically and if two or more consecutive empty pixels are evaluated and this process is repeated until the next empty pixels are found. The text between these empty pixels is considered as a single word. However, Gurmukhi is a cursive script language[40] which causes complexity to identify the empty pixels and scanning noise also may degrade the performance but applied image binarization scheme helps to reduce the unwanted noise and achieves a clear image.

Input: pre-processed and cropped typewritten text line as the input image ($\boldsymbol{I}$)
Output: Current letter($\boldsymbol{C_w}$) and remaining letters($\boldsymbol{R_w}$)

Step 1: Give input image as $\boldsymbol{I}$
Step 2: line crop using PDF estimation
Step 3: do while $\boldsymbol{untiltheendofline}$
Step 4: find the positions of pixel i.e. total rows and column in the text line as $[\boldsymbol{f}, \boldsymbol{c}] = \boldsymbol{find}(\boldsymbol{I})$
Step 5: Crop the identified positions as $\boldsymbol{Crop} = \boldsymbol{I}(\boldsymbol{min}(\boldsymbol{f}):\boldsymbol{max}(\boldsymbol{f}), \boldsymbol{min}(\boldsymbol{c}):\boldsymbol{max}(\boldsymbol{c}))$
Step 6: find the total number of columns as $[\boldsymbol{c}] = \boldsymbol{size}(\boldsymbol{Crop}, \boldsymbol{2})$
Step 7: for $\boldsymbol{i} = \boldsymbol{1}:\boldsymbol{C}$
Step 8: find the sum of total pixel values $\boldsymbol{S} = \sum_{i=1}^{r}\sum_{c=1}^{c}\boldsymbol{I}(\boldsymbol{i})$
Step 9: if $\boldsymbol{S} = \boldsymbol{0}$ then
Step 10: current word cropped as $\boldsymbol{C_w} = \boldsymbol{I}(:, \boldsymbol{1}:\boldsymbol{i}-\boldsymbol{1})$ and remaining words are as $\boldsymbol{R_w} = \boldsymbol{I}(:, \boldsymbol{s}:\boldsymbol{end}-\boldsymbol{1})$
Step 11: crop the non-zero pixel values from $\boldsymbol{C_w}$ and $\boldsymbol{R_w}$
Step 12: end
Step 13: end
Step 14: end

IV. EXPERIMENTAL STUDY

This section deals with the complete experimental study for Gurmukhi typewritten documents. The complete experimental study is carried out using MATLAB simulation tool running on Windows platform with Intel i5 processor. In this work, we have collected Typewritten Gurmukhi script data from the government official organizations. The proposed approach is applied using these datasets. In order to show the robust performance, we have considered 100 documents and evaluated the performance in terms of correct segmentation of text-lines and word.A complete process of the proposed approach is depicted in the below-given Fig.7 Document processing is shown where (a) Input image, (b) Binarized image, (c) Skew correction, and in Figure 8 show (a) Line and (b) Word segmentation is presented.

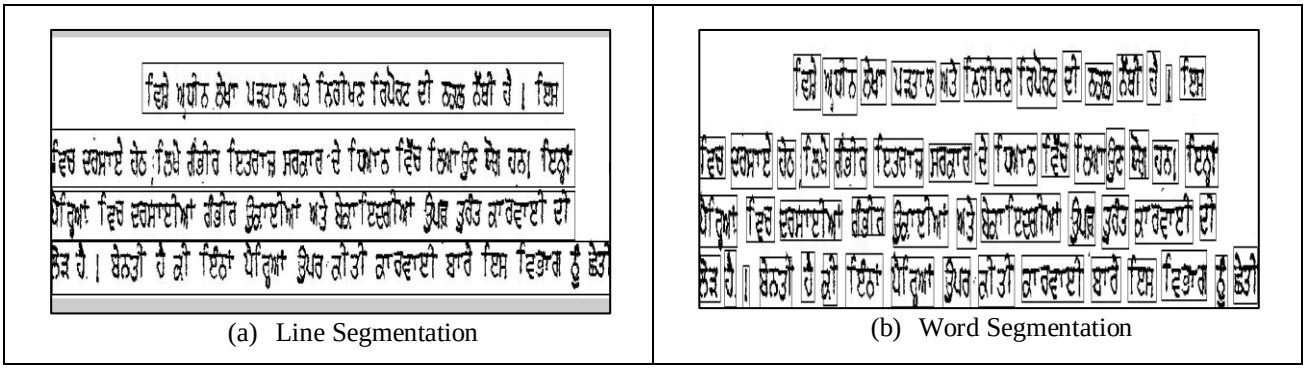


(a) Input Image (b) Binarized image

(c) Skew detection

FIGURE 7 DOCUMENT PROCESSING; (A) INPUT IMAGE, (B) BINARIZED IMAGE, (C) SKEW DETECTION.

In this experimental analysis, we have considered 100 typewritten scanned documents which are acquired from the Punjab Government Organization. According to the proposed approach, skew correction, binarization, connected component block identification, text density estimation, and probability density functions are computed which helps to perform the line and word segmentation. To compute the performance, we identify the total number of lines and words in the given document and compared the outcome of the proposed segmentation with these measurements. However, these measurements of groundtruth parameters are performed manually. Line and word segmentation performance are depicted in the below-given Fig.8.



(a) Line Segmentation (b) Word Segmentation

Fig. 8. Segmentation process; (a) Line segmentation, (b)Word segmentation.

Similarly, we performed these experimental analyses and measured the performance of line and word segmentation using the proposed approach for typewritten Gurmukhi scanned documents. The performance measurement is carried out by computing the following parameters: (a) Total number of lines segmented (b) Total number of words segmented. Here we present an experimental analysis using the proposed approach and measured the performance by using sample images.

Table-I Line segmentation performance comparison

Sample image	Number of Lines	Line Segmentation		
		Using PPA (Alaei et al., 2011)	Proposed Method	Accuracy using proposed
1	15	11	13	86.66
2	29	21	25	86.20
3	24	16	19	79.16

This study is carried out by considering a different type of scanned document samples. First of all, we evaluate the performance of line segmentation as given in Table I. In this experiment, the first document is considered which contains a smaller number of lines with proper line spacing. This image is processed through the various stages of proposed approach and the outcome is obtained that the proposed approach is able to segment 86.66% lines accurately whereas conventional scheme achieves line segmentation accuracy of 73.33%. The second image contains morenumber of lines but the noisy pixels are removed with the help of pre-processing and binarization method. Later, the proposed approach is applied which shows that the proposed approach achieves line segmentation accuracy as 86.20% and conventional method achieves 82.74% accuracy. For the next sample, a total number of lines identified as 24 where the proposedapproach segments the total 19 lines with the accuracy of 79.16% whereas PPA method segments 16 lines with the accuracy of 66.66%.Similarly, we compared the performance of word segmentation and compared with the PPA algorithm.

Table-II Word segmentation performance comparison.

Sample image	Number of Lines	Word Segmentation		
		Using PPA (Alaei et al., 2011)	Proposed Method	Accuracy using proposed
1	106	74	91	85.84
2	165	138	151	91.51
3	206	168	179	86.89

Next experiment we performed for word segmentation process for the same samples which are considered for the previous experiment and presented a comparative study in terms of correct word segmentation as given in Table II. In this study, sample image 1 is processed where a total number of words are identified as 106 and proposed approach segments 91 words accurately with the accuracy of85.84 whereas conventional PPA algorithm segments 138 words with the accuracy of 69.81%. For the second sample, 165 words are identified and the proposed approach is able to segment 151 words accurately whereas PPA algorithm segments 138 words, the word segmentation accuracy is obtained as 91.51% and 83.63% respectively. Finally, we evaluate performance for 3rd sample where total words are 206; a proposed approach successfully segments 179 words whereasconventional technique able to segment 168 words with slight over segmentation performance with the accuracy of 86.89% and 81.55% respectively.

Fig.9 shows experimental results for two case studies whose performance is analyzed with the help of parameters which are discussed before. Performance of the proposed approach is compared with the existing PPA [17] of line and word segmentation is performed.

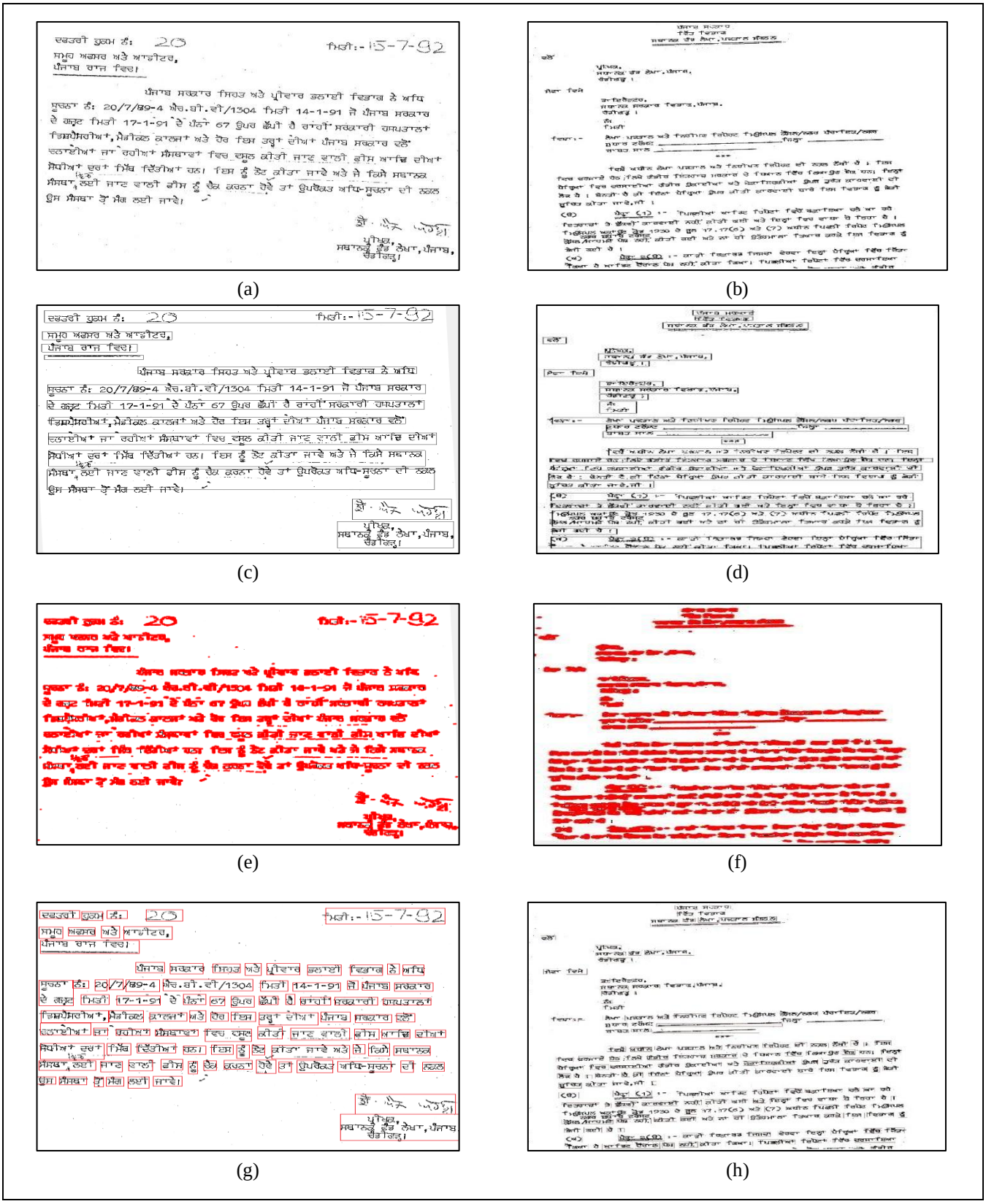


Fig. 9 Document processing using proposed approach for two test cases; (a, b) input images, (c, d) Line segmentation,(e, f) connected component, (g, h) word segmentation

V. CONCLUSION

In this work, we have presented a novel approach for line and word segmentation using Gurmukhi script for improving the OCR system. The proposed work is carried out using multiple stages such as image pre-processing where skew correction and image thresholding schemes are applied. Later, probability density function-based map generation strategy is applied to identify the text distribution in the document and finally 2D Gaussian model is applied for representing the complete text line. The proposed approach does not use any script-specific knowledge and hence can be implemented for other languages. Experimental study shows that the proposed approach achieves better performance when compared with the state-of-art techniques.

REFERENCES

1. Alotaibi, F., et al., Optical Character Recognition for Quranic Image Similarity Matching. IEEE Access, 2018. **6**: p. 554-562.
2. Radwan, M.A., M.I. Khalil, and H.M. Abbas, Neural networks pipeline for offline machine printed Arabic OCR. Neural Processing Letters, 2017: p. 1-19.
3. Nashwan, F., et al., A Holistic Technique for an Arabic OCR System. Journal of Imaging, 2017. **4**(1): p. 6.
4. Zhang, X.-Y., Y. Bengio, and C.-L. Liu, Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark. Pattern Recognition, 2017. **61**: p. 348-360.
5. Tang, Y., et al. Semi-Supervised Transfer Learning for Convolutional Neural Network Based Chinese Character Recognition. in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on. 2017. IEEE.
6. Chaudhuri, A., et al., Optical Character Recognition Systems for Hindi Language, in Optical Character Recognition Systems for Different Languages with Soft Computing. 2017, Springer. p. 193-216.
7. Johnson, K., et al., OCR for devanagari numerals using zonal histogram of angle. Journal of Statistics and Management Systems, 2017. **20**(4): p. 519-534.
8. Sarkhel, R., et al., A multi-objective approach towards cost effective isolated handwritten Bangla character and digit recognition. Pattern Recognition, 2016. **58**: p. 172-189.
9. Prasad, J.R. and U. Kulkarni, Gujrati character recognition using weighted k-NN and mean χ 2 distance measure. International Journal of Machine Learning and Cybernetics, 2015. **6**(1): p. 69-82.
10. Kumar, M., M. Jindal, and R. Sharma, Offline handwritten Gurmukhi character recognition: Analytical study of different transformations. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 2017. **87**(1): p. 137-143.
11. Sahare, P. and S.B. Dhok, Multilingual character segmentation and recognition schemes for Indian document images. IEEE Access, 2018. **6**: p. 10603-10617.
12. Naz, S., et al., The optical character recognition of Urdu-like cursive scripts. Pattern Recognition, 2014. **47**(3): p. 1229-1248.
13. Neumann, L. and J. Matas. Real-time scene text localization and recognition. in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. 2012. IEEE.
14. Desai, A.A., Gujarati handwritten numeral optical character reorganization through neural network. Pattern recognition, 2010. **43**(7): p. 2582-2589.
15. Lázaro, J., et al., Neuro semantic thresholding using OCR software for high precision OCR applications. Image and Vision Computing, 2010. **28**(4): p. 571-578.
16. Pai, Y.-T., Y.-F. Chang, and S.-J. Ruan, Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images. Pattern Recognition, 2010. **43**(9): p. 3177-3187.
17. Alaei, A., U. Pal, and P. Nagabhushan, A new scheme for unconstrained handwritten text-line segmentation. Pattern Recognition, 2011. **44**(4): p. 917-928.
18. Aradhya, V. and C. Naveena. Text line segmentation of unconstrained handwritten Kannada script. in Proceedings of the 2011 International Conference on Communication, Computing & Security. 2011. ACM.
19. Koo, H.I. and N.I. Cho, Text-line extraction in handwritten Chinese documents based on an energy minimization framework. IEEE Transactions on Image Processing, 2012. **21**(3): p. 1169-1175.
20. Nikolaou, N., et al., Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths. Image and Vision Computing, 2010. **28**(4): p. 590-604.
21. Jindal, S. and G.S. Lehal. Line segmentation of handwritten gurmukhi manuscripts. in Proceeding of the workshop on Document Analysis and Recognition. 2012. ACM.
22. Ye, Q. and D. Doermann, Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence, 2015. **37**(7): p. 1480-1500.
23. Pandya, M.D. and R.P. Jay. A Survey: Artificial Neural Network for Character Recognition. in Proceedings of Fifth International Conference on Soft Computing for Problem Solving. 2016. Springer.
24. Amir, M. and A. Jindal. New Fast Content Based Skew Detection Algorithm for Document Images. in International Conference on Pattern Recognition and Information Processing. 2016. Springer.
25. Soora, N.R. and P.S. Deshpande, Novel geometrical shape feature extraction techniques for multilingual character recognition. IETE Technical Review, 2017. **34**(6): p. 612-621.
26. Chamchong, R. and C.C. Fung. Character segmentation from ancient palm leaf manuscripts in Thailand. in Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. 2011. ACM.
27. Chen, K., et al. Robust text line segmentation for historical manuscript images using color and texture. in Pattern Recognition (ICPR), 2014 22nd International Conference on. 2014. IEEE.
28. Baechler, M. and R. Ingold. Multi resolution layout analysis of medieval manuscripts using dynamic mlp. in Document Analysis and Recognition (ICDAR), 2011 International Conference on. 2011. IEEE.
29. Wei, H., et al. Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents. in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. 2013. IEEE.
30. Yu, L. and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. in Proceedings of the 20th international conference on machine learning (ICML-03). 2003.
31. Likforman-Sulem, L., A. Zahour, and B. Taconet, Text line segmentation of historical documents: a survey. International Journal of Document Analysis and Recognition (IJDAR), 2007. **9**(2-4): p. 123-138.
32. Amin, A. and S. Wu. Robust Skew Detection in mixed Text/Graphics Documents. in null. 2005. IEEE.
33. Brodić, D., Extended approach to water flow algorithm for text line segment[ation. Journal of Computer Science and Technology, 2012. **27**(1): p. 187-194.
34. Duda, R.O., P.E. Hart, and D.G. Stork, Pattern classification. 2nd. Edition. New York, 2001. **55**.
35. Wang, X., K. Zheng, and J. Guo, Inertial and big drop fall algorithm. International Journal of Information Technology, 2006. **12**(4): p. 39-48.
36. Lacerda, E.B. and C.A. Mello, Segmentation of connected handwritten digits using Self-Organizing Maps. Expert systems with applications, 2013. **40**(15): p. 5867-5877.
37. Roy, P.P., et al., Multi-oriented touching text character segmentation in graphical documents using dynamic programming. Pattern Recognition, 2012. **45**(5): p. 1972-1983.
38. Xu, L., et al., An over-segmentation method for single-touching Chinese handwriting with learning-based filtering. International Journal on Document Analysis and Recognition (IJDAR), 2014. **17**(1): p. 91-104.
39. Niblack, W., An introduction to digital image processing. Vol. 34. 1986: Prentice-Hall Englewood Cliffs.
40. Lehal, G.S. and C. Singh. A Gurmukhi script recognition system. in Pattern Recognition, 2000. Proceedings. 15th International Conference on. 2000. IEEE.

AUTHORS PROFILE

Rajan Goyal received his B.E in Computer Engineering from Bharati Vidyapeeth University, Pune in 2008 and M.Tech in Computer Science and Engineering from Department of Computer Science, Punjabi University, Patiala, India in 2011. He is working as an Assistant Professor in Yadawindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo. He is currently pursuing his Ph.DdegreeinComputerScienceandEngineeringfromI.K. GujralPunjab Technical University Kapurthala, Punjab, India. His research interests include Character Recognition

Dr. Rajesh Kumar Narula received his Bachelor's degree in science in 1993 and Post Graduate degree in M.Sc Mathematics from Guru Nanak Dev University, Amritsar. He received his PhD degree in Mathematics from University of Bikaner, Bikaner (Raj.) in December 2007 presently he is working as a Assistant Professor in Mathematics in I.K.Gujral Punjab Technical University, Jalandhar. His research interests is Information & Coding Theory.

Prof.ManishKumarJindalreceivedhisBachelor'sdegree in science in 1996 and Post Graduate degree in Computer Applications from Punjabi University, Patiala, India, in 1999. He holds a Gold Medal inhisPostgraduation.HereceivedhisPh.D.degreeinComputerScience & Engineering from Thapar University, Patiala, India in 2008. He is working as Professor in Panjab University Regional Centre, Muktsar, Punjab, India. His research interests include Character Recognition.