Journal of Theoretical and Applied Information Technology
15th January 2021. Vol.99. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED TEXT LINE SEGMENTATION AND SKEW ESTIMATION FOR HANDWRITTEN KANNADA DOCUMENT

[1] SHAKUNTHALA B S, [2] Dr. C S PILLAI

[1]Research Scholar, VTU-RC, ACS College of Engineering, Department of CSE, Bangalore, India

[2]Professor. ACS College of Engineering, Department of CSE, Bangalore, India

E-mail: [1]shakukit@gmail.com, [2]pillai.cs5@gmail.com

ABSTRACT

Abstract. When Handwritten Kannada document undergoes text line segmentation, the process is referred to as Text line segmentation and skew correction. This is quite essential for the HCRS (Human Character Recognition System). The process of text line segmentation and skew estimation tends to be quiet challenging during document analysis. The proposed system presents improvised text-line segmentation along with skew estimation for which the handwritten Kannada document forms the dataset. Following are the three methods for carrying out preprocessing, namely: (i) filtering (ii) gray scale conversion and (ii) Binarization. The ESLD (Enhanced Supervised Learning Distance) algorithm is being adopted for the assessment of distance amidst text lines and G_Clustering aids in grouping of words or the Connected Components. Also, by computing skew angle with respect to the gap, Skew estimation can be performed. It's elucidated from the output that the proposed system exhibits higher performance.

Keywords: *Segmentation, Skew Correction, Filtering, Gray Scale, Binarization.*

1. INTRODUCTION

The process of Handwritten Character Recognition helps in the identification of characters written by the writers. The most important process in HCR is the segmentation wherein text is transformed into lines. There are two types of handwritten text: first being the offline HCR which means that the writers utilize pen/pencil for writing on papers. The second is the online HCR which means that the writers utilizes digital tools like the electronic pen and all for writing purpose.

Since different individuals have varying writing styles, handwriting recognition becomes quite a tedious task. Many sorts of machine learning techniques are being employed and implemented for procedures pertaining to offline and online HCR which includes the SVM (Support Vector Machines), Gaussian Mixture Models, ANN (Artificial Neural Network), Fuzzy Logic etc.

Following are the different types of features an HCR system must include:

1. *Flexibility*: this particular feature must take into account different sorts of writing patterns from different people.

2. *Customization*: The handwritten styles of any writer must be easily comprehended by the OCR.

3. *Efficiency*: Online HCR systems should have good efficiency pertaining to time and space.

4. *Automatic Learning*: The OCR system must be trained via automatic learning mechanism so that the customization feature can be enabled.

The text-based recognition system comprises of a significant process of Segmentation which works on separating text lines, words and finally the characters thus ensuring effective classification and recognition. The output obtained from the segmentation phase highly ensures the character recognition's accuracy. Because of incorrect segmentation there can be an issue of false recognition. There exist numerous approaches which caters to printed text segmentation and handwritten text segmentation too. Printed text

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

segmentation involves consistent font size and line spacing, which makes it an easier methodology. In contrast, there is complexity in handwritten text segmentation because of discrepancies in writing patterns, overlapping of characters and also due to uniform/non-uniform skew in the entire text. These reasons prohibit implementing or expanding segmentation methods concerning machine printed documents to handwritten documents.

Text-line segmentation holds crucial for automatically recognizing off-line handwritten text documents. But this tends to be a challenging process due to issues such as differences in inter-line distance, existence of non-uniform baseline skew as well as touching and overlapping text-lines. The degree of righteousness of text-line segmentation is directly proportional to preciseness of the word or character segmentation thereby affecting the accuracy of word/character recognition. [1][2] Have suggested a variety of approaches pertaining to text-line segmentation which are classified as following: (i) projection profile techniques, (ii) Hough transform techniques, and (iii) smearing techniques.

The projection-profile technique involves dividing the document image into different column chunks having as 5% of overall document's width. For every chunk there is generation of horizontal projection profiles of foreground pixels. Based on the smoothed projection profiles, valleys in every chunk having the least foreground pixels amidst two consecutive peaks are identified based on the smoothed projection profiles thereby indicating the separation position of the two text lines. Extraction of the initial evaluated text lines is done by connecting every profile's valley with the nearest valley in the previous profile. Separating lines are drawn horizontally from left to right. Similarly, separating lines are drawn horizontally at the same position for unused valleys as drawn in the previous profile.

The process of Hough Transform based method includes enhancement and binarization of document images and thereafter extracting of CC (connected components). Classification of connected components is done with respect to their average height and width into the following types: large components, small components including accents, and third being the normal sized components which being the core part of the text lines. There is equal size partitioning of very connected component in the third subset. Hough transform is imposed to the gravity center points of all blocks. Next, in case half of the points are allocated to a text line then a CC is allocated to it as per the accumulator array. Then is the post processing, wherein the components in the second subset gets allocated to the nearest text lines and those in the first subset gets allocated to the text line they reside upon or are split and allocated to different parts.

Printed and handwritten documents implement the Smearing based techniques. The issue of text line segmentation pertaining to printed documents can be attended using the technique of Grouping or bottom-up approaches which groups the CCs in accordance with the geometrical and topological features. Following is the description of the technique employing the neighborhood CC analysis.

Unlike the printed document, process of handwritten text line segmentation and skew correction tends to be quite complicated. HCRs efficiency can be enhanced through the process of segmentation. The issue of concern is segmentation of handwritten Kannada document. Text line segmentation in a handwritten document is complex because of inconsistent, issue of skewness and also touching and overlapping of text lines. These issues can be resolved via proposed model which offers solution for both skew correction and text line segmentation.

There is a proposal of text line segmentation and skew correction techniques taking the Kannada handwritten document into consideration. The handwritten document image acts as input which undergoes preprocessing in following three steps: (i) Filtering (ii) Gray conversion and (iii) Binarization. The process of text line segmentation splits the document image into a set of segments following which there is identification of the edge and the CC (connected components) within the text line. Next is the task of gap estimation which involves measurement of gap amidst any two text lines or words by the means of ESLD (Enhanced Supervised Learning Distance) algorithm. Then is the computation of Mean height and average width of the CCs and also their grouping is performed via R_Clustering algorithm. The method groups the related connected components to build clusters so that the words are segmented effectively. Then there is extraction of lines/words and measurement of skew through optimum skew angle amidst the lines or word with respect to the gap. At last, the resultant deskewed text is written in a different document image.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The research paper is classified in 5 sections: Section 2 presents overview of certain related approaches for handwritten text line segmentation. Section 3 elaborates the proposed techniques of text line segmentation method and skew estimation. Section 4 illustrates Experimental output and the comparison in contrast to prevailing line segmentation approaches. Section 5 presents the conclusion.

2. RELATED WORKS

There is no doubt that the techniques of skew detection and correction holds reliable and successful enough in carrying out further processing in HCR (handwritten recognition system) and the research in the similar domain has been intensified for developing more skillful skew detection and correction approaches. Following are certain such techniques. There is recommendation of a technique to perform skew estimation from the scanned Persian document. It involves transforming the document image into a text block image by the means of connected component (CC) analysis and morphological closing accompanied by thinning operations. In addition, rectangular patch covering thin line is associated with the thinning operations. The slopes of the thinned lines aids in estimating the skew angle via Hough transform which is illustrated through the algorithm.

The above is experimented on the document image comprising of Chinese text. Yet another approach for estimating the skew angle pertaining to the handwritten English document image is proposed that relies upon the orthogonal projection. The author assures that the accuracy is retained even if the method is employed on varied languages. [3] Has proposed a basic technique for skew detection wherein all the components associated with the same line are clustered together. The skew angle acts as the peak angle between the centroid of the connected components. There is less computational overhead using this approach. Another method for skew and slant correction is recommended in [4] that relies upon geometrical model and projection. The experiment is performed with the help of handwritten English document. There is illustration of a technique for skew angle estimation of both printed and handwritten document. A technique has been proposed by [5] for segmentation of the handwritten text lines with respect to the historical document images. The line regions within the image are identified by estimating black or white transition maps. There is a discussion of a methodology which helps in skew identification and correction of the printed document via bilinear interpolation method. The computation rime can be reduced by transforming the discrete cosine. In addition, the skew angle can be identified by implementing fft to the four quadrants of the image. An effective and basic approach is highlighted that relies on boundary growing, thinning and moment for the estimation of skew angle. The proposed method yields at par accuracy and computational time. A technique based on random transform is put forth for skew estimation that has been experimented on printed Kannada documents. There is enhanced accuracy and execution speed with the proposed method.

Numerous research works are into practice that caters to handwritten documents and various basic datasets are being built that aid the researchers for the result sharing and comparing the classifier's performance, as put forth by [6]. The literature inculcates various standardized datasets for the following script like Roman, Chinese, and Korean. In addition, there is availability of good enough volume of running English text in the NIST SD3 and TD1 datasets. The other three NIST datasets (SD11-SD13) includes examples of phrases. A total of 91500 handwritten phrases underwent scanning at 200 dpi in binary mode. The CENPARMI dataset comprises of 17000 isolated digits fetched from images of nearly 3400 postal ZIP Codes through manual segmentation as indicated by [7]. There were about 28000 isolated handwritten characters and digits included in the CEDAR English dataset that were being derived from the images of US postal addresses as per [8]. Using a semi-automatic process, individual characters and digits were segmented from the address images. There were near about 5000 ZIP Codes, 5000 city names, and 9000 state name images.

Though India is considered a multilingual, multi-script country, availability of standard datasets pertaining to the Indian languages' domain is weak in contrast to other languages. Certain datasets with respect to off-line handwritten Bangla and Devanagari numerals and characters have been provided by [9]. Herein, four types of text categories are taken into account namely, stories and general news, sports news sentences, movie and Kannada medical texts. The data collection is performed by distributing the texts to 51 individuals with varying educational background and age. They are given plain A4 sheets for writing down the given text using a variety of pens of their own choice but with a calm and composed mindset. The handwritten texts obtained from the participants

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

undergo scanning in gray-scales via flatbed scanner having the resolution as 300 dpi. [10] Employs the Otsu's technique for obtaining the binary images of the scanned documents.

The technique of Offline handwriting recognition identifies the different set of letters or words prevailing in a digital image of handwritten text. There are two types of recognition system: "online" and "offline". It's observed by [11] that the Kannada handwriting becomes complex to pre-segment because of its vernacular which leads to difficulty in performing automatic off-line recognition. For resolving this issue, various recognition systems are built on the structure of HMM (hidden Markov models) which support joint segmentation and recognition. Images undergo pre-processing prior to training and testing, following which extraction of feature sequences takes place.

The feature extraction process works by extracting a group of related features in an ordered manner apt to Markovian modelling thus minimizing the redundancy factor in the word image

Pr-processing stage is then followed by feature extraction according to [12] while preserving the discriminative information for recognition. Following techniques are adopted for the process of feature extraction namely, SIFT (Scale Invariant component Transform), SURF (Speeded Up Robust Features) and ORB (Feature Extraction and descriptors). The SURF technique is a nearby element locator and a descriptor employed for assignments such as identification of an object/3D remaking/enrollment or characterization. The SIFT (scale-invariant element change) descriptor is the motivation behind the SURF technique. Unlike the SIFT technique, the SURF is much swifter and is proclaimed to be robust in contrast to different sort of image changes in SIFT. For recognizing interest focuses, SURF considers a number estimate of that determinant of a Hessianblob finder, having 3 whole number operations along with a pre-computed standard image. Its element descriptor relies upon the completeness of the Haar wavelet reaction (one of the wavelet method) centring on the purpose of interest.

Thereafter, analysis of Line segments of neighbouring partitions is performed in order to merge into correct text lines. Though, touching lines resulting in over or under segmentation fails to get segmented. Yet there is another line segmentation method that is oriented on piece-wise horizontal projections of vertical stripes as indicated by [13]. Here, there is computation of piece-wise separating lines which are employed in line segmentation and the vertical projection being utilized in word segmentation. The CTM (Cut Text Minimization) technique separates text lines in such a manner that the number of text points cut is reduced. The ascenders and descenders are being tracked to prevent cutting through lines with varying slope. Usage of projection profile technique is done in segmentation of skewed and overlapped Devanagiri script. Height of the line segments is obtained via piece-wise projection profile which id examined for identifying touching and overlapping lines. A water flow technique has been proposed for multi-skewed document images by [14].

Though effective working has been achieved with this method in collaboration with conjunct character recognition, it demeans certain segment of characters because of segmentation. Expectation-maximization algorithm has been proposed by [15] to understand mixtures of Gaussians [15]. By the means of Connected Components Labeling algorithm, the unwetted bands can be segregated according to [16], for acquiring the text lines. Moreover, the candidate points can also be determined against black and white blocks throughout the text line. These candidate points, helps in identifying the baseline which can be then stretched straight horizontally. There is a proposal of a novel methodology based on PPA (piece-wise painting algorithm) which deals with the segmentation of unconstrained handwritten text line as put forth by [17]. This methodology enables text line detection by improvising the separability between the foreground and background portions.

[18] Has put forth the approach of text line extraction from multi-skewed handwritten documents which relies upon the hypothetical water flows, from the image's left and height sides confronting obstruction from the text line characters. The text lines can be extracted by labelling theunwetted stripes on the image frame. [19] Has proposed segmentation of handwritten Chinese text line through the approach of clustering with distance metric.

The Mumford shah model employs morphing for eliminating overlaps amidst textlines and connected broken in the handwritten document. Combination of different on-line and off-line MCS (Multiple Classifier System) based systems has been proposed by [20] for handwritten text line recognition. Following is the description related to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

segmentation of Bangla unconstrained handwritten text. Using the horizontal histogram of stripes and association among the minimal values of histogram, extraction of text line is performed. The line segmentation algorithm employs Viterbi algorithm for locating the optimal succession of text and gap regions within the vertical zones. There exists certain praiseworthy work presented by [21][22]. [23] Have introduced text line detection of handwritten documents that relies upon block-based Hough transforms. Their assessment has resulted in outstanding output. Both the existing and prior results have been compared using the similar dataset of 152text-pages.

The technique employs the topological idea that there exists a path for every single text line from one side of the image to the other, traversing in that text line only. [24] Have recommended segmentation of handwritten document image into text lines and words.

3. PROPOSED WORK

3.1 Overview

The research presents an improvised version of text line segmentation and skew estimation of Handwritten Kannada document. The recommended method identifies and segregates handwritten text lines thereby computing the skew for the skewed text lines by the means of ESLD (Enhanced Supervised Learning Distance) and R_Clustering algorithms. The ESLD (Enhanced Supervised Learning Distance) algorithm performs the task of gap estimation which involves measurement of gap amidst any two text lines or words. Then is grouping of the connected components that is performed via R_Clustering algorithm. The techniques comprise of the following stages (i) input the document image (ii) preprocessing (iii) text line segmentation and (iv) skew estimation.

3.2 Proposed Methodology

3.2.1 Preprocessing:

The handwritten document image acts as input which undergoes preprocessing. Preprocessing is carried out with the help of (i) Filtering (ii) Gray conversion and (iii) Binarization. This process ascertains that the segmentation accuracy is at par or highest. The received document image depicts the row and column of the image.

Filtering: This process is essential in removing any undesired pixels or noise from the text lines of the input handwritten image so as to carry out the segmentation of the text lines effectively.

Gray scale conversion: There is processing of the document image in order to extract the text line and words. The resultant processed image is in the form of a gray scale image.

***Binarization*:** Transforming the grayscale images into binary images is performed using binarization wherein the foreground or the text is separated from the background information. In other words: In case the intensity value of the image is more than the existing threshold value then the pixels values are changed to 1 else to 0. Thereafter, regions without any text are eliminated from the binary image by traversing the image in top, bottom, left and right directions. This also involves removal of any white pixel areas from the image. Resultant, only the textual part of the image is obtained.

3.2.2 Text line Segmentation

Initially, there is segmentation of the handwritten texts into lines which are then segmented into single words by determining the CCs (Connected Components). Subsequently, there is Connected Components by computing their average width and height.

3.2.2.1 Find the Average width and height of CC

The algorithm employed work by taking the height and width of the complete handwritten word into consideration. Usually, if there is no skew then there should be at least one min value referring to the word's height and one max value referring the word's width. Post skew correction with estimated skew angle repetition of the same process, the busy zones are taken into account for carrying out the skew correction accurately.

Computation of average width, height and line height is performed from the anticipated text lines. To achieve this, Imag1 is vertically split into n equal parts and the avglinehyt is calculated for each text line by considering the text line's heights for the entire document image. For computing the height, total no: of data pixel per row of each vertical partition of array Imag1 is considered. This data is inserted row wise in array CPY for available coordinates. In array CPY, the zero data pixel location refers to the space amidst two consecutive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

text lines which helps in computing the height for the same:

$$SegHytSum(j) = Seghyt1 + Seghyt2 + Seghyt3 + \ldots + Seghytn$$

The max height (maxlinehyt) value and min height (minlinehyt) value of every single text along with its location information is recorded in separate arrays. The minimum width (minwidth) and minimum height (minhyt) related to the character/symbols has been assumed for disconnected diacritics.

- Calculate the height (HYT) and width (WID) of the segment,
- If WID < = minwidth, then the segment is connected, so cut the segment at AvgLinehyt and place it on Image and return

With respect to the handwritten document, the data is referred by the black pattern whereas the background is depicted by the black one. Issue pertaining to overlapped and connected text lines can be resolved by identifying the gap amidst the text lines. On receiving a segment, the height is verified and if the height is less than or equal to the average height of the line then the text line is single else if the height is more, it indicates that the line segment has two or more text lines.

This applies for all the subsequent vertical segmentation and the process is repeated for all the segments individually. The segment is diacritic if its height is less than the min height (minhyt) and is positioned to the nearest segment. If the segment's width turns out to be less or equal to the minimum after applying the vertical segmentation, then it indicates that the two text lines are connected at a particular point. This must be then reduced to a point where there is least data loss since such a wrong text line segmentation can result in incorrect word segmentation. This can cause erroneous feature extraction as well as recognition. Situations where the placement of the connected character is incorrect, maximum height (maxlinehyt) or minimum height (minlinehyt) can be considered for reducing the segmentation errors.

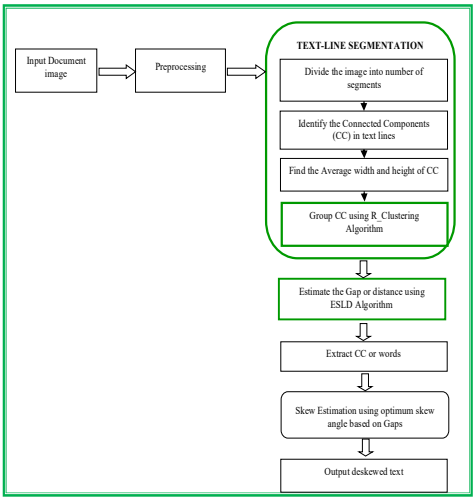

Figure 3.1: Overall Proposed Segmentation Architecture

3.2.2.2 Group CC using R_Clustering Algorithm

The learned distance metric enables components of the same text line to be connected in a sub tree, yet there exist components from different text lines that are connected. These between-line edges are obscure as their lengths may not be more compared to the within-line edge lengths. This situation can be handled with the help of hypervolume wherein the sum of the cluster's hypervolumes of connected components is considered for computing the partition:

$$G_u = \sum_{i=1}^{m} [\det(C_i)]^{1/2}$$

Here $\det(C_i)$ refer to the determinant of the covariance matrix C_i of cluster *I*. It's calculated considering the constituent black pixels of the CCs in the cluster. In the beginning, the spanning tree's components acts as a single cluster wherein every single edge is deleted so that the cluster gets divided into two sub clusters which refers to a disjoint subtree. Edge having the highest reduction of G_V measure is chosen and eliminated so as to reduce the total G_V measure of the document. This is referred as the maximum hyper-volume reduction criterion and indicated as:

$$edge_{deleted} = \arg\max_{edge} \Delta G_u = arg^{max}_{edge}[G_u(F_k) - G_u(F_{k+1})]$$

Where $F_k = \{T_1, T_2, \ldots, T_K\}$ represents the partition of k disjoint subtrees and (F_1 depicts the initial spanning tree).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Evaluation of no: of clusters is not feasible for the G_{tl} measure. The reason being that it always decreases with the increase in the number of clusters. Mostly, the text lines can be thought of having rectangular shapes. If it would be curvilinear, it had to be possibly split into various straight sublines. It's estimated that with an appropriate quantity of clusters (partitioned text lines),the measure of the straightness of text lines is the highest. Following depicts the total straightness measure:

$$F_{um}=\sum_{i=1}^{k}\left(\frac{\lambda_{i1}}{\lambda_{i2}}\right)^2$$

With k resembling the no: of clusters, λ_{i1} and λ_{i2} ($\lambda_{i1} \geq \lambda_{i2}$) are the eigen values of the covariance matrix of each cluster.

3.2.2.3 Connected Components (CC) using R_Clustering Algorithm

Function R_Clustering D (D,S,F, U_{max},V_{max},t): Grouping

Step 1: While the time bound t is not exceeded do

Step 2:% select a 'source' group d_s in random

Step 3: (d_s,F_s)= RANDOM MEMBER (D)

Step 4: D'=D\{(d_s,F_s)}

Step 5: % R clustering the contents of d_s into randomly selected groups

Step 6: For sϵ d_s do

Step 7: (d_t,F_t)= RANDOM MEMBER (D')

Step 8: d_t= $d_t \cup \{s\}$

Step 9: %D' is quite likely not feasible, try to improve it

Step 10: Repeat

Step 11: n_1=unplaced components (D')

Step 12: For (d_i,F_i) ϵD' do

Step 13: F_i =HIBOXES (d_i, F_i, F, U_{max},V_{max})

Step 14: MoveJobs (D')

Step 15: n_2=unplaced Components (D')

Step 16: Until $n_1 \leq n_2$

Step 17: If IS R_clustering(D') then D=D'

Step 18: return D

3.2.2.4 Estimate the Gap or distance using SLD Algorithm

Apparently, various clustering algorithms highly depend on a good metric amidst the pairs of input units hence defining the distance amidst the CCs is essential for ensuring the generated spanning tree comprises components of the same line in a subtree and different subtrees for different lines. Encouraged with the concept of distance metric learning, a distance metric approach is designed to carry text line segmentation using the supervised learning.

To imply the self-designed distance metric among the CCs connected components, training samples of component pairs are taken into account and are tagged as "within line" and "between lines". This is done by marking certain training document images with the help of the truthing tool named as TTLC (Truthing tool for Text Lines and Characters). Basically, it helps in labeling the text lines and characters by automated transcript alignment and hand correction.

Consider a training document with a set of connected components say, F={a_1, a_2,…..,a_n}, with n denoting the no: of components. From this two sets of component pairs are acquired which acts as the samples for metric learning:

S= {(a_i,a_j)|a_i and a_j belong to the same line},

D= {(a_i,a_j)|a_i and a_j belong to different line}

Since it's certain that just the spatially neighboring components are linked in the minimal spanning tree, various component pairs can be discarded from the sample set to increase metric learning. This is achieved by building the area Voronoi diagram of the training document that depicts the spatial adjacency amidst the components. Component a_i can be said to neighbor of the a_j component, if the components share a Voronoi edge on their boundaries. The non-adjacent pairs in the Voronoi diagram are eliminated from S and D. Metric learning emphasizes in reducing the distance amidst the components in S and amplify or increase the distance amidst the components in D pertaining to the learned metric. Therefore, the metric learning issue can be deduced in form of a convex programming problem.

$$\min_{B \in R^{m \times m}} \sum_{(a_i,a_j)} \|a_i - a_j\|_A^2 \quad B \geq 0,$$

$$\sum_{(a_i,a_j)} \|a_i - a_j\|_A^2 \geq 1$$

Here, the distance metric is defined by the matrix mm $B \in R^{k \times l}$(where k denotes the dimensionality

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the feature space describing the component pairs).

$$d(a_i, a_j) = d_B(a_i, a_j) = \left\| (a_i, a_j) \right\|_A = \sqrt{u_{ij}^T * B * u_{ij}}$$

The feature vector being denoted by v_{ij} describes the relation between the two points a_i and a_j. B is achieved by computing the convex programming problem.

3.2.2.5 Supervised Learning Distance (SLD) Algorithm

Input: *A set individual components $d_1, \ldots, d_q$, the training set {(x,y)},the learning rate ϵ , the number of iterations T, the number of unique character k*

Output: *The Gap metric* $G=\sum_{i=1}^{q} w_i d_i(x, x')$

Step 1: Randomly assign initial weights $w_1, \ldots, w_q$

Step 2: for t=1 to T do

Step 3: Let $G^{(t)} = \sum_{i=1}^{q} w_i^{(t)} d_i (x,x')$

Step 4: Apply the SLD Algorithm with $G^{(t)}$ *to get machine learning*

Step 5: Calculate the accuracy $acc^{(t)}$

Step 6: for i=1 to q do

Step 7: Update $w_i^{(i+1)}$

Step 8: end for

Step 9: end for

Step 10: Let t be the round that produces the highest* $acc^{(t^*)}$

Step 11: return G (t^*)

3.2.2.6 Extraction of Connected components (CC)

Each component C_i belonging to the set $\{S\}$ is a part of the sequence. Let the group of components be denoted as $\{C_iN\}$ within the neighborhood NC_i of C_i. The C_i component searches for the neighboring components that can fulfill any one of the conditions:

1. C_i spans at least a fraction $\eta_1$of the height of C_{ij}; j ∈ N, height wise or vice versa.
2. The height-wise midpoints of C_i and C_{ij}; j∈N, have a vertical distance less thana threshold η_2 of the height of C_i or C_{ij}.

If the component C_{ij}, fulfills any of the above-mentioned conditions, the component is assigned to the line belonging to C_i. The boundary lines of $C_i$are initialized with the respective bounds of C_i. Also, the line's area spreads according to the no: of assigned components to the same line. Say for a text line, all the components reside tentatively or completely inside its bounding box. These components can be labelled similar to the text line's label. By doing this, small components can be properly allocated to their respective text lines.

3.2.2.7 Skew Estimation using Optimum Skew Angle

This methodology involves determining the optimum skew angles of CCs by identifying the connected components. At first the Edge Detector is utilized for preprocessing and extracting the edges of the document image. Then there is dilation of the extracted edges via circular structuring element in order to determine the connected components.

Mathematical Model for Optimum Skew Angle

Skew angle estimation is performed by determining the centroid of every single connected component in the document and plotting ellipse on it. Orientation of a CC indicates the angle between the reference axis and principal or major axis around which CC revolves with least possible inertia. Essentially, it should be ensured that the 2nd moment of the region must be similar to the ellipse.

The centroid *MID (K, L)* of Connected Components (CCs) is determined as

$$MID(K, L) = \left(\frac{\sum Orig\, K_i}{N}, \frac{\sum Orig\, L_i}{N} \right)$$

Where$(Orig\,K_i, Orig\,L_i)$depicts the co-ordinates values of pixels within corresponding CCs and N depicts the total no: of pixels within in each Connected Components, 1st order moments are calculated:

$$(L_I) = (Orig L_i - L)$$

Using above mathematical equations, 2nd order moments are

$$\mu_{kk} = \left(\sum (K_i)^2 / N\right)$$

Journal of Theoretical and Applied Information Technology
15th January 2021. Vol.99. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\mu_{ll} = \left(\sum \frac{(L_i)^2}{N}\right)$$

$$\mu_{kl} = \left(\frac{\sum(K_i * L_i)}{N}\right)$$

The moments μ_{ll} and μ_{mm}depicts the variances of l and m respectively. The moment μ_{mm} *is the covariance between l and m.*

$$\text{Skew } (\mu_{ll} > \mu_{mm}) = \frac{\mu_{mm} - \mu_{ll} + \sqrt{(\mu_{mm} - \mu_{ll})^2 + 4 * \mu_{lm}{}^2}}{2 * \mu_{lm}}$$

$$skew(\mu_{mm} > \mu_{ll}) = \frac{2 * \mu_{lm}}{\mu_{ll} - \mu_{mm} + \sqrt{(\mu_{mm} - \mu_{ll})^2 + 4 * \mu_{lm}{}^2}}$$

Optimum skew angle (θ) = (180/Π)$\tan^{-1}(skew)$

4. RESULTS AND DISCUSSION

The proposed deskewing approach is evaluated on handwritten Kannada document images that includes 250 lines and 754 words. The average line segmentation yields an accuracy of 97.13% and the word segmentation yields an accuracy of 93.48%. It's elucidated that the proposed methodology ensures at par performance in carrying out the segmentation of skewed lines and words. The method effectively performs line segmentation by the means of SLD algorithm. It's observed that due to inconsistent spacing between the words and broken characters, there happens to be degradation in the performance of the word segmentation process. Table 1 lists down the output of line and word segmentation.

The proposed algorithm has been evaluated for the handwritten text wherein the data is collected from individuals with diverse backgrounds. Nearly, 4600 Kannada handwritten words are utilized for the experiment out of which 42% are employed for classifier training and the rest being deployed for the purpose of testing. Near about 2700 Kannada words have undergone testing, yielding an accuracy of 96.15%.

Table 2 lists down skew correction accuracy that the classifier exhibits.

Table 1: *Overall Performance of the Proposed Segmentation*

Approaches	*Text lines*	*Words*	*Accuracy*
Input	**250**	**754**	**97.13%**
Proposed Segmentation	**241**	**694**	**93.48%**

The result of text line segmentation is given in Figure 2

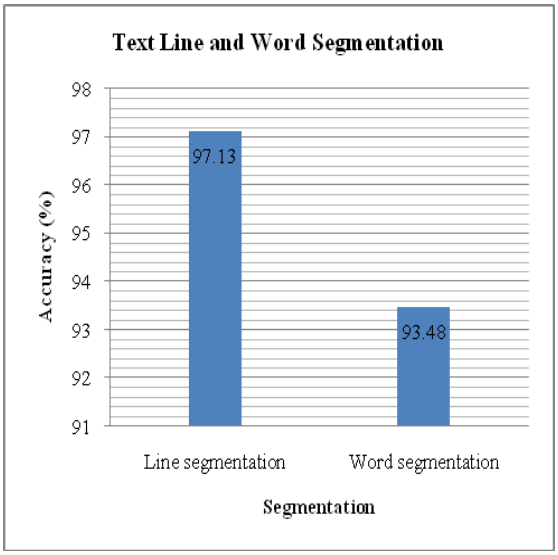

Figure 2: *Overall Results of Proposed Method for Segmentation of Lines and Words*

Table 2: Results of Skew Correction Algorithm

Total Words	*Skewed words*	*Corrected*	*Incorrected*	*Skew Correction Accuracy*
4600	**2700**	**3180**	**87**	**96.15%**

Five documents are chosen for distance metric learning and for testing, rest of the 100 documents comprising of 1106 text lines are considered. Table 3 depicts the actual rates of text line detection using the SLD algorithm including and excluding the metric learning. It's clearly proven that there is significant increase in the performance of text line segmentation with the usage of distance metric learning. Albeit, performance of SLD algorithm along with metric learning is remarkable enough

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

but it does confront certain text line detection failures. These failures can be either of the two: i) Error line splitting: an actual text line is divided into two or more lines (depicting multiple clusters); ii) Error line merging: two or more actual text lines are combined to form a single cluster.

Table 3: Corrected rates of text line detection using SLD

Different Metric	*Detected Text Lines*
With learned metric	**1087 (96.56%)**
With metric by hand	**897 (87.18%)**

5. CONCLUSION AND FUTURE WORKS

The present research has proposed the approach of SLD and R_Clustering algorithm for line and word segmentation in context to an unconstrained hand written Kannada document. Here the words belonging to the text lines are grouped with the help of an intelligent technique. The identified words are then extracted and stored in a new image. While extraction, it's ensured that there is no overlapping of words and that the unwanted information is eliminated properly. There are other proposed and thoroughly experimented methods mentioned in the research to improvise the process of text line segmentation and skew correction of Kannada handwritten documents. The average segmentation rate reported is 97.13%.

The proposed method is capable and effective enough in determining the optimum angle, deskewing of the word as well as storing the words in the line appropriately. The methods pertaining to line and word segmentation ascertains at par performance in spite of character size variation and existence of consonant and vowel modifiers. Although, the word segmentation process can be hampered with existence of space resemblance between the words and characters.

REFERENCES

[1] Manmatha R, Rothfeder J.L (2005), "A scale space approach for automatically segmenting words from historical handwritten documents",*IEEE Transaction on Pattern Analysis and Machine Intelligent*, Vol. 27, No. 8, pp. 1212–1225.

[2] Li Y,Zheng Y,Doermann D and Jaeger S (2008), "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 8, pp. 1313-1329.

[3] Avanindra and Subhasis Chaudhuri, (1997) "Robust Detection of Skew in Document Images," *IEEE Trans on Image Processing*, Vol.6, No.2 , pp. 344-349.

[4] LioliosN, Fakotakis N, and KokkinakisG (2001) "Improved document skew detection based text line connected component clustering" *IEEE Vol.1,* pp. 1098-1101.

[5] Nagabhushan P, Angadi SA, and Anami,B.S (2007) "Geometrical model and projection based algorithms for tilt correction and extraction of ascender descenders for cursive word recognition "*IEEE ICSCN*, pp. 488-491.

[6] Sanchez A, Suarez P.D, Mello C.A,Oliveira A.L.I, and Alves V.M.O(2008) "Text Line Segmentation in Images of Handwritten Historical Documents," *IEEE Computer Society*.

[7] LeCun Y, Bottou L, Bengio Y, andHaffiner P (1998) "Gradient based learning applied to document recognition", *Proceedings. of IEEE*, Vol.86, pp. 2278–2324.

[8] Suen CY, Nadal C, Legault R, Mai T, and Lam L (1992) "Computer Recognition of Unconstrained Handwritten Numerals,"*Proceedings of the IEEE*, Vol.80, pp. 1162–1180.

[9] Hull.J(1994)"A database for handwritten text recognition research," *IEEE Transactions on PAMI*, Vol.16, 1994, pp. 550–554.

[10] Bhattacharya.U, and Chaudhuri.B.B (2009), "Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals," *IEEE Transactions on PAMI*, Vol. 31, pp. 444–457.

[11] Otsu N,(1979) "A threshold selection method from gray-level histograms",*IEEE Transactions on System, Man. and Cybernetics*, Vol.9, pp. 62–69.

[12] Xiang D, Yan H, Chen X, and Cheng Y (2010), "Offline Arabic Handwriting Recognition System Based On HMM," *IEEE*, pp. 526-529.

[13] EI Yacoubi A.M, Gilloux M, Sabourin R, and Suen C.Y,(1999) "An HMM based Approach for Offline Unconstrained Handwritten Word Modeling and Recognition" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 8,.PP.752 -760.

[14] Pal, U., and Datta, S. "Segmentation of Bangla unconstrained handwritten text." *IEEE*, 2003.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] Basu, Subhadip, et al, (2007) "Text line extraction from multi skewed handwritten documents."*Pattern Recognition* Vol.40, pp.1825-1839.

[16] Naveena C, and Manjunath Aradhya VN, "Handwritten character segmentation for kannada scripts."Information and Communication Technologies (WICT), World Congress on *IEEE*, 2012. PP.144 - 149

[17] Park, June-Me, Carl G. Looney, and Hui-Chuan Chen. "Fast connected component labeling algorithm using a divide and conquer technique", *Computers and Their Applications,* 2000.

[18] Alaei A, Pal U, and P. Nagabhushan P,(2011) "A new scheme for unconstrained handwritten text-line segmentation", Pattern Recognition, Vol.44, pp. 917–928.

[19] Basu S. C, Chaudhuri C, kundu M, Nasipuriand Basu D.K (2007),"Text line extraction from multi-skewed handwritten documents",*Pattern Recognition*, Vol.40, pp. 1825–1839.

[20] Yin F, and C.L Liu (2009), "Handwritten chinese text line segmentation by clustering with distance metric learning",*Pattern Recognition*, Vol. 42, pp. 3146–3157.

[21] Liwicki M, and Bunke H (2009), "Combining diverse on-line and off-line systems for handwritten text line recognition", *Pattern Recognition*, Vol.42, pp. 3254–3263.

[22] Louloudis G, Gatos B, Pratikakis I, and Halatis C (2009), "Text line and word segmentation of handwritten documents",*Pattern Recognition*, Vol.42, pp. 3169–3183.

[23] Louloudis G, Gatos B, Pratikakis I, and Halatsis C(2009), "Text line detection in handwritten documents",*Pattern Recognition*, Vol.41, pp. 3758–3772.

[24] Papavassiliou V, Lakis T.S, Kastsourous,V, and Carayannis G(2010),"Handwritten document image segmentation into text lines and words, *Pattern Recognition*, Vol.43, pp. 369–377.