International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 23 Issue 6 –OCTOBER 2016 (SPECIAL ISSUE).

DESKEWING ALGORITHM FOR UNCONSTRAINED HANDWRITTEN KANNADA DOCUMENTS LEADING TO LINE AND WORD SEGMENTATION

S.A.Angadi [#1] and *Sharanabasavaraj.H.Angadi* [*2]

[#1] *Professor of Computer Science and Engineering, Visvesvaraya Technological, University, Belagavi, Karnataka, India*

[*2] *Department of Computer Science and Engineering, Rural Engineering College, Hulkoti, Karnataka, India*

Abstract— **Extraction of lines and words from handwritten document images containing skewed text is one of the most difficult and challenging problem. In this paper, a new deskewing algorithm leading to line and word segmentation from an unconstrained hand written Kannada documents is proposed. The method employs preprocessing, dilation and labelling the connected components of input image as initial step. Then an intelligent technique is used to group the words belonging to a text line. The extracted words are subjected to removal of unwanted information that pertains to adjacent words. Further the skew angle computation and rotation operation (when angle is other than zero) are performed for purpose of deskewing of extracted words. Then deskewed words are intelligently written into new image without overlapping of words in text line. The method also takes care of detecting text lines containing consonant modifiers. Inter word and intra character gap variations are also taken care at the time of word segmentation by the proposed method. The method is evaluated for 20 images having 166 lines and 823 words. A line segmentation accuracy of 96.38% and word segmentation accuracy of 92.10% is achieved. The method works efficiently for document images containing skew upto four degrees. The method also works correctly even in the presence consonant and vowel modifiers, compound characters and uneven spacing between characters and words.**

***Index Terms*—** Connected components, Kannada hand written document image**,** Labels, Projection profile features, Segmentation, Skew detection and correction.

I. INTRODUCTION

Human beings are perennial optimizers hence they are on the lookout of technologies for making their tasks easier. To fulfill such requirements automation is one of the foremost tools. Automation systems are required for various applications that are essential in our daily routine. One such application is an optical character recognition (OCR) system that identifies characters present in digital images of either printed or handwritten text. Such OCR systems find various applications in the field of banking, security, postal system, mobile computing systems and so on. Generally the recognition of unconstrained handwritten characters by an OCR system is quite difficult as compared to printed characters. This may be due to many challenges such as size variations, variations of space between the characters, line variation, and presence of skew, presence of noise and so on. Many researchers have proposed various techniques towards the development of efficient hand written recognition system. The literature survey indicates the existence of large amount of work with respect to the recognition of European languages and to a little extent for the recognition of Indian languages like Hindi, Bangla, Gurumukhi, Tamil etc, where as comparatively less work is reported for Kannada language. Therefore scope exists for the development of an efficient automatic handwritten recognition system for Kannada language. The handwritten recognition system comprises of pre-processing, feature extraction and recognition stages. The pre-processing stage performs digitization, skew detection and correction, segmentation of lines, words and characters. The feature extraction stage is used to extract the unique features from the pre-processed document image. Finally character identification is performed by the recognition /classification stage. Amongst these, pre-processing stage plays an indispensive role for the satisfactory performance of the later stages of hand written recognition system. One of the challenges faced by the pre-processing stage is the writing of majority of people will not be in a straight line but will generally have certain inclination. The inclination with respect to horizontal is referred to as skew. The existence of skew in a document leads to difficulty in segmentation of lines and words. Therefore, a proper deskewing mechanism is required for reliable line and word segmentation of hand written document to improve the recognition rate. Deskewing of hand written Kannada document is a difficult and a challenging task due to presence of inflection in Kannada script. The Kannada script contains vowels, consonants, modifiers and also compound characters. The task of segmenting the line from the document image is also difficult due to the presence of top and bottom modifiers. Also word segmentation is more challenging due to the various sizes of characters as well as the varying inters a word gap that exists between the words. The next section describes the various methods for skew detection, correction and segmentation of lines and words which are reported in literature. Not many

works related to deskewing and segmentation of lines and words from handwritten Kannada document images are found in literature. The proposed research addresses the said issue of deskewing and segmentation of lines/words of handwritten Kannada document. In this paper, a deskewing algorithm leading to line and word segmentation from an unconstrained hand written Kannada documents is proposed. The proposed method employs preprocessing, dilation and labeling the connected components of input image as initial step. Then a intelligent technique is used to group the words belonging to text lines. Further the skew angle computation and rotation operation (when angle is other than zero) are performed for purpose of deskewing of extracted words. Then deskewed words are intelligently written into new image without overlapping of words in text line. The method also takes care of detecting text lines containing consonant modifiers. Inter word and intra character gap variations are also taken care at the time of word segmentation by the proposed method. The methodology is evaluated on the document images containing handwritten Kannada text written by different persons. The document contains 166 lines and 823 words. The document images are created by scanning documents using the HP flat bed scanner with 300 dpi. The performance of the method is found to be best in terms of segmentation of skewed lines and words of handwritten Kannada document images. The method works well for segmentation of lines. However, the inconsistent spacing between the words and presence of broken characters may affect the performance of word segmentation process. It is found that the accuracy of the system is 96.38% for text line segmentation and 92.10% for word segmentation. The method works intelligently even in the presence of consonant and vowel modifiers and uneven spacing between characters and words. This paper is organized into five sections. Detailed literature survey related to the skew detection and correction and also the line and word segmentation is brought out in Section-2. Section-3 presents the proposed methodology. The experimental results are brought out in Section-4. Section-5 gives the conclusion.

II. LITERATURE SURVEY

The reliable and efficient skew detection and correction method is indispensible for the satisfactory performance of later stages of handwritten recognition system. Researchers have been contributing various methods to develop a robust and intelligent skew detection and correction schemes. Some of such techniques are summarized below. A novel approach to estimate the skew from the scanned Persian document is proposed in [1]. The method transforms the document image to the text block image using the connected component analysis and morphological closing followed by thinning operations. Further rectangular patch covering thin line is subjected to the thinning operations. Then the skew angle is estimated from the slopes of the thinned lines. An algorithm to estimate the skew angle in a document image using Hough transform is presented in [2]. The method is experimented on the document image containing Chinese text. a scheme based on orthogonal projection to estimate the skew angle from the hand written english document image is proposed in [3]. The author states that the method gives the same accuracy even when it is applied to the different languages. A simple method for skew detection is proposed in [4]. The method clusters the components belonging to the same line. The peak angle between the centroid of the connected components will be treated as a skew angle. The method operates with less computational overhead. A new algorithm for skew and slant correction based on geometrical model and projection is presented in [5].The method is experimented using hand written English document. A method for skew angle estimation of printed or hand written document is presented in [6]. It uses wigner-ville distribution for horizontal projection profile. Atomic decomposition and energy distributions are represented by using Wigner ville distributions. A method for detecting and correcting the skew of printed document using bilinear interpolation method is discussed in [7]. Discrete cosine transform helps to reduce the computation time and further the skew angle is detected by applying fft to the four quadrants of the image. An efficient and simple method based on boundary growing, thinning and moment to estimate the skew angle is presented. The method claims to be best in terms of accuracy, computational time and so on. [8] Discusses a novel scheme based on radon transform to estimate the skew. The method is experimented on printed Kannada documents. The method claims to provide better accuracy and execution speed. A robust method for text line segmentation is proposed in [9]. the method generates a modified histogram from the run length smearing. The proposed method separates lines and words of bangla, Devanagari and Telugu scripts and results are promising. Line and word segmentation of document images using Hough transform technique is proposed in [10]. The technique is experimented on the hand written as well as printed document images. But the technique fails to segment lines properly when there is a large and nonuniform separation of lines and even in case of narrow spacing between the lines. Error occurs during segmentation of words due to the variation of inter character spacing between the words and also due to the existence of very closely spaced characters. A method to segment hand written text document into lines, words and characters is discussed in [11]. The peak valley points of the histograms are obtained by dividing text into vertical stripes and are used for segmenting text lines. Segmentation of the words are achieved by employing the vertical projection profile and structural features. Narrow spaced characters (touching characters) are segmented using water reservoir, structural and topological features. The method has been experimented on oriyan text. A robust segmentation method for handwritten word is presented in [12]. The method works on Arabic words. The extraction of words is achieved by finding peak segment borders from chain code and skeleton. A text line segmentation based on Hough transform is presented in [13]. The technique is experimented on hand written English and Hindi text. A robust scheme to segment line, word and characters of bangla language is described in [14]. Water reservoir principle is employed for segmentation purpose. a new approach uses neighbourhood tracing algorithm and projection profile to extract characters from the word. The method is examined on hand written Devanagari document. From the literature survey, it can be deduced that most of the work that has been reported on segmentation of lines and words of documents pertain to roman/english script documents. Few works have been reported on Indian languages like bangla, Tamil, Hindi and so on. However, not much work is reported in segmentation of skewed lines and words of document images containing Kannada script.

Therefore there is a scope towards the development of proper deskewing mechanism for reliable line and word segmentation of hand written Kannada document to improve the recognition rate. The proposed methodology for deskewing algorithm leading to segmentation of lines and words from handwritten document images is described in the ensuing section.

III. PROPOSED METHODOLOGY

The proposed deskewing method leading to line and word segmentation from an unconstrained hand written Kannada document image consists of following processing steps such as preprocessing, dilation and labelling, grouping of words of text lines, deskewing of words belonging to a text line and inserting words to the new image. The description of each of these stages is given in the following subsections. The block schematic diagram of the proposed method is given in figure 1

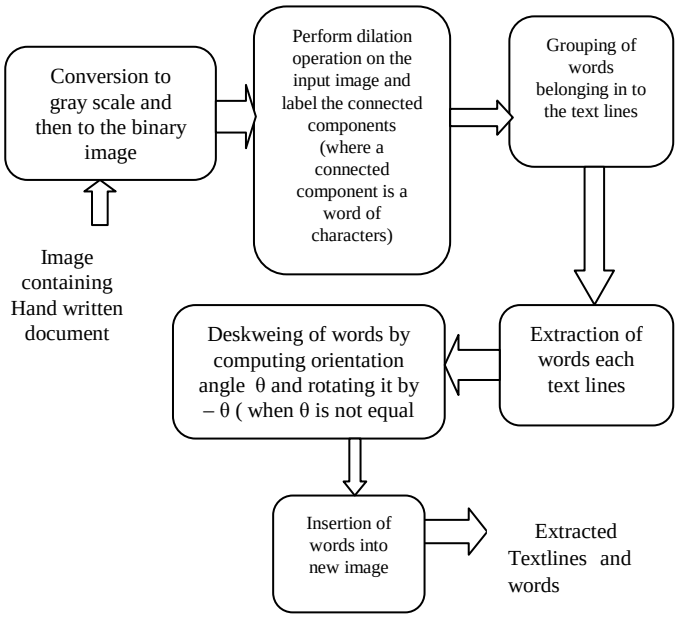

Figure.1. Block schematic diagram of the proposed method of deskewing text line & word segmentation model.

A. *Conversion to Gray scale and Binary image*

The document image is to be processed to extract text line and words. First the image is converted to gray scale image, which is thresholded to obtain a binary image.

B. *Dilation and labelling the connected component*

For the purpose of proper segmentation of lines and words, the determination of skew (orientation) angle present in the hand written text document is important. The angle can be determined by considering individual character or isolated word or entire line as a image. Each of these approaches have some advantages and disadvantages. Finding orientation angle by using individual character creates complexity in the process of saving images as a word due to the presence of compound characters. On other hand determination of orientation angle for the entire line creates a difficulty due to the variation of the orientation angle between the words. A sample input image containing handwritten Kannada text is shown in figure 2a. The better option is to use word to determine the skew angle. For this purpose, the dilation process is performed on the binary image so that the words in the document become connected components and is shown in figure 2b. Further the dilated images are labelled, which are used in further processes.

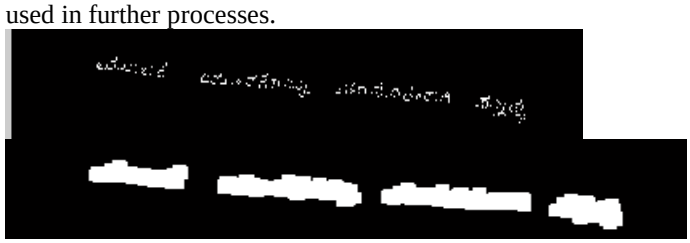

Figure 2a. A sample handwritten Kannada text line Figure.2b A sample dilated image of figure 2a

C. *Grouping of words*

The labeling process assigns the labels column wise so in this stage, a new technique to identify the words belonging to a text line is employed. For the sake of grouping the words, every dilated component is processed to find the boundary coordinates such as rmin, rmax, cmin, cmax and also the label number. Where rmin is the minimum row, rmax is maximum row, cmin is minimum column, and cmax is the maximum column.

The boundary coordinates for all the connected components are stored in the information base IB. The rmin value in a set of boundary coordinates represents the position of word in the document image and this can be used for grouping of words belonging to same text line. Hence in this work a new mechanism is devised that exploits the rmin values for grouping of words belonging to text line. Initially sorting operation is done on all the values of rmin attribute and stored in the new vector KB1. By comparing absolute difference of two successive rmin values with the experimental observed threshold value (79), the words are grouped into a line/row. If the difference is less than the threshold then the rmin value is stored in the same row of matrix KB2. Otherwise rmin value is stored in the next row of the matrix KB2. The procedure is repeated for the remaining dilated components. The grouped words belonging to text line is shown in figure 3.

D. *Extraction of words*

In this stage, words belonging to the text line identified by the row element of matrix KB3 are

Figure.3 A sample image comprising words belonging to a text line

Extracted to determine the skew. The extraction process is achieved by applying the boundary coordinates over the input document image as indicated below.

E. *Deskewing of words*

In this stage computation of the orientation (skew) angle θ for the labeled component is carried out. If is not equal to zero

then is rotated by an angle of –θ. The figure 4a shows the word containing skew and deskewed word after rotating by an angle of - θ is shown in figure 4b.

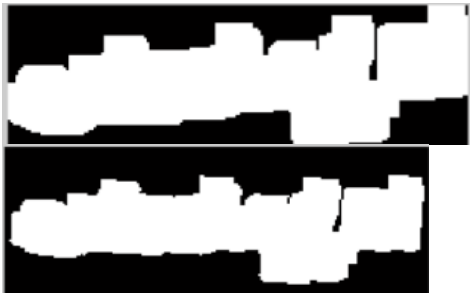

Figure 4a Skewed sample image
Figure.4b A deskewed sample image

F. Insertion of words into a new image

During the process of deskewing of word, coordinate values of deskewed image may alter. Therefore, care is taken in this stage to avoid the overlapping of words in to a new image by comparing the boundary coordinate cmax of previous word with the cmin of current word image. If (cmax is less than cmin) then cmin is updated as below.

$Cmin_{j+1}$ =cmaxj+constant. 1<j<=wc (2)

The words/connected components belonging to a text line are regrouped on a new image by making all the rmin equal to the minimum amongst all the rmin values of the components belonging to the text line. The other parameters of the component are accordingly translated using deskewing process so that all the words/connected components lie on the same horizontal row.(refer figure 5,6a and 6b).

Figure.5 A sample image containing words positioned at at different location due to the skew
Figure.6a. A sample image containing words positioned at same row location
After deskewing

The procedure is repeated for the remaining words of text line. Figure 6b shows image containing words situated at same position.

Figure.6b A sample image containing words belonging to text line after deskewing

The procedure repeats for the remaining text lines. The detailed experimental analysis is provided in the next section

IV. EXPERIMENTATION

The proposed deskewing algorithm leading to line and word segmentation is implemented and tested on hand written Kannada document images. The document is written on A4 size plain paper in an unconstrained manner. It may contain top, bottom and base modifiers and also includes compound characters. Such document is used to evaluate the system performance. The document is scanned on a hp scanner with resolution of 300 dpi. The images in bmp format are employed. A sample document image having the skew is shown in figure 7. The document containing skewed line is deskewed using proposed novel deskewing algorithm and is shown in figure 8. The words are segmented using horizontal projection profile features. The image contains 166 lines and 823 words. Due to the unconstrained writing, the document may have skewness within the lines, uneven spacing between words and presence of compound characters..

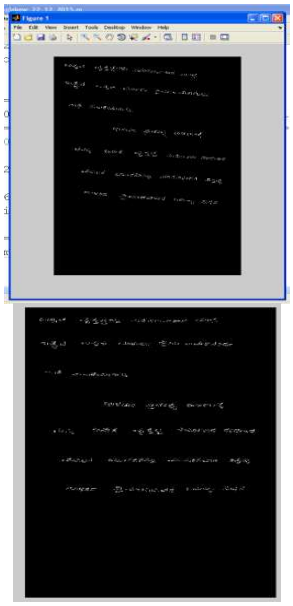

Figure 7 . Skewed document Image
Figure 8. Deskewed document image

Initially the hand written Kannada document image undergoes dilation operation after preprocessing. Then the labels are assigned to the connected components which facilitate the identification of the words belonging to particular line and are shown in figure 9 and figure 10 respectively.

Figure 9. A sample of dilated hand written kannada document image
Figure 10. Labelled handwritten kannada document image

The boundary coordinates of each of the labelled components are computed and updated in the information base IB=[IB:rmin,rmax,cmin,cmax,rmax,centroid,label number].
The information base is shown in figure 11.

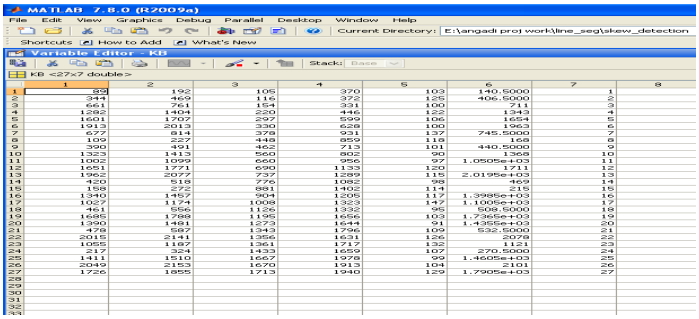

Figure 11. Information base of sample input image containing boundary coordinates

Further the information base is sorted using rmin as key. The sorted values are stored in KB1[] and shown in figure 12.

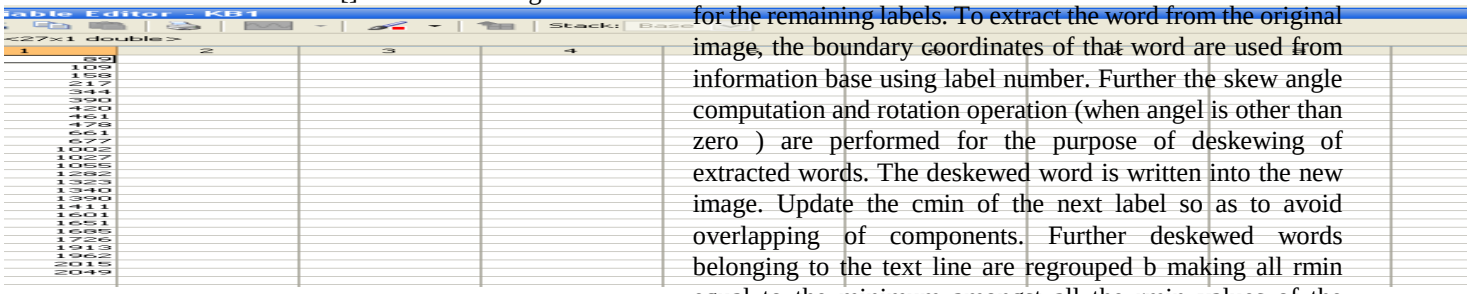

Figure 12. Information base of sample input image containing sorted rmin

Then labels that are belonging to particular line are grouped by comparing successive rmin with threshold. The threshold is selected through the experimentation. The label belonging to the each row is identified and is shown in Table 1.

TABLE 1. THE IDENTIFIED CONNECTED COMPONENTS BELONGING TO DIFFERENT ROW

1st row (rmin)	**89**	**109**	**158**	**217**	
2nd row (rmin)	**344**	**390**	**420**	**461**	**478**
3rd row	**661**	**677**			
4th row	**1002**	**1027**	**1055**		
5th row	**1282**	**1323**	**1340**	**1390**	**1411**
6th row	**1601**	**1651**	**1685**	**1726**	
7th row	**1913**	**1962**	**2015**	**2049**	

The computation is done to identify the corresponding labels and is shown in Table 2

TABLE 2. THE IDENTIFIED CONNECTED COMPONENTS LABELS BELONGING TO DIFFERENT ROW (ROW WISE)

1st row	1	8	15	24	
2nd row	2	9	14	18	21
3rd row	3	7			
4th row	11	17	23		
5th row	4	10	16	20	25
6th row	5	12	19	27	
7th row	6	13	22	26	

Once labels related to a line are identified, the words in a text line are obtained. Further determination of skew angle for labelled component is carried out. The challenging task is to identify whether line is having positive or negative inclination. To solve these problems, consider the lowest and highest label among the labels of that particular line. Check rmin between these two labels. Fix the lowest rmin as top row for the remaining labels. To extract the word from the original image, the boundary coordinates of that word are used from information base using label number. Further the skew angle computation and rotation operation (when angel is other than zero) are performed for the purpose of deskewing of extracted words. The deskewed word is written into the new image. Update the cmin of the next label so as to avoid overlapping of components. Further deskewed words belonging to the text line are regrouped b making all rmin equal to the minimum amongst all the rmin values of the components belonging to a text line. The other parameters of the components are accordingly translated to the deskewing process so that all the words/connected components lies on the same row. The text line is intelligently written in to new image. The procedure is repeated for the other labelled components.

The proposed methodology is implemented on Matlab software and evaluated for handwritten Kannada text document images on Intel Pentium processor (2.1GHz, 830MHz FSB) computer. A sample line and word segments are shown in figure 13 and figure 14.

Care is taken to avoid this kind of undesirable information in this proposed work by employing dilation and label assignment i.e, when labels are more than one so it contains other undesirable information.So technique is used to extract desired word only. Find orientation angle θ. Rotate it by an angle- θ in the case when angle θ is not equal to zero. Place the the content in a new image with lowest rmin among the labelled components of that current line. Update the cmin of next label so as to avoid overlapping of components. The procedure is repeated for the other labelled components. Once the line is extracted then proceed for the segmentation of the word using horizontal projection profile. Set the threshold by comparing gap that exists between inter character and inter word to identify word. The proposed methodology is implemented on Matlab software and evaluated for handwritten Kannada text document images on Intel Pentium processor (2.1GHz,830MHz FSB) computer. A sample line and word segments are shown in figure 13 and figure 14

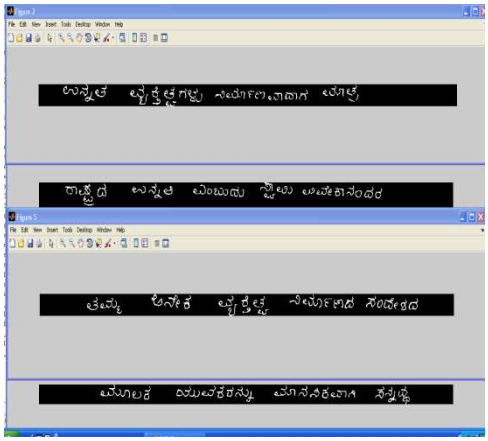

Figure 13 Sample Hand written Kannada document image with segmented lines.

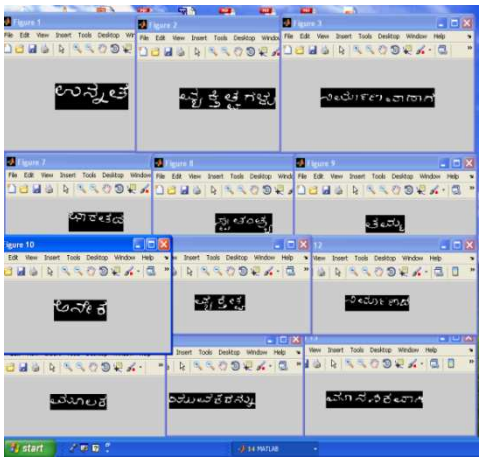

Figure 14 Sample Hand written Kannada document image with segmented words

The proposed deskewing approach is evaluated on handwritten Kannada document images containing 166 lines and 823 words. A sample test document image is shown in Figure 7. The average line segmentation accuracy of 96.38% and word segmentation accuracy of 92.10% is achieved by the proposed method. The performance of the proposed method is found to be efficient in segmentation of skewed lines and words. The method works well for segmentation of lines containing variability of skew upto 4 degree and also in presence of compound characters. However, the inconsistent spacing between words and broken characters may affect the performance of word segmentation process. The line and word segmentation results are shown in Table.3

Table.3 Overall system performance of the proposed method

#lines in text document images	#segmented lines using proposed method	Accuracy of the proposed method
166	**160**	**96.38%**
#words in text document images	#segmented words using proposed method	Accuracy of the proposed method
823	**758**	**92.10%**

The results of text lines segmentation is given in the Figure 15.

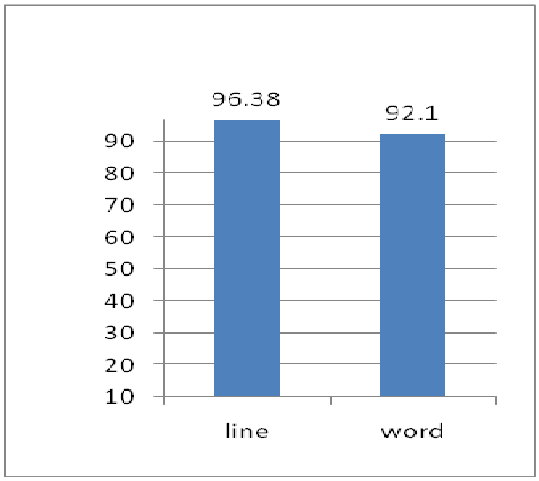

Figure 15. Overall result of proposed method for segmentation of lines and words.

V. CONCLUSION

In this paper, a deskewing algorithm leading to line and word segmentation from an unconstrained hand written Kannada documents is presented. The method employs an intelligent technique to group the words belonging to text line. The identified words are extracted and stored in a new image. Care is taken to remove the unwanted information at the time of extraction of words by bounding box approach and also to avoid the overlapping of words during the storing in a new image file. Similarly the method works efficiently in detecting the orientation and in deciding the deskewing of the oriented word and also in storing words in a proper aligned line. Even with the variation of character size and also presence of consonant and vowel modifiers, the proposed line and word segmentation method gives good performance. The existence of space resemblance between words and characters will affect during word segmentation process. The proposed methodology is experimented on hand written Kannada document images written by different people with different skews. It gives line segmentation accuracy of 96.38 % and word segmentation accuracy of 92.10 %. The method is found to be efficient as results in Table.3 demonstrates viz, the method reliably segments 160 lines from 166 lines and 758 words from 823 words. The proposed method segments the lines and words from the documented image having four percentage of skew. Further it is planned to extract the compound characters as well as isolated characters.

REFERENCES

[1] Ashkhan. M.Y, D.S.Guru, P Punitha, 2006,"Skew estimation in Persian documents: A novel approach", proceedings of international conference on computer graphics, imaging and visualization,2006

[2] Tian Jipeng, G Hemanthkumar, H.K.Chetan, 2011,"Skew correction for Chinese character using hough transform", international journal of computer application Vol. 21 , No. 2, pp. 33-36 May 2011

[3] Subhas Panwar, Neeta Nain,2012,"A novel approach for skew normalisation for handwritten text lines and words", eighth international

conference on signal, image technology and internet based systems, pp. 296-299, 2012

[4] N Liolios, N Fakotakis and G Kokkinakis ,2001, "Improved document skew detection based text line connected component clustering",IEEE pp 1098-1101,2001

[5] P Nagabhushan ,S A Angadi , B.S.Anami, 2007, "Geometrical model and projection based algorithms for tilt correction and extraction of ascender/descenders for cursive word recognition ",IEEE ICSCN 2007,MIT Campus,Anna University ,Chennai,pp. 488-491,Feb 22-24 2007.

[6] E Kavallieratou,N Fakotakis and G Kokkinakis,2002,"skew angle estimation for printed and hand written documents using Wigner Ville distribution",ELSEVIER,image and vision computing, Vol.20, pp. 813-824 2002.

[7] Mandip kaur,simple Jindal,2013,"An integrated skew detection and correction using FFT and DCT", international journal of scientific and technology research ,Vol. 2, issue 12, pp. 164-169, Dec 2013

[8] Prakash K Aithal,Rajesh G ,Siddalingswamy P C ,Dinesh V Acharya,2011," A novel skew estimation approach using Radon transform",11th international conference on Hybrid intelligent system pp. 1-4. IEEE 2011.

[9] Nallapareddy Priyanka,Srikant Pal,Ranju Mandal,2010," line and word segmentation approach for printed document",IJCA special issue on recent trends in image processing and pattern recognition,pp 30-36,2010.

[10] Satadal Saha,Subhdip Basu,Mita Nasipuri,Dipak Kr Basu, 2010,,"A Hough transform based technique for text segmentation",journal of computing,vol-2,issue-2,pp 134-141 Feb 2010.

[11] N Tripathy, U Pal ,2006,"handwriting segmentation of unconstrained Oriyan text", Sadhana, Vol. 31, Part-6, pp. 755-769,Dec 2006.

[12] Satwan W Shah,Zhixin Shi and Venu Govindaraju,2009,"segmentation of Arabic handwriting based on both contour and skeleton segmentation",10th international conference on document analysis and recognition ,pp 793-797,2009.

[13] Sunanda Dixit, Sneha, Nilotpal Utkalit,Suresh H N , 2014,"Text line segmentation of handwritten documents in Hindi and English ", international journal of Innovation trends in computing and communication ,Vol.2, issue 4, pp. 733-739 April 2014

[14] U Pal and Sagarika Datta ,2003,"segmentation of Bangla unconstrained handwritten text", proceedings of seventh international conference on document analysis and recognition 2003