[VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY] Vol.4. No. 1, July, 2015
83 - 94,ISSN : 0975-5446

Words Extraction from printed Bilingual Gujarati-Roman Script Documents

CHAUDHARI Shailesh A.
M.Sc.(I.T.) Programme
Veer Narmad South Gujarat University
Surat
Shailesh_mca@yahoo.com

GULATI Ravi M.
Department of Computer Science,
Veer Narmad South Gujarat University
Surat
rmgulati@gmail.com

Abstract
Accuracy of any OCR directly depends on correct segmentation of various parts of document image (like line, word and character). Word extraction is difficult specifically, when document is multi-script and varying font size and style is used. In this paper we present a script independent word extraction technique from printed bilingual Gujarati-Roman documents. This approach is based on size of the structuring element for morphological dilation. In this approach first we count space between characters and words. Then we use three different methods median, average and standard deviation to calculate the size of the structuring element. Moreover we also use different shapes of structuring element like disk, diamond, and square along with aforementioned three methods. We report results for Gujarati, Roman, and bilingual Gujarati-Roman scripts.

Keywords:- Script, Morphological Dilation, Structuring Element, Segmentation, intra-spaces.

1. Introduction

In a multilingual country like India, one can notice Roman mixed with regional scripts and occasionally both Roman and Gujarati are mixed with the regional script. Segmentation of a printed text line into words is an important component of a variety of document manipulation systems, such as optical character reader (OCR), document imaging system, keyword spotting system for digital library, and so on. Existing methods for word segmentation accept a gap-based approach which assumes that there is a significant gap between adjacent words. But, in practical situations, there are many difficulties. First, the magnitude of word gap varies considerably from one image to another depending on the scanning resolution, language in the document, word processor and font style, and so on. The size of a word gap is quite different between Gujarati and Roman text lines. This fact implies that accurate gap-based word segmentation should be required to these variations. In this paper we propose a word segmentation method composed of gap

clustering techniques. This method is evaluated for printed bilingual Gujarati and Roman documents, and an encouraging accuracy is achieved.

2. Related Work

Document Image Analysis is concerned with the problem of transferring the document images into electronic form. This would involve the automatic interpretation of images of printed and handwritten documents, including text, forms, postal envelopes, bank cheques, engineering drawings, maps etc [1]. Several systems which work in specific domains, like the ones mentioned above, have been developed. Document Image Analysis can be defined as the process that performs the overall interpretation of document images [2]. The main problem is to segment a document page into text, figures, tables, etc known as page segmentation. Lot of research has done into solving the problem of page segmentation and a number of algorithms have been proposed for the same. Page segmentation algorithms can be categorized into three classes: top-down approaches, bottom-up approaches and hybrid approaches [3]. Top-down algorithms start from the document image and iteratively split it into a number of smaller images. The splitting procedure stops when some criterion is met. Examples of top-down approaches are X-Y cut [4] and the shape-directed-covers-based [5] algorithm. Bottom-up algorithms start from document image pixels and cluster the pixels into connected components which are then clustered into words, lines, or final zone segmentations. Examples of bottom-up approaches are the Docstrum algorithm [6], the Voronoi diagram based algorithm [7] and the run-length smearing algorithm [8] and the text string separation algorithm [9]. Hybrid approaches are a mixture of the above two approaches. The split-and-merge algorithm [10] is one such algorithm. A gap clustering technique is used to identify the gaps between words regardless of the gap-size variations among different document images [11]. A complete line and word segmentation system for some popular Indian printed languages is presented [12]. A system for Gujarat handwritten numerals using four different profiles namely; vertical, horizontal and two diagonal profiles feature extraction techniques with multilayer feed forward neural network [13]. A line extraction and line wise script identification based on statistical features is presented [14].

3. Properties of Gujarati Script

Gujarati language stands at the 26th position among the most spoken native language in the world and nearly 50 million people throughout the world speak Gujarati. The basic direction of writing Gujarati is from left to right and top to bottom, the same as English. Altogether Gujarati alphabets utilize 94 symbols, which can be categorized into different groupings. Gujarati character set provides 34 (+2 compound *ksha, gna*) consonants, 14 vowels which are represented by a single symbol as shown in fig. 1 (a, b).

ક	ખ	ગ	ઘ	ઙ
ચ	છ	જ	ઝ	ઞ
ટ	ઠ	ડ	ઢ	ણ
ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ
ય	ર	લ	વ	
શ	ષ	સ		
હ	ળ	ક્ષ	જ્ઞ	

Figure 1.a: Gujarati Alphabets

Figure 1.b Gujarati alphabet with vowel modifiers

4. Properties of Roman Script

The Roman script uses two bipartisan spellings for each character, one called lowercase, the other called uppercase or capital as shown in fig. 2.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z

Figure 2. Roman Alphabets

Roman alphabet consists of two types of graphemes, vowels and consonants. In general, each grapheme possesses these two types of spelling with few exceptions changing from script to another. The Roman script is written from left to right, it's a non-cursive; its letters are isolated from one to another, separated by intra-words spaces in its printed form. The Roman alphabet has also many loops that can have different forms.

5. Morphological Dilation

Dilation is one of the basic operations in mathematical morphology. It is originally developed for binary images. Morphological dilation on a binary image gradually increases the boundaries of the image region. Let us assume that A is an image and B is set of coordinate points known as structuring element. Then the dilation of A by B can be denoted as eq.1

$$A \oplus B = \bigcup_{b \in B} A_b \tag{1}$$

This means that A is translated by every point of the B. dilation can be considered as a union operation of all the translations of the image A caused by the elements specified in the structuring element B.

The role of the structuring element is very crucial. The presence of a pixel to the right of the origin grows a layer of pixels on the right side of the object. The presence of 1 on the left side of the origin of the structuring element grows a layer of pixels on the left side of the resultant image.

6. Proposed Methodology

The approach developed for the word extraction from printed bilingual documents, includes several steps as shown in fig. 3. First the paper based document is scanned at 300 dpi with a flat-bed scanner to create a digital image of the document. Then we begin with a preprocessing to prepare the scanned document to the segmentation. Next we perform the detection and extraction of text lines with horizontal projection. After that we analyze each line separately to calculate size

of structuring element and perform morphological operation to extract the different words presents in each line.

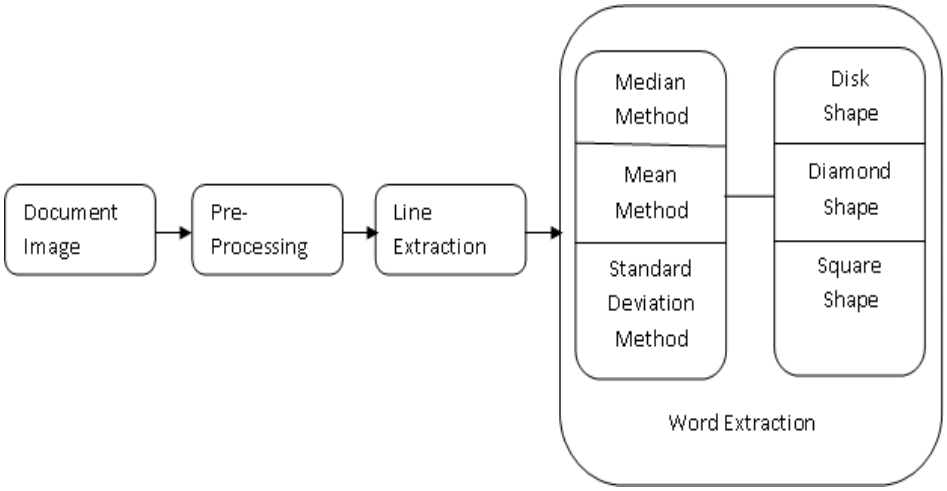

Figure 3. Block Diagram of proposed System

6. 1 Pre-processing

The digitized document image is the result of an acquisition step using a scanner. Our approach doesn't give much importance in the pre-processing because the proposed system should work with images preprocessed in advance by a dedicated pre-processing documents system such as the elimination of graphics objects, the skew correction, etc.

However, our approach preserves the amount of information present in the text because we work with document images containing vowel modifiers, given their important role in the understanding of the text, although their presence may increase the complexity of the segmentation task because of problems encountered in the detection and extraction lines. Indeed, in some writing styles, vowel modifiers may exceed the upper or the lower limit of the line, which can error the step of detecting lines.

In our case, the pre-processing step is limited to the binarization and noise removing of the document image and the values' inversion of black and white pixels in order to prepare the document to the step of detecting lines.

6. 2 Lines Extraction

This phase is quite difficult in the case of bilingual documents because of the large variability between Gujarati and Roman printed scripts, and it becomes more complex with the presence of vowel modifiers in Gujarati script. We have chosen to use the horizontal projection profile method to draw up the boundaries of the horizontal lines. The horizontal projection profile is a histogram of a number of ON pixels along every row of the image. This method corresponds to the needs of document's segmentation because we handle text documents with a simple structure. Our proposed approach goes through the image horizontally and calculates the value of black pixels in each row of the matrix representing the image. Next, we have analyzed the histogram of projections, if the number of black pixels has changed its value from 0 to a positive one then this position is the lower limit of a line. Line segmentation is done at this point. We kept, each time,

the positions of the white areas that will be used for cropping the image into different lines as shown in fig. 4 Segmented text lines further used as input for word extraction.

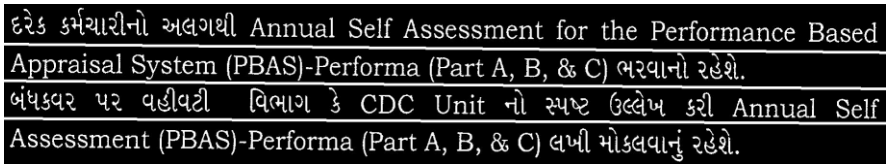

Figure 4. Line Extraction

6. 3 Words extraction

The document segmentation allows to segment documents at different levels, either characters, or words. This level of segmentation is the most difficult among the others, given that the segmentation has to differentiate between different types of spaces between characters, and between words, which is not always obvious to a word extraction system. The objective of the proposed approach is to segment the document image in order to separate and extract the words of a bilingual printed document. Segmentation methods segment documents into connected components; either character or word.

Our approach uses mathematical morphology for the elimination of intra-word spacing and the building of connected components formed by different words in the bilingual printed document. We used the morphological dilation to enlarge the image by filling the holes corresponding in our case to the intra-word spaces. To accomplish this, we must determine the best structural element able to stick the different characters of a word, without sticking words together. At this level, two major problems appear. The first is the size of the structural element and the second is its shape. The determination of these two characteristic features of the structuring element is the foundation of our work.

Choosing only one or fixed size of the structuring element for each document cannot give a performing segmentation because the intra-word spaces differ from one font to another, and depend on the size of the font. Similarly, the spaces between words depend on the text alignment, especially in the case of justified text. Moreover, a document can contain different fonts and sizes. The shape of the structuring element solves the problem of extracting modifiers along with words. A solution of this problem is to stick modifiers to their words.

The approach developed proceeds line by line to find each time, the size and shape of the structuring element of the dilation that can separate and extract correctly and with minimal changes the different words in a document line. Indeed, we proposed three methods to calculate the size of the structuring element and selected three specific shapes for the structuring element. Our approach consists in testing all the combinations of methods for calculating the size of the structuring element and its forms. In fact, we fix each time the calculation method and we vary the shape. We begin by applying a combination of Gujarati and Roman printed documents. If we get good results, we continue testing on mixed documents. Otherwise, we consider it unnecessary to apply the combination to bilingual documents.

6. 4 Structural Element Size Calculation

To calculate the size of the structuring element, we started by determining the list of spaces in each line, then, we developed three different methods to solve this problem. Our approach proceeds by analyzing the extracted lines. This analysis is based on the calculation of the vertical

projection to determine the values of the different spaces in the document. The next step is to analyze each line vertically and to determine the positions of the spaces inter and intra words, if the number of black pixels becomes zero after a sequence of non-zero black pixels then this change corresponds to the presence of a space in the line. We store its position and calculate the length of this area. We obtain at the end a list composed of space values present in the considered line.

Let us consider a Roman printed line as shown in fig. 5 which is both left-justified and right-justified between the fixed left-hand and right-hand margins and in which each pair of mutually adjacent words is separated by some distance. Firstly, we scans the line from the left-margin to the right-margin in the direction of the line height (indicated by Arrow), each time scanning a width and sequentially stores the results in the logical product array as shown in fig. 6 in terms of "O" (or "black bit") indicating that a black section was detected by that scan and "l" (or "white bit") indicating that the scanned area was all white.

It has been observed that for some time

Figure 5. Roamn text line

Figure 6. Logical representation of line

After storing the result of one line in the logical product array, we begins to examine this array sequentially from the left-hand end as shown by Arrow D in fig. 7 and the number of white bits (or "white bit number") in a continuous array sandwiched between black bits "0" on both sides is counted as shown also in fig. 7. A new list is created which contains only consecutive space values within the inspected line.

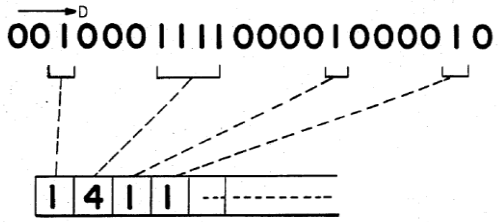

Figure 7. Encountered space list from line

Then this list is used for further processing to identify the size of structuring element for morphological dilation operation.

6. 5 Methods for calculating the size of the structuring element

After the detection of spaces in each line of the document, we determine the size of the structuring element according to the three proposed methods.

6. 5.1 Method based on median calculation

The median is the numerical value separating the higher half of a data sample from the lower half. The *median* of a finite list of numbers can be found by arranging all the data from lowest value to highest value and picking the middle one as eq. 2.

$$Mdn = l + \left(\frac{\frac{N}{2} - \Sigma fo}{fw} \right) i \tag{2}$$

Where,
L is the lower limit of the interval containing the median.
N is the total number of scores.
Σfo is the sum of the frequencies or number of scores up to the interval containing the median.
fw is the frequency or number of scores within the interval containing the median.
i is the size or range of the interval.

This method proceeds by elimination of redundant spaces values presents in the considered line then sort the new list of spaces in ascending order to permit the interpretation of these values. This list reflects the nature of spaces contained in the processed image. It begins with the relatively small areas, which actually represent the spaces between characters in a word, and reaches the largest gap present in the line. This method is based on the fact that the threshold space, able to stick the characters of a word without sticking the words together, has an intermediate value between the lower and upper bound of the new list of spaces. The median value of this list is considered as the size of the structuring element of the dilation.

6. 5.2 Method based on the Mean calculation

Mean is arithmetic average of a range of values or quantities, computed by dividing the total of all values by the number of values.

$$\text{Mean} = 1/N \sum_{i=1}^{N} f_k(i) \tag{3}$$

Where,

fk is vector of space list,

k is vector index which takes values k=1,2,----,N where N is the total no of elements in a vector.

This second method is similar to the previous one in the determination of distinct values of spaces. It is based on the fact that the threshold value of the structural element of the dilation is proportional to the number of spaces present in the image given and their lengths. Indeed, this method sets the size of the structuring element to the average lengths of different spaces in the line introduced.

6. 5.3 Method based on the standard deviation

The standard deviation tells you how spread out the numbers are in your sample data. The standard deviation of a data set is the square root of its variance. It is calculated using eq.4.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{4}$$

This method is based on the detection of spread out the numbers between spaces. It works on the entire list of spaces in line to be processed. From this list, it calculates the mean. Then, it calculates variance which is the average of the squared differences from the mean. Finally calculate the square root of the variance to get the standard deviation value. This calculation method of the structuring element is to generate a list of how much spread occurs between different lengths spaces. The size of the structuring element corresponds to value.

6. 6 Structural element's shape

After calculating the size of the structuring element, next phase is the choice of suitable form which allows segmenting the bilingual printed document correctly. The structuring element can have several forms such as square, diamond, polygon, Euclidean disc, line, point pairs, rectangle, etc. In this approach we are interested to three specific forms of the structuring element, the first is the diamond shape, and the second is the square, the third is the disk. We used these shapes because of the presence of modifiers in the documents to be segmented.

7. Experimental Analysis and Discussion

Due to lack of standard databases, the proposed algorithm has been applied on printed documents from various sources. We have used 3000 Gujarati test samples from printed Gujarati documents, 3000 Roman test samples from printed Roman documents, and 2500 mixed test samples from printed bilingual Gujarati-Roman documents.

For each line, the diameter of the diamond is equal to the size of the structuring element determined by one of the three calculation methods proposed later. The table 1 and fig.8 shows the results found by combining the diamond shape with the three methods for calculating the structural element.

Table 1 Accuracy of Diamond shape

Structuring Element Shape	Method of Size Calculation	Script	Word Extraction Accuracy
Diamond	Median	Gujarati	56.44%
		Roman	83.49%
	Average	Gujarati	97.99%
		Roman	99.44%
		Gujarati	94.40%
	Standard Deviation	Roman	98.42%

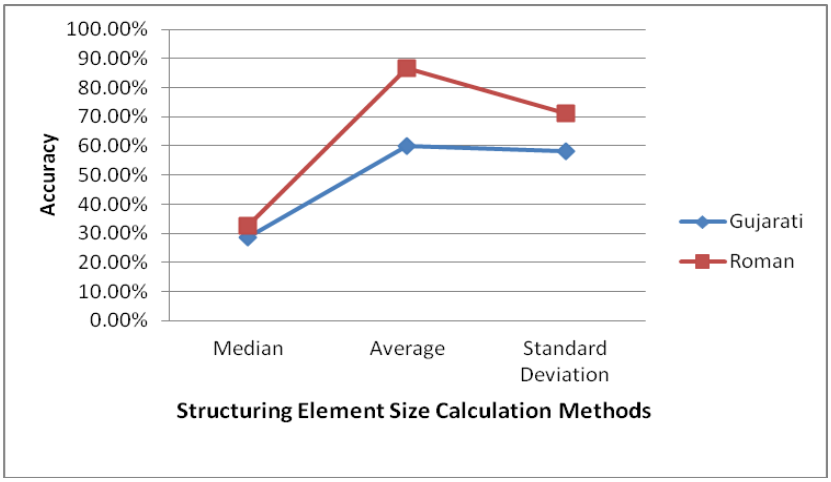

Figure 8. Accuracy of Diamond shape with different methods

For each line, the length of the square is equal to the size of structuring element determined by one of the three calculation methods proposed later. The table 2 and fig. 9 shows the results found by combining the square shape with the three methods of calculating the structural element.

Table 2 Accuracy of Square shape

Structuring Element Shape	Method of Size Calculation	Script	Word Extraction Accuracy
Square	Median	Gujarati	28.72%
		Roman	32.74%
	Average	Gujarati	86.66%
		Roman	59.91%
	Standard Deviation	Gujarati	71.10%
		Roman	58.15%

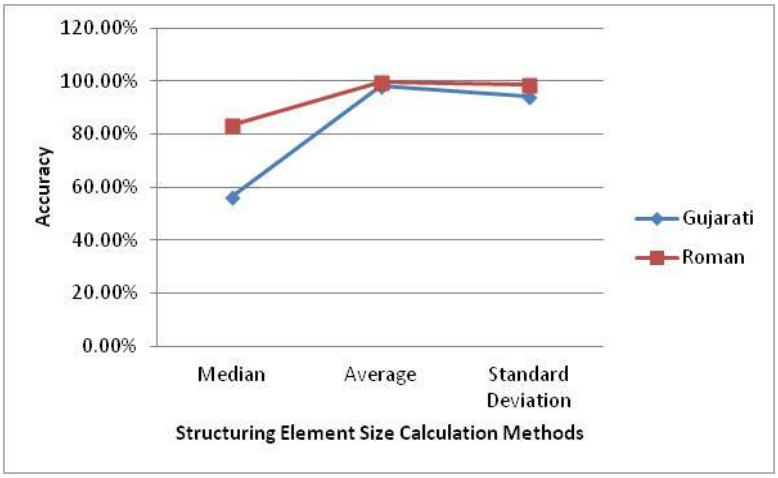

Figure 9. Accuracy of Diamond shape with different methods

For each line, the width of the rectangle is equal to the size of the structuring element determined by one of the three calculation methods proposed later and height is equal to two times this value. The table 3 and fig. 10 shows the results found by the disk shape

Table 3 Accuracy of Disk shape

Structuring Element Shape	Method of Size Calculation	Script	Word Extraction Accuracy
Disk	Median	Gujarati	50.87%
		Roman	75.38%
	Average	Gujarati	95.77%
		Roman	99.09%
	Standard Deviation	Gujarati	92.46%
		Roman	98.87%

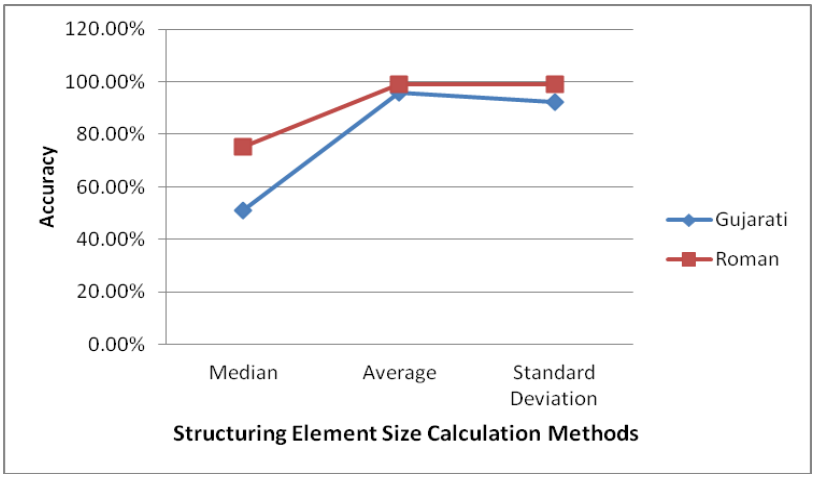

Figure 10. Accuracy of Diamond shape with different methods

The table 4 and fig. 11 represents the best rates achieved for each form of the structuring element.

Table 4 Best Accuracy of different shape

Structuring Element Shape	Method of Size Calculation	Script	Good Extraction Accuracy
Diamond	Average	Gujarati	97.99%
		Roman	99.44%
Square	Average	Gujarati	86.66%
		Roman	59.91%
Disk	Average	Gujarati	95.77%
		Roman	99.09%

We note that the best good extraction rates are obtained for 97.99% and 99.44% Gujarati to Roman. These rates are achieved by the average method of calculating the structuring element's size based on the spaces values.

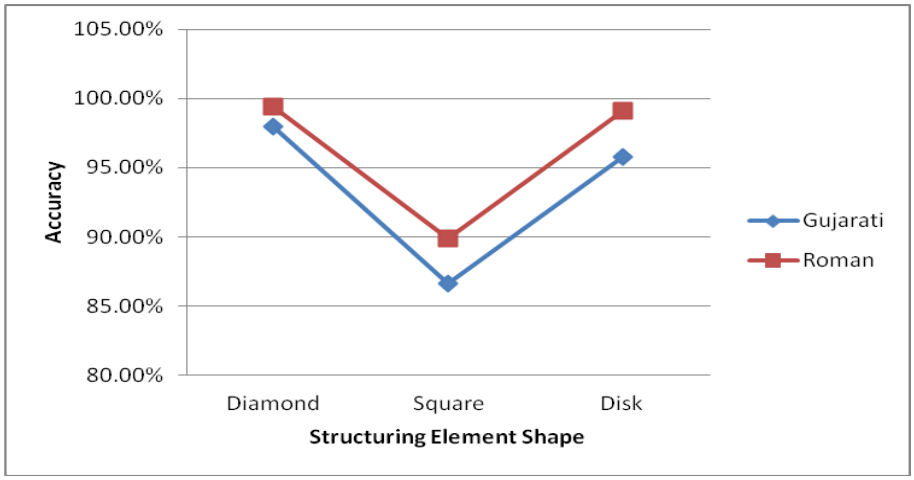

Figure 11. Best Accuracy different shapes

The application of this method on the sample printed bilingual documents gave a good extraction rate equal to 98.85%. This result is explained by the sufficiency of method of calculating the size of the structuring element to changes in the lengths of spaces between the words. The fig. 12 shows a sample run of a line from a printed bilingual document with Gujarati vowel modifiers.

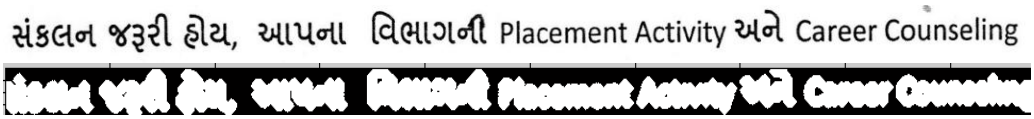

Figure 12. Example of word segmentation from a bilingual printed document

The word segmentation of the printed bilingual document gave 10 words, which correctly corresponds to the words found in the line of the document introduced.

8. Conclusion

A word extraction method composed of gap clustering techniques has been proposed. The gap clustering technique is immune to the variations of scanning resolutions, languages in the document, font styles, and so on. The separation and extraction of words in a printed bilingual document constituted the main contribution of our recognition's area, its different stages, and the various available methods of documents segmentation into words.

We have developed different methods for calculating the size of the structuring element of morphological dilation, combined with different forms and tested on samples of printed Gujarati and Roman documents. After that, we have compared the results, and the best performing was chosen to testing printed bilingual documents in our study. The proposed method has been applied to the segmentation of Gujarati and Roman documents and proven to be effective.

We also plan to extend our method to the processing of textual handwritten bilingual documents, to mixed bilingual documents (both handwritten and printed forms in the same document) as well as treatment of bilingual documents of any kind.

9. References

[1] Y. Y. Tang, M. Cheriet, J. Liu, J. N. Said, and C. Y. Suen, "Document analysis and recognition by computers," Handbook of Pattern Recognition and Computer Vision.

[2] T. Watanabe, ``Document analysis and recognition," IEICE Transactions Inf. and Syst., vol. E82-D, No. 3, March, 1999.

[3] S. Mao and T. Kanungo, "Empirical performance evaluation methodology and its application to page segmentation algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, March, 2001.

[4] G. Nagy and S. Seth, "Hierarchial representatoin of optcially scanned documents," Proceedings of International Conference on Pattern Recognition, vol. 1, pp. 347-349, July, 1984.

[5] H. S. Baird, S. E. Jones, and S. J. Fortune, "Image segmentation by shape-directed covers,"Proceedings of International Conference on Pattern Recognition, pp. 820-825, June, 1990.

[6] L. O. Gorman, "The document spectrum for page layout analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pp. 1162-1173, 1993.

[7] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," Computer Vision and Image Understanding, vol. 70, pp. 370-382, 1998.

[8] F. Wahl, K. Wong, and R. Casey, "Block segmentation and text extraction in mixed text/image documents," Grahical Models and Image Processing, vol. 20, pp. 375-390, 1982.

[9] L. O. Gorman and R. Kasturi, "Document image analysis," IEEE Computer Society Press, 1995.

[10] T. Pavlidis and J. Zhou, "Page segmentation and classification," Graphical Models and Image Processing, vol. 54, pp. 484-496, 1992.

[11] Soo H. Kim, Chang B. Jeong, Hee K. Kwag, and Ching Y. Suen "Word Segmentation of Printed Text Lines Based on Gap Clustering and Special Symbol Detection", ICPR-2002.

[12] Nallapareddy Priyanka, Srikanta Pal and Ranju Mandal, "Line and Word Segmentation Approach for Printed Documents", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition", RTIPPR, 2010.

[13] Apurva A. Desai. , "Gujarati handwritten numeral optical character reorganization through neural network", Pattern Recognition, Vol. 43, no.7 pp. 2582-2589, 2010.

[14] Shailesh Chaudhari and Ravi Gulati, "Script Identification from bilingual Gujarati-English Documents", International Journal of Computer Applications (IJCA), Vol. 93 No. 17, 2014.