Segmentation of Characters from old Typewritten Documents using Radon Transform

Apurva A. Desai
Department of Computer Science
Veer Narmad South Gujarat University,
Surat, Gujarat, India

ABSTRACT

Optical character recognition is a very challenging area. Many works have been done and still being done for many languages across the world. For many Indian languages too good amount of work has been done. However, Gujarati is a language for which hardly any work can be found. Gujarati has a rich literary heritage, and therefore it is important to preserve it for the next generation. In this paper an attempt has be done to segmenting out the words and characters from old typewritten Gujarati documents. Here an algorithm is presented which makes use of global threshold for converting scan RGB documents to blank and white documents. Noise removal has also been applied. Here Radon transform is utilized for skew detection. The novel concept of using Radon transform is presented here in this work. Here Radon transform is used for segmenting documents into lines and then vertical profiles has been used for further segmentation of lines in characters. At last this segmentation algorithm is also tested for the documents typewritten in Hindi. The algorithm presented here gives very good results.

Keywords* : *Segmentation; Radon transform; skew correction; digitization; noise removal

1. INTRODUCTION

Gujarati is a language with very rich literary, cultural, and historical heritage. One can find very old historical documents, literature, books and manuscripts written in Gujarati language. These documents carry lots of knowledge within them. Unfortunately, most of these high valued literatures is not digitized, as a result it will be difficult to preserve this knowledge and also difficult to pass this knowledge on to the next generations to come.

Gujarati is a language belonging to Devnagari family of Indian languages. This is a language unlike many other Devnagari languages without having shirolekha over its characters. It is a language which is having 12 vowels (swars), 34 consonants (vyanjans), and many modifiers (matras). Also Gujarati script is having many composite characters like u,Â etc. All these things make Gujrati Optical Character Recognition very difficult. In addition to complexity of language itself, complexity also increases if the script or document is handwritten or typewritten. Complexity also increases if the physical documents are too old and degraded. If the documents are typewritten it is possible to have different inter-character and inter-word spacing. Also due to long time the background of the document has also got flawed and possibilities of vanishing of the text are also very high. In such scenario the segmentation of characters from old documents becomes very challenging. Figure 1 shows a sample of a Gujarati document.

The document shown in Fig. 1(a) is approximately 35 to 45 years old document. Observe that the background and printed characters. The quality of the documents is really week for optical recognition. Here in this paper the segmentation of old Gujarati documents in to lines and characters has been addressed. In this work some really exciting image processing techniques like Radon transform, skew detection and correction, noise removal, digitization and projections are employed. This paper is divided in five sections. In section two introduction of Radon transform is given. Section three is addressing issue of digitations, section four is discussing issue of skew detection and correction.

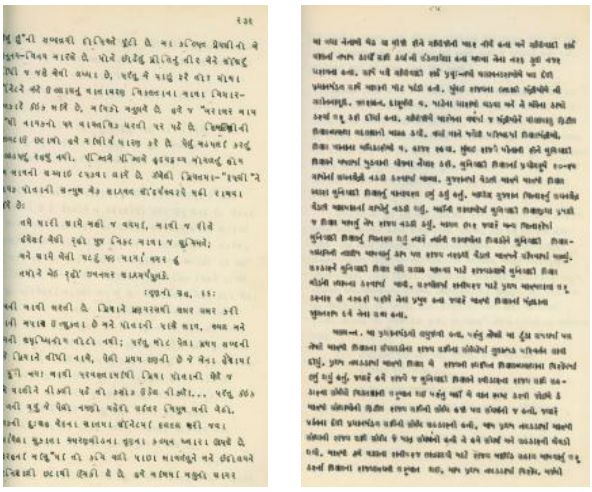


Figure 1 : Scanned old Gujarati documents

2. RADON TRANSFORM

Here, Radon transform is utilized for skew correction and also for line segmentation. In recent time Radon transform has attracted attention of researchers working in the area of image processing and tomography. Radon transform computes line integral from multiple line sources along a parallel path in a direction. That means, it computes projections of an image from a specific direction, taken from different angles by rotating the source of projection around centre of image. The projection of a two dimensional image $f(x, y)$ can be calculated for any angle θ using;

$$RR_{\theta}\ (x') = \int_{-\infty}^{\infty} f(x^{'}cos\theta - y^{'}sin\theta, x^{'}sin\theta + y'cos\theta)dy$$

Where, $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

Fig 2 shows Radon geometry from two different angles, horizontal and vertical, of a Gujarati digit.

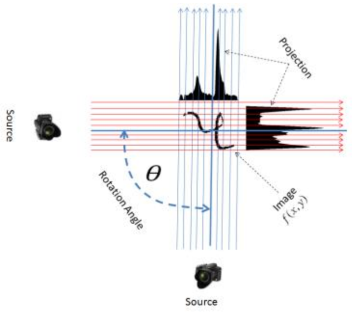


Figure 2: Radon geometry for Gujarati digit five (5)

The strongest property of Radon transform is its ability to extract lines from the image, even from a noisy one. Radon transform represents such lines in form of peaks. The Fig. 3 shown Radon transform taken by 0 to 179 degree of a Gujarati digit five (5) of which Radon geometry is shown in Fig.2.

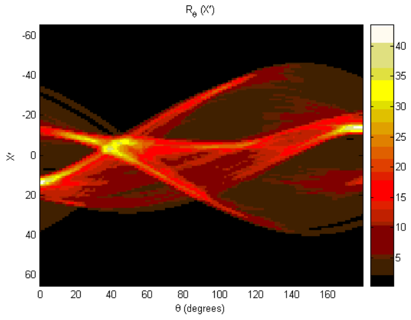


Figure 3: Radon transform of Gujarati digit five (5)

3. RELATED WORK

The work presented here in this paper is mainly addresses three issues of optical character recognition, digitization of the old degraded documents, skew detection and correction and then the segmentation of the text into sentences and then up to single characters or composite characters. Here in this section some of the attempts have been mentioned which address these issues.

In 2005, Ntzios et al. [1] presented a work dealt with old Greek handwritten OCR system. In this paper they made use of adaptive binarization technique for binarizing old documents. In the pre-processing stage they did skeletonization of characters in the document. Here authors presented OCR system without performing segmentation on the documents. Water reservoir technique was an instrument for feature extraction in this paper. The position and the place of cavities of characters constitute feature set for the classification. This work has claimed accuracy level up to 98.94%. In 2006, Kavallieraton et al. [2] worked on noise removal from old historical degraded documents. In their work they presented an algorithm which was having three stages. In the first stage iterative global threshold (IGT) was applied. In the second stage, using a fact that even after performing IGT still degraded portion of the document has more black pixels, they isolated yet noisy portion. In the last and third stage local threshold was performed on the portion isolated in the second stage. In this work they obtained precision up to 100%. Again in 2006, Kavallieraton et al. [3] has proposed a novel technique for binarization especially for degraded documents. Here in this work a hybrid approach based on both iterative global threshold (IGT) and local threshold have been used. Authors have claimed 97.6% success rate using this approach. Gatos et al. [4] also in 2006, presented a very good paper explaining adaptive binarization of degraded documents. In that algorithm authors had used low-pass Wiener filter as a pre-process and then interpolating neighboring background intensities were used for background rough calculation. In the next step of algorithm threshold value was calculated. Here threshold was based on background calculation done in the earlier step and the background of the original image. The last step of the algorithm was improving the quality of the text, for doing this stroke connectivity was taken into consideration. Agam et al. [5] presented a work to enhance quality of degraded documents. Here in the algorithm foreground separation was performed first, and then linear blending was used to improve the quality of image. The algorithm presented in this paper was then compared with other algorithms and proved better than them.

Skew detection and correction is a problem which has attracted attention of many researchers. Manjunath et al. [6] presented a work to estimate skew from binary documents. The algorithm presented here was based on skeletonization, moments and boundary growing (BG) approaches. Also in this work a novel four stepped approach has been presented for thinning. The presented algorithm was compared with many other approaches and proved better than many of them. In 2001, Das et al. [7] also worked on correcting skew in the documents. In their work they used mathematical morphology for estimating skew in the documents. In this work mathematical morphological methods, in specific closing and opening, were used to constitute lines for each of them skew was measured and the median angle of all those skew angles was considered for skew correction. In 2007, Banshree et al. [8] presented an OCR system for printed Kannada documents. Here, Hough transform technique was utilized for skew detection and then segmentation for lines and words, vertical and horizontal projection were utilized. Murtoza Habib el al. [9] used Radon transform for detecting skew in the documents in Bangla scripts. Before using Radon transform to find skew angle, the authors have done pre processing for filtering out isolated dots, punctuations etc. and also found upper envelops. These upper envelopes were then utilized in Radon transform. This algorithm was tested for angle up to ±45 degree.

Like binarization and skew corrections there are many researchers presented their work on segmentation, and therefore many methods can be traced on it. Looking to the available literature it is found that the horizontal and vertical profiles are the most widely used methods for segmentation. However, here few attempts are presented for segmentation on documents in Indian languages. Bhagvati et al [10] in 2002 used projection profiles for segmenting lines, words, and characters in Telugu and Gurumukhi scripts. They also noted that Run-Length smoothing algorithm is an effective way for segmenting words from text. In 2008 Sagar et al. [11] presented their work on Kannada text. In this work they did segmentation of Kannada script using the pixel information of the document. In 2007 Jindal et al. [12] presented segmentation of horizontally overlapping lines. This work addressed eight Indian languages. Here they used inter line gaps and projection for overlapping line segmentation. They considered Gurumukhi, Devnagari, Bangla, Gujarati, Kannada, Tamil, Telugu, and Malayalam printed text.

Seethalakshmi et al also made use of vertical and horizontal profiles for segmentation in printed Tamil text.

4. DIGITIZATION

For success of any optical character recognition digitations is an initial and the most important phase. Here for segmentation of Gujarati characters from over 35 to 45 years old documents, they are scanned using standard flat-bed scanners in 300 dpi resolutions in RGB format. A couple of sample scanned images is shown in Fig. 1. For segmenting theses documents into lines and characters they are required to be converted in black and white. There are many methods used for converting RGB or grayscale image to black and white image. Global threshold and adaptive threshold are two major groups of methods. Global threshold method is a method in which one optimum value is calculated and then entire image converted to black and white, where as in adaptive threshold method various threshold values are calculated for different part of an image and then RGB or grayscale image is converted into black and white. Looking to the nature of the scanned document global threshold is adopted to convert RGB images to black and white images. Since the documents are old and typewritten after conversion into binary images scattered noise remains there in the images. This noise creates problem at the time of segmentation. To remove this noise median filter of 3 x 3 size has been employed.

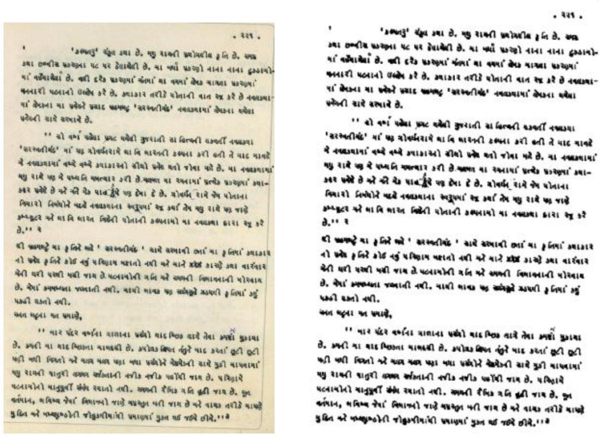


Figure 4 The RGB document and its binarized document

Once RGB document has been converted into BW image, it has been observed in many cases that due to the conversion process some the parts of the text got isolated and some of the part of the text or word got so thin. To take care of such issues the image is dilated using morphological dilation. Let us assume that A, be the image and B be a smaller 2x2 set of pixels. Then the dilation would be;

$$A \bigoplus B = \cup_{x \in B} A_x \tag{1}$$

5. SKEW DETECTION AND CORRECTION

In the digitized form of very old type-written documents skew occurs because of two things, one, the a document was type written with some skew and second, the document was scanned with some skew within it. In both of these two situations the skew generated remains uniform across the page. With this assumption Radon transform was used for finding skew angle. Here Radon transform was calculated on a BW image with 0 to 179 degrees of rotation angle around the centre of the image. This transform would give the angle of a vertical line and as we were looking for a skew in the horizontal line the angle of vertical line would be subtracted from 90 to get skew in the horizontal line. The sign of subtracted angle would give the direction of skew as well. After calculating the skew in the document the document was rotated to remove skew from the document. Fig. 6b is a Radon transform of a document shown in Fig. 5a. here light colours shows the place of straight lines which is at 89^0. Thus the rotation angle is of 1^0. Fig.6 shows the skewed document and skew-less document.

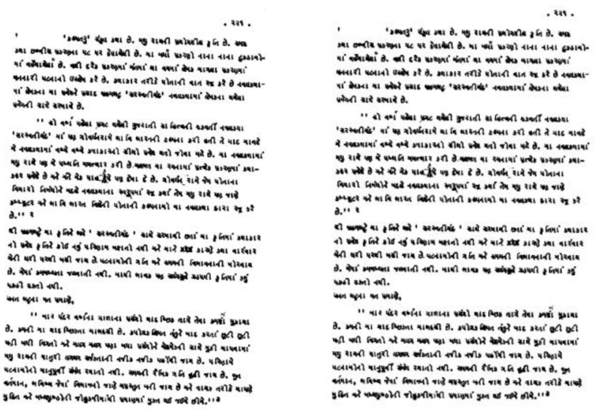


Figure 5 Skew detection and correction

After skew correction the document image is ready for the segmentation.

6. SEGMENTATION

The prime objective of this work was to perform segmentation on the Gujarati documents. By segmentation we wanted to separated out each individual characters and composite characters. If all individual characters and composite characters got separated they can then be classified or identified. Here for us there were two tasks to perform. One, separating lines form the documents and second is separating words and characters. If typewritten documents are observed carefully one might observe a fact in Gujarati typewritten documents that there is considerable space between two characters in a word. This fact was utilized and that helped in segmenting characters directly without segmenting the words. In any language two lines are separated by some bank space between them. In other words, there is a line of background pixels between two lines. In the introduction section we have noted that Radon transform has got strong characteristic to identify and locate a straight line from a figure. Exploiting this feature of a Radon transform we separated lines from the document. Fig.6b shows Radon transform on a Gujarati Document shown in Fig 6a with 0^0 to 179^0 degrees of projection angle. Also observe that there are many light coloured spots, which are the peaks of Radon transform, there in Radon transform at 90^0 and these spots are same as total number of lines in the document.

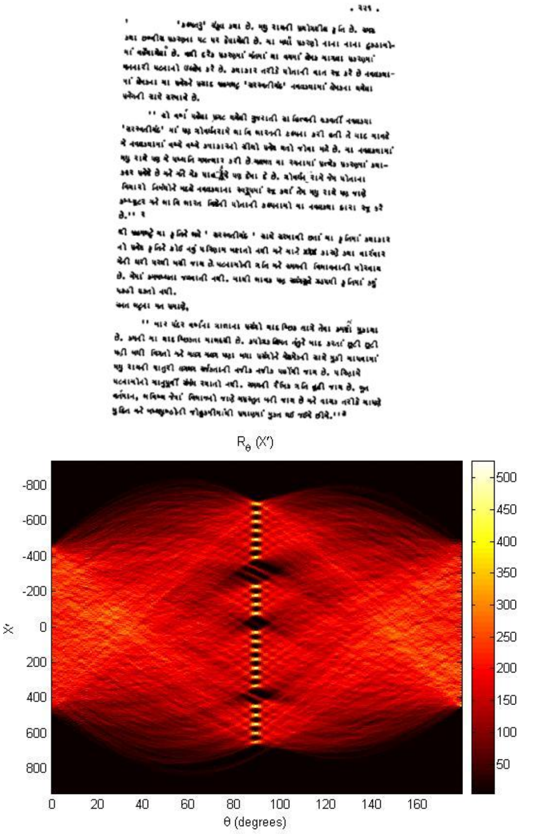


Figure 6 : Gujarati document and its Radon transform

Now to identify the blank lines, that are line separators, an angle was found with the maximum valleys. Here valleys were nothing but the place where no foreground pixels were there, in other words they were the blank spaces. Observe Fig. 7, where projection obtained by Radon transform at the identified angel, in this case it is 90^0, is shown, which shows the peaks with the written texts and valleys along with the blank spaces between the lines.

After identifying the place of lines, those lines are cropped out, and as a result all individual lines were segmented out. In Fig. 8 one of the segmented lines is illustrated.

The next step was to segmenting characters from these separated lines. For doing this vertical projection was used on each of the separated lines. The vertical projections too will give the places of foreground pixels and background pixels. The place where we get peaks that indicates the presence of characters and the valleys indicate the blank spaces. In a line, spaces between characters are the separators between the characters; therefore the valleys are in fact the separators. We identified the places of valleys and then the characters are cropped out. Fig.8 shows a line and its vertical projection. Fig 9 demonstrate the complete algorithm for segmentation of old Gujarati character segmentation.

The algorithm presented here is tested for Hindi documents also. Hindi is also a Devnagari language unlike Gujarati language having shirelekha (headlines) above all the characters. Because of these shirolekhas this algorithm does not segmented out the Hindi characters from the documents, however, this segmented out Hindi documents up to words. For further segmentation from words to Hindi character separate segmentation process is needed. Refer Fig. 10 which shows the result of this algorithm on Hindi document. It is observed that the algorithm works very effectively for separating Gujarati characters, right and left modifiers, and composite characters. However, when the inter-character space is not sufficient, the algorithm identifies two characters as a single character. On an average in a line, which contains more than 30 characters, 2 characters are wrongly separated. In Hindi documents most of the time all the words are separated correctly.

Gujarati and Hindi documents considered here, for testing this algorithm, are having on average 24 lines. On an average there are twelve words and total thirty characters or composite characters with modifiers. It is observed that on an average fifteen words and eleven words are wrongly segments in Gujarati and Hindi documents respectively. Similarly on an average two characters or composite characters are wrongly segmented per line in Gujarati document. The proposed algorithm segments lines and words of documents very efficiently. Whereas at the character level the algorithm works well with the Gujarati language, but it does not give satisfactory results at all for the documents in Hindi language. Table 1 shows the summarized results.

Table 1 : Summarized results of proposed algorithm with Gujarati and Hindi Languages

Lang.	Segmentation Level					
	Line	Effi. (%)	Words	Effi. (%)	Char.	Effi (%)
Gui.	YES	100	YES	94.79	YES	65.75
Hindi	YES	100	YES	96.18	NO	-

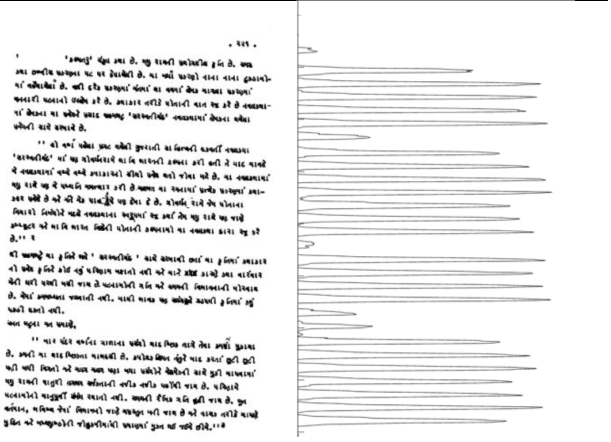
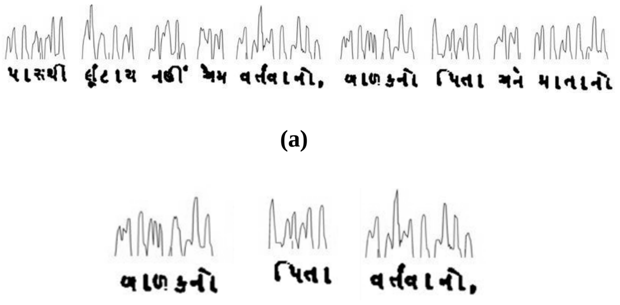
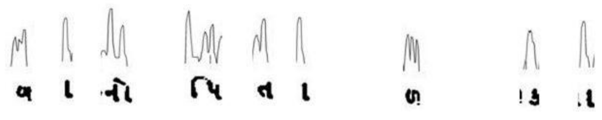


Figure 7 : Document and Radon transform



(a)

(b)



(c) (d) (e)

Figure 8: (a) line segmentation (b) words segmentation (c) characters segmentation (d) under segmented character (e) over segmented characters with their vertical projection

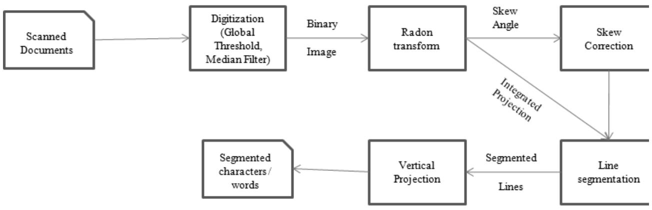
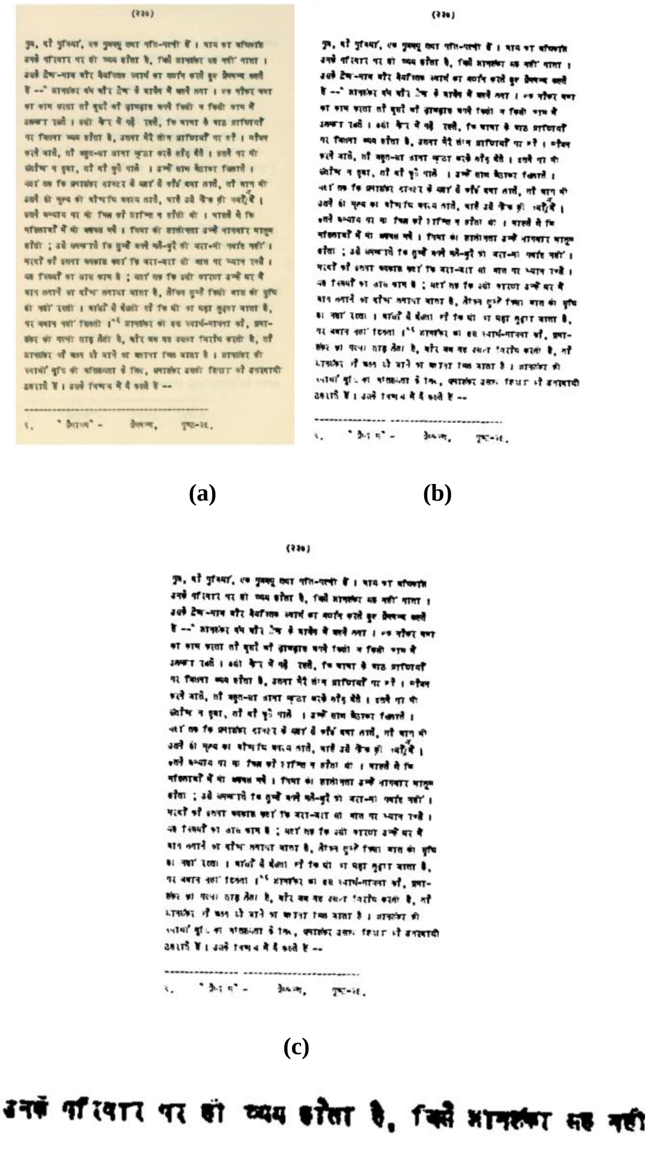


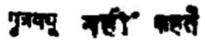
Figure 9 : Block diagram of proposed algorithm



(a) (b)

(c)

(d)



(e)

Figure : 10 (a) A scanned Hindi document (b) binary document without noise (c) document without skew (d) a line segmented from the document (e) some of the words segmented from the document

7. CONCLUSION

Optical character recognition for old and historical document is always a challenging work. It needed very efficient digitization process and then segmentation process for classification of characters. For old and typewritten Gujarati documents, the algorithm presented here works satisfactory. However, because of irregularities of type writers causes unequal inter-character and inter-words spacing the words and composite characters needs more processing to separate. At the same time, when this algorithm is applied to a language like Hindi, it works very fine to segment out documents up to word level. As a whole the algorithm presented gives very satisfactory result, which may be used for more than one language as well.

8. REFERENCES

[1] E. Kavallieratou, E. Stamatatos, Improving the quality of degraded document images, proceedings of the second international conference on document image analysis for libraries (Dial '06), 2006, 330-349

[2] K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris, S.J. Perantonis, an old greek handwritten ocr system, proceedings of the 2005 eight international conference on document analysis and recognition, (ICDAR 05), 2005, 64-69

[3] E. Kavallieratou, S. Stathis, Adaptive binarization of historical document images, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), 2006, 742-745

[4] B. Gatos, I. Pratikakis, S.J. Perantotis, Adaptive Degraded Document Image Binarization, Pattern Recognition, 39, 2006, 317-327

[5] G. Agam, G. Bal, G. Frieder, O. Frieder, Degraded document image enhancement, http://ir.iit.edu/publications/downloads/doc_enhancement.pdf

[6] Manjunath Aradhya V N, Hemlatha Kumar G, Shuvkumara P., skew estimation technique for binary document images based in thinning and moments, Engineering Letters, 14, 1 (Advance Online Publication)

[7] N.P. Banshree, R. Vasanta, OCR for script identification of hindi (devnagari) numerals using feature sub selection by means of end-point with neuro-memetic model, Proceedings of World Academy of Science, Engineering and Technology, 22,(2007,78-82

[8] S.M. Murtoza Habib, N. Ahmed Noor, M. Khan, Skew angle detection of bangla script using radon transform, http://www.bracuniversity.net/research/crbpl/papers/paper_ICCT06_skew_Murtoza.pdf

[9] A.K. Das, B. Chanda, A fast algorithm for skew detection of document images using morphology, International Journal of Document Analysis and Recognition, 4, 2001, 109-114

[10] C. Bhagvati, T. Ravi, SM. Kumar, A. Negi, On developing high accuracy ocr system for telugu and other indian scripts, Proceedings of the Language Engineering Conference (LEC’02), 18-23

[11] B.M. Sagar, Shobha G., R. Kumar, OCR for kannada text to machine editable format using database approach, WSEAS Transactions on Computers, 6(7).2008, 766-769

[12] M.K. Jindal, R.K. Sharma and G.S. Lehal, Segmentation of horizontally overlapping lines in printed indian scripts, International Journal of Computational Intelligence Research, 3 (4), 2007 277 – 286

[13] Sheetalkumari, Shreeranjani, Balachandar, A Singh, M. singh, R Ratan, S. Kumar, Optical character recognition for printed tamil text using unicode, Journal of Zhejang University SCI, 64(11), (2005), 1297-1305

9. AUTHORS PROFILE

Prof. Apurva Desai is a professor in Computer Science. His area of interest is Optical Character Recognition (OCR), Computer Graphics, Data Mining, and Artificial Intelligence. He has published over forty national and international papers and authored four books. Prof. Desai has delivered many talks in national and international events and visited many countries for academic purpose. So far six students have acquired Ph.D. degree under his guidance.