International Journal of Science Research
Volume 01, Issue 01, June 2012, pp. 18-23

A Survey of Methods and Strategies in Handwritten Kannada Character Segmentation

M. Thungamani[1*] and P. Ramakhanth Kumar[2]

[1]*Tumkur University, Tumkur, Karnataka, India*

[2]*Department of Information Science and Engineering, R.V. College of Engineering, Bangalore, Karnataka, India*

**Author for correspondence: thungamani_k@rediffmail.com*

Received on: 15th December 2011, Revised version received on: 10th February 2012, Accepted on: 30th February 2012

Abstract: A survey on different Segmentation methods listed here for Handwritten Kannada characters; Segmentation is an important task of any optical character recognition (OCR). Character segmentation is considered one of the main steps in pre-processing. Therefore, there are many techniques developed for character segmentation. This paper provides a review of these advances. The aim is to provide segmentation techniques that have been developed; the "classical" approach consists of methods that partition the input image into sub-images, which are then classified. The operation of attempting to decompose the image into classifiable units is called "dissection." The second class of methods avoids dissection and used Holistic approaches that avoid segmentation by recognizing entire character strings as units are described. Next strategy is that line segmentation is carried out by calculating the horizontal projection profile of the whole document, and exhibits valleys of zero height corresponding to white space between the text lines. Vertical projection profile is used to do the word and character level segmentation. Accuracy depends upon the segmentation algorithm. The connected components for segmentation to calculate the ratio between height and width, and the ratio between black and white pixels inside the bounding box to cluster the connected components.

Key Words: Segmentation technique, Horizontal projection, Vertical projection, Connected components

1. INTRODUCTION

Kannada is a language spoken in south India predominantly in the state of Karnataka, Kannada whose native speakers are called Kannadigas roughly 50 million, Kannada script has large number of characters with similar looking shapes among characters, and characters belonging to same class have higher variability across different set of fonts.

Kannada script is written horizontally from left to right and an absent of lower and upper case as like in English language [1], Moreover the Kannada characters are formed by combination of basic symbols, recognition of the Kannada character is complex and challenging task & increased character set, it contains Vowels, Consonants & Compound characters. Some of the character may overlap together. Kannada text is difficult when compared with Latin based languages because of its structured complexity. Moreover, Kannada language uses 49 phonemic letters, shown in Figure 1 it is divided into 3-groups,Vowels (Swaragalu- Anusvara (o), & Visarga (:)15), Consonants (Vyanjanagalu-34) and modifier glyphs (Half-letter) [2] from the 15 vowels are used, to alter the 34 base consonants, creating a total of (34*15) +34=544 characters, additionally a consonants emphasis glyph called Consonant conjuncts in Kannada Figure 2 [vattakshara], exists for each of the 34 consonants. This gives total of (544*34) +15=18511 distinct characters.

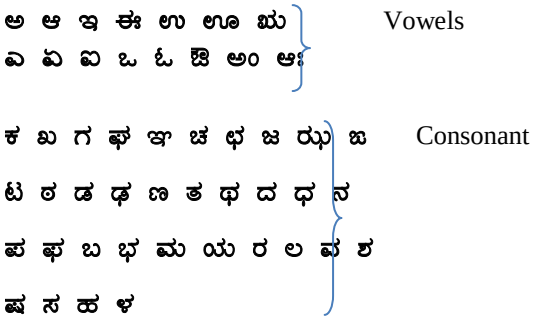


Figure 1. Consonant conjuncts in Kannada

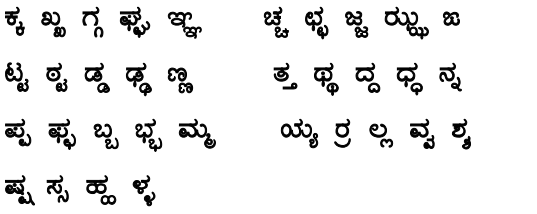
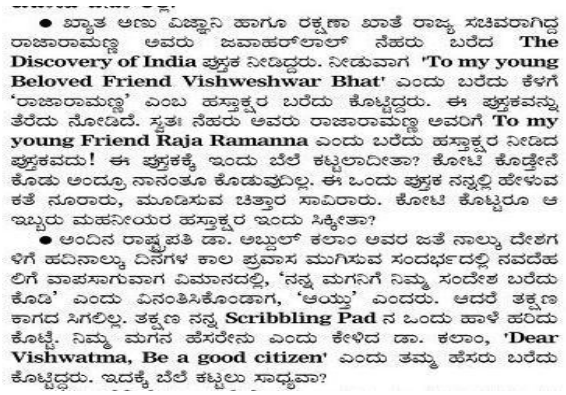


Figure 2. Consonant conjuncts in Kannada (vattakshara)

The peculiar nature in which one or more consonants combine with vowels to produce a compound character in Kannada language shown in the Figure 3 results in a huge number of character combinations, Sample Kannada script that has half letters/subscript (vattakshara) and shirorekha.



● ಖ್ಯಾತ ಅಣು ವಿಜ್ಞಾನಿ ಹಾಗೂ ರಕ್ಷಣಾ ಖಾತೆ ರಾಜ್ಯ ಸಚಿವರಾಗಿದ್ದ ರಾಜಾರಾಮಣ್ಣ ಅವರು ಜವಾಹರ್‌ಲಾಲ್ ನೆಹರು ಬರೆದ **The Discovery of India** ಪುಸ್ತಕ ನೀಡಿದ್ದರು. ನೀಡುವಾಗ **'To my young Beloved Friend Vishweshwar Bhat'** ಎಂದು ಬರೆದು ಕೆಳಗೆ 'ರಾಜಾರಾಮಣ್ಣ' ಎಂಬ ಹಸ್ತಾಕ್ಷರ ಬರೆದು ಕೊಟ್ಟಿದ್ದರು. ಈ ಪುಸ್ತಕವನ್ನು ತೆರೆದು ನೋಡಿದೆ. ಸ್ವತಃ ನೆಹರು ಅವರು ರಾಜಾರಾಮಣ್ಣ ಅವರಿಗೆ **To my young Friend Raja Ramanna** ಎಂದು ಬರೆದು ಹಸ್ತಾಕ್ಷರ ನೀಡಿದ ಪುಸ್ತಕವದು! ಈ ಪುಸ್ತಕಕ್ಕೆ ಇಂದು ಬೆಲೆ ಕಟ್ಟಲಾದೀತಾ? ಕೋಟಿ ಕೊಡ್ತೇನೆ ಕೊಡು ಅಂದ್ರೂ ನಾನಂತೂ ಕೊಡುವುದಿಲ್ಲ. ಈ ಒಂದು ಪುಸ್ತಕ ನನ್ನಲ್ಲಿ ಹೇಳುವ ಕತೆ ನೂರಾರು, ಮೂಡಿಸುವ ಚಿತ್ತಾರ ಸಾವಿರಾರು. ಕೋಟಿ ಕೊಟ್ಟರೂ ಆ ಇಬ್ಬರು ಮಹನೀಯರ ಹಸ್ತಾಕ್ಷರ ಇಂದು ಸಿಕ್ಕೀತಾ?

● ಅಂದಿನ ರಾಷ್ಟ್ರಪತಿ ಡಾ. ಅಬ್ದುಲ್ ಕಲಾಂ ಅವರ ಜತೆ ನಾಲ್ಕು ದೇಶಗಳಿಗೆ ಹದಿನಾಲ್ಕು ದಿನಗಳ ಕಾಲ ಪ್ರವಾಸ ಮುಗಿಸುವ ಸಂದರ್ಭದಲ್ಲಿ ನವದೆಹಲಿಗೆ ವಾಪಸಾಗುವಾಗ ವಿಮಾನದಲ್ಲಿ, 'ನನ್ನ ಮಗನಿಗೆ ನಿಮ್ಮ ಸಂದೇಶ ಬರೆದು ಕೊಡಿ' ಎಂದು ವಿನಂತಿಸಿಕೊಂಡಾಗ, 'ಆಯ್ತು' ಎಂದರು. ಆದರೆ ತಕ್ಷಣ ಕಾಗದ ಸಿಗಲಿಲ್ಲ. ತಕ್ಷಣ ನನ್ನ **Scribbling Pad** ನ ಒಂದು ಹಾಳೆ ಹರಿದು ಕೊಟ್ಟೆ. ನಿಮ್ಮ ಮಗನ ಹೆಸರೇನು ಎಂದು ಕೇಳಿದ ಡಾ. ಕಲಾಂ, **'Dear Vishwatma, Be a good citizen'** ಎಂದು ತಮ್ಮ ಹೆಸರು ಬರೆದು ಕೊಟ್ಟಿದ್ದರು. ಇದಕ್ಕೆ ಬೆಲೆ ಕಟ್ಟಲು ಸಾಧ್ಯವಾ?

Figure 3. Sample Kannada script [scanned from prajavani, dated 29th Dec 2011].

The recognition of hand writing by machines has been research topic for over 40 years. Character segmentation has long been a critical area of the OCR process. Optical character recognition (OCR) is a program that translates scanned document into a text document; once it is translated into text it can be stored in ASCII format. Most OCR systems use a combination of hardware and software to recognize character, in computer software any character symbol that requires One-byte (8-bit) of storage. This includes all the ASCII & Extended ASCII character set uses the number 0 through 127 to represent all English characters, letters, and numbers punctuation including space character. The process of Character Recognition shown in Figure.4 can be classified as Pre-processing, Segmentation, Extraction and Recognition are important task of any OCR system, and it separates the image text document into lines, words & character. The accuracy of OCR-system mainly depends on the segmentation algorithm to get high recognition rate.

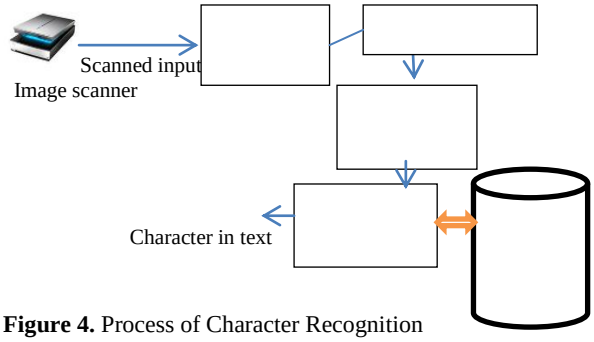


Figure 4. Process of Character Recognition

As Shown in Figure 4.The hand written document is chosen for scanning. It is placed over the scanner [3], Scanner software is invoked which scans the document. The document is sent to a program that saves it in preferably TIF, JPG of GIF format, so that the image of the document can be obtained when needed. This is the first step in OCR [4] the size of the input image is as specified by the user and can be of any length but is inherently restricted by the scope of the vision and by the scanner software length.

OCR Methods

The OCR method consists of a number of steps are listed below

- Binarization,
- Noise removal,
- Thinning,
- Skew Detection & Correction,
- Segmentation,
- Feature Extraction and selection,
- Classification

System first scan the image of Kannada script, using scanner then to pre-processing step, [1] binarization, is the process of converting a gray scale image (0.255 pixel values), into binary (0 & 1 pixel values) by thresholding. Binary document requires less space to store, this techniques remove the majority of the noise. Noise introduced during scanning or due to poor quality of page may also contain blur image has to be cleared before further processing.Thinning-reduce the size of the character by using Skeletoning the image by different technique example Otsu [22] method. Skew correction – Angle detection cumulative scalar products of windows of text blocks with the Gabor filters at different orientation are calculated. Maximum Cumulative scalar products give the skew angle. Next process is Handwriting recognition is the ability of a Computer to receive & interpret input from sources such as paper documents, photographs, touch-screens and other devices.

2. BACKGROUND STUDY

In the late 1950s and early 1960s, in order to get good segmentation results, strict constraints were placed on the input documents. The writer or printer was constrained to place characters into printed boxes [5] to ease segmentation. Fonts of printed materials were chosen with strong leading edges for easier detection. Due to the constrained inputs and the research focus on classification, segmentation was not well developed before the 1970s. A good part of recent progress in reading unconstrained printed and Hand written Kannada text may be described to more insightful handling of segmentation. Segmentation methods are listed under different headings. The term “classical” in 1996 approach consists of methods that partition the input image into sub images, which are then classified. The profile based methods can only segment non-overlapping lines and characters. The operation of attempting to decompose the image into classifiable units is called “dissection”. The second class of methods avoids dissection, and segments the image either explicitly, by classification of pre-specified windows, or implicitly by classification of subsets of spatial features collected from the image as a whole. The

third strategy is a hybrid of the first two, employing dissection together with recombination rules to define potential segments, but using classification to select from the range of admissible segmentation possibilities offered by these sub images. Finally, holistic approaches that avoid segmentation by recognizing entire character strings as units are described. The aim is to provide an appreciation for the range of techniques that have been developed.

2.1 Dissection Approach

A single partitioning of the image into sub images based on character-like properties followed by classification of the sub images. Dissection [6] is the decomposition of the input image into a sequence of sub images. The criterion for good segmentation using the dissection approach is the agreement of character properties in the segmented sub image and the expected symbol. The dissection method makes use of the character properties like height, width, space, separation from neighboring components, disposition along the baseline, etc. This method is suitable for printed image documents in which each character image is well spaced.

2.2 Holistic Approach

Segment and recognize words as single units. The task of segmenting characters varies in difficulty based on the input type. Fixed pitch machine printed text can typically be segmented fairly easily using simple projection analysis. However, written text must be segmented using more advanced methods such as [7] Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs), and contextual methods. In order to achieve high accuracy on complex problem domains, segmentation and recognition cannot be treated independently. However, a simple problem domain can use more basic techniques and still achieve high accuracy.

3. SEGMENTATION TECHNIQUES

After scanning the document, the document image is subjected to pre-processing. For background noise elimination [8], skew correction and binarization to generate the bit map image of the text. The pre-processed image is then segmented into lines, words and characters, discussed in the following sections.

3.1 Line segmentation

To separate the text lines, from the document image, the horizontal projection profile of the document image is found. The horizontal projection profile [9] is the histogram of the number of ON pixels along every row of the image. White space between the text lines is used to segment the text lines. The line segmentation block diagram Figure.5 shown below. The architecture [10] can be divided into four steps, temporary input image, histogram calculation, draw the histogram of the number of pixels in each row, find a row with no active pixel, false minimum elimination and temporary output memory. System first load data into temporary input memory. When the first 32 bit input is detected it will enable the histogram calculation process. If the histogram calculation detects drastic changes in the histogram, it will enable the false minimum elimination entry. Elimination process is done until it reaches the point of separation or segmentation point.

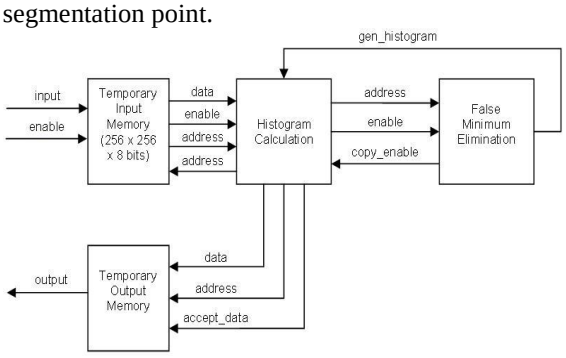


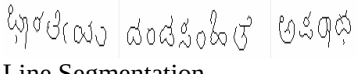
Figure 5. Line segmentation Block Diagram

3.2 Word segmentation

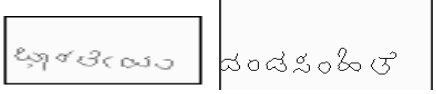
The spacing between the words is used for word segmentation [11]. For Kannada document spacing between the words is found by taking vertical projection profile of an input text line. Vertical projection profile is the sum of ON pixels along every column of an image. The projection profile is the histogram of the image. In the profile, the zero valley peaks may represent the character or word space. It differentiates whether it is character or word spacing, find the maximum character space cluster and use it for separating the words.

3.3 Character segmentation

Character segmentation is the decomposition of an image into sub images in Figure.6, which only contain a single character. It is a critical step in most [13,14] OCR systems, and typically the cause of a high proportion of OCR errors. The overlapped text uses projection profile algorithm, connected components and also uses nearest neighborhood method to cluster the connected components.



Line Segmentation



Word Segmentation

Character segmentation

Figure 6. Segementation stages

3.4 Projection analysis

In the projection profile method, the horizontal and vertical profiles are computed. Projection profile is the histogram of the image i.e the number of ON (black) pixels along every row of the image. The projection profile exhibits valleys of zero height corresponding to white space between the text lines. Line segmentation is done at these points. When the projection profiles are plotted we can see peaks and valleys in the plot. The Pre-processing (noise removal, skewing, digitization) Segmentation (lines, words and characters) zero valued valleys are identified to separate the lines, words and characters.

3.4.1 The Horizontal projection

To separate the text line the horizontal projection profile [10][12] of the document image is found, the horizontal projection profile is the histogram of the number of ON pixels along every row of the image. White space between text lines is used to segment the text lines. Figure.7 shows a sample Kannada document along with its horizontal projection along with the histogram for line 1 and line 2.

Horizontal projection can't deal well with skewed, curved & fluctuating lines, Hough transform [18] considers any image to compose of straight lines, and it creates an angles offset plane in which the local maxima are an assumed to correlate with text lines.

Figure 7. Kannada text lines with Horizontal projection profile and line segmentation with histogram

3.4.2 Algorithm for segmentation

Step 1: The binarized image is checked for white space between the text lines.
Step 2: If white space between the texts lines are detected then the image is segmented into sets of paragraphs across the white space between the text lines.
Step 3: The lines in the paragraphs are scanned for horizontal space intersection with respect to the Background. Histogram is used to detect the image.

3.4.3 A vertical projection

Vertical projection shown in figure.8 is simply a running count of black pixels in each column. If the count falls below the predefined threshold, the column is a candidate for splitting the image. Peaks of the derivative of this data indicate a potential split. Projection analysis, a one-dimensional analysis, works well on good quality machine printed / hand written documents.

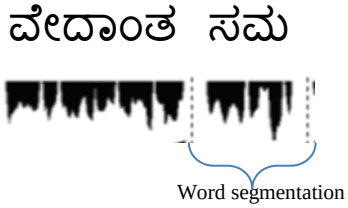


Figure 8. Vertical Projection profile for word segmentation.

3.4.4 Segmentation strategy

Kannada text is composed by attaching to the glyph of a consonant the glyphs of the vowel modifiers and the glyphs of the consonant conjuncts. Due to the large number of letter combinations possible, building a classifier to recognize OCR system for printed Kannada documents a whole letter is very difficult. Here segmentation strategy is based on the following observations [12] The Kannada akshara shows three distinct vertical regions shown in Figure 9. The top zone, which ends just below the short head line of the consonant symbol, contains the vowel modifiers and sometimes parts of the base consonant. Some letters may not have a head line, and, in that case, the location of the top zone of neighboring letters can be used. The middle zone contains the consonant glyphs and some vowel modifiers. The vowel modifier glyphs may appear as either connected or disconnected components to the right of the base consonant. The middle zone ends where the base consonants end. The bottom zone which extends below the base consonant consists of glyphs for the consonant conjuncts and the glyphs for some vowel modifiers. These glyphs generally appear disconnected from the base consonant and the vowel modifiers present in the middle zone. The words are first vertically segmented into three zones. This segmentation is achieved by analyzing the horizontal projection profile of a word. Separating the middle zone from the bottom zone is easier because of the fact that the consonant conjuncts are always disconnected from the base consonant. Separating the top zone from the middle zone is more difficult. There are a few situations where the top zone as segmented may contain some of the base consonant or the middle one may contain some of the base consonant or the middle zone may contain a little bit of the top vowel

modifier. The next task is to segment the three zones horizontally. The middle zone is the most critical since it contains a major portion of the letter. The middle zone is therefore the first to be segmented. The aim is to separate the vowel modifier from the consonant. To achieve this segmentation we follow an over segment-and-merge approach. The middle zone is first over segmented by extracting points in the vertical projection (of the middle zone) showing drops in the histogram value exceeding a fixed threshold.

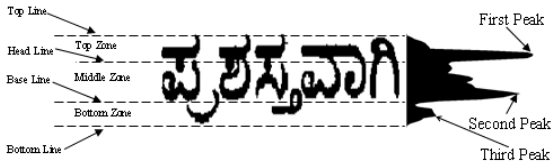


Figure 9. Three horizontal zones in Kannada word with subscript (vattaksharas).

3.5 Connected component analysis

Segmentation process consists of finding all connected component (CC) in the binary image of the document, for each CC, shape features are calculated such as the ratio between height and width, and the ratio between black and white pixels inside the CC bounding box. These features and all kind of graphics present in the document. Intuitively, most characters are connected components because all pixels touch each other.Connected component analysis is a two dimensional analysis that works well on proportional component fonts and handwritten characters. Connected component analysis can also be used to segment blocks of characters into individual words. Connected component analysis is 4 times faster [7]. And yielded half as many errors as projection analysis others have proposed a recognizer that does not perform character classification, but instead determines if the connected component is a single character or not [15]. The connected component method [16] first labels the pixels in the image. The pixels that are connected are labeled with the same blob. After labeling, the labeled components are extracted from the image. The CC method solves the overlapping character segmentation problem, but the separates simple character into their constituent glyphs which may increases the recognition complexity. For example the character will be segmented in to two glyphs [8] each. These glyphs are to be reassembled to preserve the character shape if the recognition phase uses the shapes of the basic characters. If the recognition system uses the basic characters then CC method is suitable. This method can solve the overlapping segmentation problem and makes use of the projection profiles and CC with some heuristics to segment the overlapped text documents in a robust way. CC pixels are connected to their close ones based on geometrical criteria to form text lines, other methods have also been proposed such as repulsive attractive network, stochastic method [10,21].

4. CONCLUSION

This paper addresses the segmentation techniques of Kannada text lines, words and characters. The proposed methods based on Horizontal and a Vertical projection profile, connected components is 4 times faster than projection analysis. Segmentation accuracy can be achieved with overlapping lines and characters. Due to many challenges in text line segmentation although methods have been proposed but the problem still remains open for subscript (vattaksharas).

ACKNOWLEDGMENTS

I owe my sincere feelings of gratitude to Dr.S. C. Sharma for his valuable guidance and suggestions which helped me a lot to write this paper. It gives great pleasure to express my feelings of gratitude to Dr. P. Ramakanth Kumar and Dr.M. Krishna for valuble guidance support and encouragement.

REFERENCES

[1] B. M. Sagar, G. Shobha and P. Ramakanth Kumar, Converting printed Kannada text image file to machine editable format using Database, *International Journal of Computers*, 2, **2008**, 173–175.

[2] M. A. Rahiman and M. S. Rajasree: A Detailed study and Analysis of OCR research in south Indian script, *International Conference on Advances in Recent Technologies in Communication and Computing*, Version 17, **2009**, 31–38.

[3] Seethalakshmi, Optical character recognition for printed Tamil text using unicode. *Journal of Zhejiang university Science* 6A, **2005**, 1297–1305.

[4] S. Mishra, D. Nanda and S. Mohanty, Oriya Character recognition using neural networks, *Special issue of International Conference IJCCT* 2, **2010**, 88–92.

[5] G. Richard, Casey and Eric Lecolinet. A survey of methods and strategies in character segmentation, *IEEE Transactions on Pattern Analysis and Machined Intelligence*, 18, **1996**, 690–706.

[6] M. S. Das, C. R. K. Reddy, A. Govardhan and G. Saikrishna, Segmentation of Overlapping Text lines, Characters in printed Telugu text document images, *International Journal of Engineering Science and Technology*, 2, **2010**, 6606–6610.

[7] R. A. Wilkinson and M. D. Garris. Comparison of massively parallel hand-print segmenters, *National Institute of Standards and Technology* (CSL), Advanced Systems Division, Gaithersburg, MD, 4923, **1992,** 196–214.

[8] S. B. Patil, Neural Network based bilingual OCR system: experiment with English and Kannada billingual document, *International Journal of Computer Applications*, 13, **2011**, 6–14.

[9] M. Thungamani and P. Ramakhant Kumar, Keshava Prasanna and S. K. Rao, Off-line handwritten kannada text recognition using support vector machine using zernike moments, *International Journal of Computer Science and Network Security*, 11, **2011**, 128–135.

[10] Z. Razak, K. Zulkiflee, R. Salleh, M. Yaacob and E. Mohd, Tamil: A real-time line segmentation algorithm for an offline overlapped handwritten jawi character recognition chip, *Malaysian Journal of Computer Science*, 20, **2007**, 171–182.

[11] R. S. Kunte and R. D. Sudhaker Samuel, A simple and efficient optical character recognition system for basic symbols in printed Kannada text: *Sadhana*, 32, **2007**, 521–533.

[12] B. M. Sagar, G. Shobha and P. Ramakanth Kumar, Character segmentation algorithm for Kannada optical character recognition, *Proceedings of the International conference on Wavelet Analysis and Pattern Recognition*, Hong Kong, 30–31, **2008**, 339-342.

[13] K. A. Kluever, Study report character segmentation and classification, http://www.tipstricks.org/example.asp, **2008,** 1–21.

[14] T. V. Ashwin and P. S. Sastry, A font and size-independent OCR system for printed Kannada documents using support vector machines: *Sadhana*, 27, **2002**, 35–58.

[15] J. Wang and J. Jean, Segmentation of merged characters by neural networks and shortest-path. In SAC '93, *Proceedings of the ACM/SIGAPP symposium on Applied computing*, **1993**, 762-769. New York, USA.

[16] P. Gader, M. Mohamed and J. H. Chiang, Segmentation-based handwritten word recognition. *Proceedings of USPS Fifth Advanced Technology Conference*, **1992**, 215–225.

[17] K. Sheshadri, T. Pavan Kumar and P. Ramakanth Kumar, An OCR system for printed Kannada using k-meansclustering, 978-1-4244-5697-0/10. *International Conference on Industrial Technology IEEE*, Vi a del Mar, **2010**, 183–187.

[18] N. Venkateswara Rao, A. Srikrishna, B. Raveendra Babu and G. R. M. Babu, An efficient feature extraction and classification of handwritten digits using neural networks, *International Journal of Computer Science, Engineering and Applications*, 1, **2011**, 47–56.

[19] B. Vijaya Kumar, Machine Recognition of printed Kannada Text. Version 3.0 Addison Wesley, IISc Bangalore, http://resources.metapress.com/pdf-review.axd? Code=09gn4jx2eqcwk9xy &size=largest. **2001,** 37–48

[20] N. Arica, F. T. Yarman–Vural, One- dimentional representation of two-dimentional information for HMM ased handwriting recognition, *Pattern Recognition Letters*, 21, **2000**, 583–592.

[21] I. Rios, A. de Souza Britto Jr, A. L. Koerich L. E. S. Oliveira, An OCR free method for word spotting in printed documents, *Journal of Universal Computer Science*, 17, **2011**, 48–63.

[22] S. Bag and G. Harit, A medial axis based thinning strategy for character images, (IIT) Karagpur,India, *In proceedings of the second National Conference on Computer Vision, pattern recognition, Image Processing and Graphics* (NCVPRIPG), Jaipur, India, http://arxiv.org/abs/1103.0738v1, **2010**, 67–72.