



www.icgst.com

GVIP

A Two-stage Character Segmentation Technique for Printed Kannada Text

R Sanjeev Kunte, Sudhaker Samuel R D
S J College of Engineering,
 Mysore-570006, Karnataka, India,
 [sanjeevkunte,sudhakersamuel]@yahoo.com,
<http://www.sjce.ac.in>

Abstract

Kannada is a widely spoken language of south India. Character segmentation in Kannada text is a difficult task since adjacent characters in a Kannada word sometimes overlap in the vertical projection profile due to the presence of bottom extension characters (Vatthus). In such cases, the usual method of character segmentation using projection profile is not efficient. In this paper we present a segmentation method in which the Vatthus of a Kannada word are segmented first by using connected component algorithm, and then the remaining characters of the word are segmented by vertical projection profile method. The segmentation method was tested on a large data set of 25000 words and a segmentation rate of 98.2% is achieved.

Keywords: *Kannada text, Projection profile, Character segmentation, Connected components.*

1 Introduction

Optical Character Recognition (OCR) systems have been effectively developed for the recognition of printed characters of non-Indian languages. Efforts are on the way for the development of an efficient OCR system for Indian languages, especially for Kannada, a popular South Indian language, which is very rich in alphabet. In OCR systems efficient character segmentation is a crucial preprocessing step for reliable character recognition. The overall success (recognition) rate of an OCR system depends heavily on the proper segmentation of characters.

A typical OCR system consists of image capturing, preprocessing, segmentation, feature extraction and recognition stages. Segmentation refers to extraction of objects of interest from rest of the image. It is one of the decision stages in an OCR system, because incorrectly segmented characters will not be

recognized properly. This reduces the recognition rate of the OCR system.

The methods of segmentation are broadly classified into three strategies [4] , [11] as follows:

- The classical or dissection approach: In this approach, the segments are identified by extracting the distinguishing attributes of the character image.
- Recognition-based segmentation: The image as a whole is searched for components that match predefined classes.
- Holistic approach: The system tries to recognize the word as a whole.

The classical segmentation approach uses several methods of segmentation [4] such as White Space and Pitch, Projection Analysis, Connected Component Processing, etc. The most common approach is to use Projection Profile Analysis since it is simple and fast.

In Kannada text, for words having bottom extension characters (Vatthus), the space between two adjacent characters does not have zero spaced valleys in the vertical projection profile, which makes it difficult to extract individual characters from the word as shown in Figure (8). In such situations the usual approach for segmentation is to use the connected component method by treating each of the individual characters in the text (both main and Vatthu characters) as separate components of the image. This however is both time consuming and computationally expensive method.

In this paper we describe a two-stage method, in which, only the Vatthus (which are usually few in number) are segmented from the word using connected component processing in the first stage. In the second stage the remaining characters from word (with Vatthus removed) are easily segmented using the traditional vertical projection profile method. As seg-

mentation of characters by projection method is simple and fast, the proposed two-stage method for Kannada character segmentation is faster than the conventional method in which all the characters from the text are segmented by connected component processing only.

The major strengths of the proposed two-stage method for segmenting Kannada characters which combines connected component analysis and projection profile is that, it works faster than classical single-stage method of segmenting characters using connected component analysis only. Also, the proposed method can be used without any modifications for segmenting the characters from documents for any other language (for example- Telugu -another widely spoken south Indian language) which have a word structure consisting of 'main' and 'Vatthu' like characters.

The rest of this paper is organized as follows: Section(2) presents a survey of work done in Character segmentation. Section (3) describes the properties of Kannada language and introduces the Kannada alphabets. Section (4) presents the segmentation procedure adopted. Section (5) provides experimental results. Conclusion is presented in section (6).

2 Literature Survey

Currently there are many OCR systems available for handling printed/handwritten English documents with reasonable levels of accuracy. Such systems are also available for many European languages as well as some of Asian languages such as Japanese, Chinese etc. However, there are not many reported efforts at developing OCR systems for Indian languages especially for a South Indian language like Kannada [7] , [8].

Some of the previous works in Character segmentation (a stage of OCR system) for different languages is given below.

Amara and Nouredine [9] have described a method for segmentation of printed Arabic characters using a modified histogram as well as the number of black segments in a line of pixels.

Veena and Mishra [3] have proposed a method for segmentation of touching and fused Devanagari characters in two stages. In the first stage the words are segmented into easily separable characters or composite characters. Statistical information about the height and width of each separated box is used to hypothesize whether a character box is composite. In the second stage the hypothesized characters are further segmented.

Pal and Sagarika [10] have proposed a method of segmentation of unconstrained Bangla characters based on piece-wise projection for the segmentation of Lines, and Water reservoir principle for segmenting characters inside a word.

In the Kannada OCR system proposed by Ashwin

and Sastry [2] a segmentation method is described in which the words are first vertically segmented into three zones. This segmentation is achieved by analyzing the horizontal projection profile of a word. Later the three zones are segmented horizontally to extract the characters into their constituents, i.e. the base consonant, the vowel modifier and the consonant conjunct.

But, there are a few situations where the segmented top zone may contain some of the base consonant or the middle zone may contain a little bit of the top vowel modifier. This reduces the recognition rate of the OCR system. This particular problem is solved in our segmentation method by extracting the character as a whole instead of its constituents.

Kimura [5] proposes a bounding box method in where, a word image is split into horizontally overlapping zones. A connected component analysis is applied to detect the boxes enclosing each connected component for segmentation of characters.

In the OCR system for Tamil, proposed by Aparna and Ramakirshnan [1], horizontal and vertical projection profiles are employed for line and word segmentation. Connected component analysis is performed on the words to extract the individual characters.

Anniwear and Yoshinao [12] proposes a segmentation technique for Uygur Scripts. Segmentation is done by line separation and word separation done using projection profile. The characters from the word are isolated using a two-step algorithm i.e, a topological segmentation, and quasi-topological segmentation. Topological segmentation is based on tracing the outer contour of a given word. Quasi-topological segmentation is based on the decision to section a character on a combination of feature-extraction and character-width measurements.

3 Overview of Kannada language

Kannada is one of the four popular Dravidian languages of South India. Kannada script is written horizontally from left to right and the concept of lower and upper case is absent. Kannada is a non-cursive script i.e. a Kannada word is written without joining the characters of the word. The characters are isolated within a word. Kannada language has 16 vowels and 34 consonants as the basic alphabet of the language as shown in Figure (1) and Figure (2), respectively.

Each vowel has a vowel sign (modifier) and each consonant has a basic form (primitive). A basic consonant can combine with the vowel sign to form another set of 16 Consonant-Vowel (CV) composite characters called as *gunithakshara*. Such an example is shown in Figure (3) for the first consonant (ka). Including the basic form of the consonant, each consonant group is a set of 17 characters.

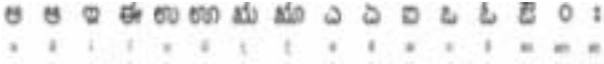


Figure 1: Kannada Vowels



Figure 2: Kannada Consonants



Figure 3: Consonant-Vowel composite characters of basic consonant

In Kannada, all the 34 consonants have a half/short form, which can be referred as half consonant (Vatthu or Subscript) as shown in Figure (4). Any half consonant can appear below any other consonant or a CV character as a bottom extension character to form a conjunct-consonant character. Examples of three such characters are shown in Figure (5).



Figure 4: Short forms/ Half Consonants (Vatthus)



Figure 5: A few examples of conjunct-consonants

In the rest of the paper Vatthus are referred as subscripts and all other characters (vowels, consonants,

CV characters) other than the subscripts are referred as main characters.

4 Segmentation Methodology

Our segmentation system uses the classical approach in which the scanned image is dissected into individual building blocks to be recognized as characters.

The proposed method starts by segmenting the lines and then the words from the scanned document image using horizontal and vertical projection profiles respectively. Each segmented word is then examined for the presence of subscript characters. If subscript characters are present in the word, then all of them are extracted using the connected component method. If subscripts are not present (or when all the subscripts are extracted) the main characters from the word are segmented using vertical projection profile.

The details of the segmentation methodology adopted for segmentation of lines, words and characters are now described.

4.1 Line Segmentation

To separate the text lines, the horizontal projection profile of the text 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 (6) shows a sample Kannada document along with its horizontal projection. The projection profile has valleys of zero height between the text lines. Line segmentation is done at these points.

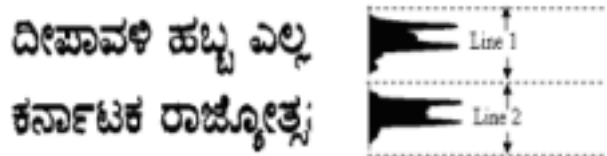


Figure 6: Text lines with Horizontal projection profiles and Line segmentation details for a sample Kannada text document

4.2 Word Segmentation

The spacing between the words is used for word segmentation. For Kannada script, spacing between the words is greater than the spacing between characters in a word. The spacing between the words is found by taking the vertical projection profile of an input text line. Vertical projection profile is the sum of ON pixels along every column of the image.

A sample input text line and its vertical projection profile is shown in Figure (7). From the vertical projection profile it can be observed that, the width of zero-valued valleys is more between the words in the

line as compared to the width of zero-valued valleys that exists between characters in a word. This information is used to separate words from the input text lines.

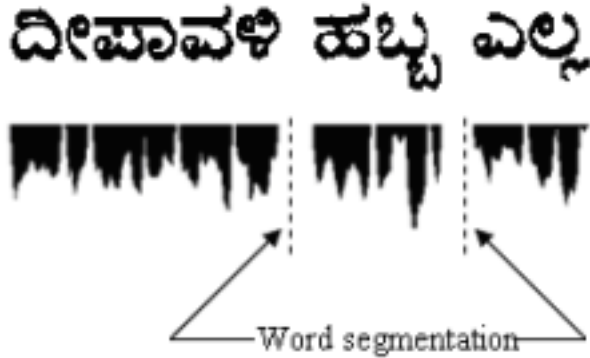


Figure 7: Input text line and its vertical projection profile indicating word segmentation for a sample document

4.3 Character Segmentation

As Kannada is a non-cursive script, the individual characters in a word are isolated. Spacing between the characters can be used for segmentation. But, sometimes in the vertical projection profile of a word, there will not be any zero-valued valleys, due to the presence of conjunct-consonant (subscripts) characters. The subscript character position overlaps with the two adjacent main characters in vertical direction as shown in Figure (8).

Hence in these cases the usual method of vertical projection profile to separate characters is not possible. In these cases the following two-stage approach is used:

Stage 1:

- Check for the presence of subscripts in a word.
- If subscripts are present, they are extracted first from the word using connected component method.

Stage 2:

Remaining characters from the word are extracted using vertical projection profile.

If subscripts are not present in a word then the characters from the word are extracted using vertical projection profile in one stage itself.

Thus, for character segmentation it is first necessary to check whether there are any subscripts in a word. For this a Kannada word is divided into different horizontal zones, as explained in section 4.3.1.

4.3.1 Zones in a Kannada word

A Kannada word can be divided into different horizontal zones. Two different cases are considered. Case (i) A word without subscripts as shown in Figure (9) (pronounced as *ramanu*) and Case (ii) A word

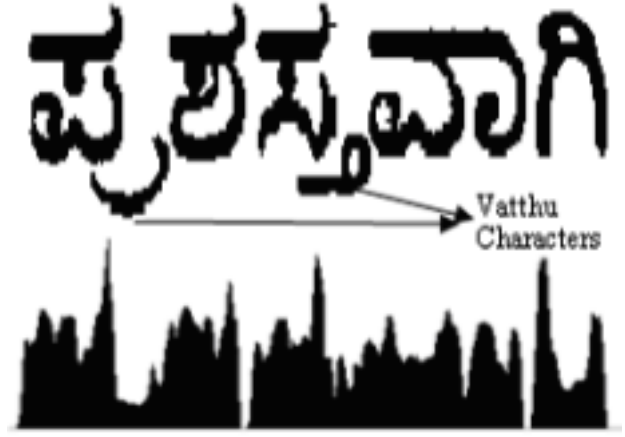


Figure 8: A sample Kannada word with subscripts along with its vertical projection profile

with subscripts as show in Figure (10)(pronounced as *prashastavaagi*).

Consider the sample word shown in Figure (9), which does not have a subscript character. The imaginary horizontal line that passes through the topmost pixel of the word is the top line. Similarly, the horizontal line that is passing through the bottommost pixel of the main character is the base line. The horizontal line passing through the first peak in the profile is the head line. The word can be divided into top and middle zones. Top zone is the portion between the top and head line and the middle zone is the portion between the head line and the base line.

For the words with conjunct-consonant characters, it is divided into three horizontal zones as shown in Figure (10) for a sample word with subscripts. The word is divided into top, middle and bottom zones. The top and middle zones are chosen similar to that of the word without subscripts. A bottom portion is chosen between the base line and the bottom line. The bottom line is the horizontal line passing through the bottommost pixel of the word.

Before character segmentation it is first necessary to find out whether the segmented word has a subscript or not. This can be detected as follows:

Close observation of Figure (9) and Figure (10) reveals that,

- There are several peaks in the horizontal projection profile of the sample Kannada word in Figure (9) and Figure (10).
- However there are only two peaks in horizontal projection profile of approximately equal size in Figure (9) and Figure (10).
- In the horizontal projection profile shown in Figure (10), there are two peaks of approximately equal size in top and middle zones of the word. Also there is an occurrence of third peak after



Figure 9: Two horizontal zones in a sample word without conjunct-consonant character

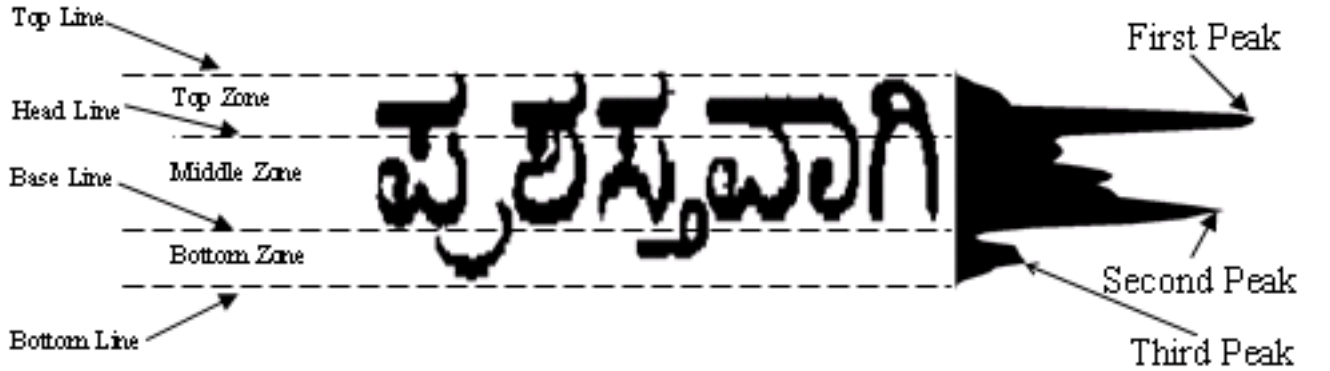


Figure 10: Three horizontal zones in a Kannada Word with subscripts

the second peak in the bottom zone of the word which is due to the subscripts in the word.

- Where as, in horizontal projection profile shown in Figure (9), there are only two peaks of approximately equal size in the top zone and middle zones of the word. The absence of the third peak, after the second peak indicates that there are no subscripts in the word.

Thus, by checking for the presence or absence of the third peak (after the first two approximately equal sized peaks) in the bottom zone of the horizontal projection profile of the segmented Kannada word, it is possible to find out whether the segmented word has a subscript or not.

4.3.2 Character Segmentation of a Word without subscripts

Let us first consider a Kannada word which does not have any subscripts. An example is shown in Figure (11) (pronounced as *deepavali*) along with its vertical projection profile.

In Fig (11), the presence of zero-valued valleys in the vertical projection profile of the word can be observed between all the characters of a word which makes the character separation easier. The word is examined row-wise. The portion of the image, which

lies between two successive zero-valued valleys of the vertical projection profile is assumed to be as a separate character and separated out.

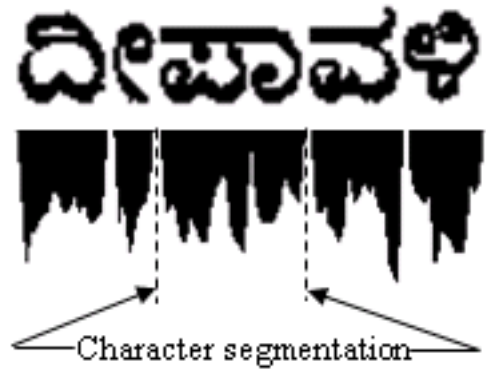


Figure 11: Vertical projection profile of a word that does not have any subscripts

4.3.3 Character Segmentation of a Word having subscripts

Let us consider a sample Kannada word as shown in Figure (12) (pronounced as *prasanna*) which contains subscripts. If vertical projection of this word is considered, then there will be no zero-valued valleys between

the first character ಷ and its subscript character ಷ. So also for the last character and its subscript. Hence, just the zero-valued valleys of the vertical projection merely do not indicate character separation. The individual characters in this case, are separated in two stages as follows:

- In the first stage, the subscripts of the word are separated out.
- In the second stage, from the plain word (in which subscripts are now removed) the individual characters are extracted as explained in section 4.3.2.

Stage1: subscript character segmentation

Consider a sample word as shown in Figure (12). The total height of the word in terms of number of rows is found. Let it be H .

- Columns of the word are scanned from left to right. Every column is scanned from bottom row to the top row to find the presence of an ON pixel p . When such an ON pixel is found, the number of rows that has gone up is counted. Let it be L .
- If L is less than or equal to some threshold value τ , the pixel p is assumed to be one of the points of the subscript character. Then, using p as the initial point, the connected component algorithm [6] is applied to extract the subscript character at that position.
- The value for τ is fixed by finding the position of the valley (minimum value) between the second peak and the third peak which will be just below the base line in the bottom zone of the word.

In Figure (12) the sample word has three main characters and two subscript characters. During initial scanning from leftmost column to right column, a ON pixel point m will be detected. Since it is at height greater than τ from bottom, it does not belong to a subscript character and hence it is discarded. As scanning proceeds, at point p a valid point for a subscript character is detected. Hence, the subscript character ಷ is extracted using connected component algorithm. Similarly, the second subscript character ಷ is extracted subsequently as the scanning proceeds towards right.

Connected Component Analysis on an image is done in order to extract a group of pixels connected by 8-connectivity. By knowing one of the inside point of a connected component in an image, it is possible to extract all the pixels of the component in an iterative manner using connected component algorithm.

The scanning process is repeated till the end of the word (right most columns) to extract all the subscript characters present in the word.

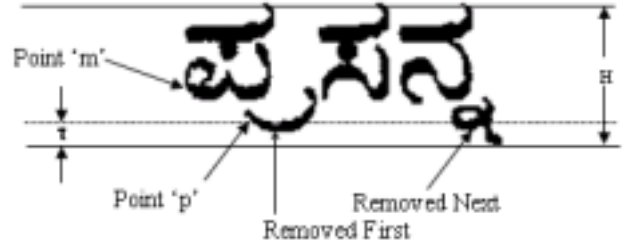


Figure 12: A sample word with subscripts and details of First stage of Character segmentation

At the end of stage 1, after separating subscripts what remains is a plain word without having any subscript characters as shown in Figure (13).

Stage 2: Main character segmentation

The output of the first stage converts the word with subscripts into a plain word without any subscript characters. Hence, during the second stage, the same method used for segmenting the characters (for a word without subscripts) is followed for segmenting the main characters as explained in section 4.3.2.



Figure 13: Plain word of Figure 12 after removal of subscripts

5 Experimental Results and Discussion

The segmentation algorithm has been tested on many printed Kannada documents for about 25000 words which contained characters of different fonts and size varying from 14 to 20 font sizes. We considered only good quality of printed documents where there are no touching or broken characters. Correct segmentation obtained for 98.2% of the characters.

The time taken for segmenting a sample A4 page of 300 Kannada words by the proposed method was about 1.25 seconds.

The same sample page was segmented using the conventional method in which all the characters of the text are treated as individual components and all of them are extracted using the method of connected component. The time taken to segment was about 18 seconds.

This shows the proposed two-stage method is about 15 times faster than conventional connected component processing method.

Time complexity:

In single-stage connected component analysis method, for the segmentation of the characters, the

method is applied for both subscripts and main characters. Where as, in the proposed two-stage method, only subscripts are extracted using connected component analysis. The main characters are segmented using the simple projection profile method which does not take much time.

Normally, the ratio of subscripts to main characters in a Kannada document is very small. Hence, the time taken for segmentation by the proposed two-stage method is much faster.

Let us consider a sample A4 page of Kannada document consisting of about 300 words, comprising of about 1400 main characters and 100 subscript characters totaling 1500 characters or about 1500 components.

If the time taken to extract a single connected component is T , then,

(i) The time taken for segmentation by the proposed two-stage method:

$$\text{Time (two-stage)} = 100 * T + t;$$

Where, $100 * T$ is time taken for extracting 100 subscripts, and t is time taken for finding the presence of subscripts in 300 words plus time taken for segmentation of 1400 main characters using vertical projection profile

(ii) The time taken for segmentation by the classical single stage method:

Time (single-stage) = $1500 * T$; corresponding to 1500 components

Extracting a character using connected component analysis is an iterative procedure compared to a single step segmentation using valleys in a projection profile. Hence, value of 't' is much less significant with the magnitude of $100 * T$.

Therefore, the segmentation time of the proposed two-stage method is far less than conventional single-stage method of connected component analysis only.

Failure Cases:

We observed two cases in which the direct adoption of our two-stage algorithm fails.

Case (i)

In some cases we observed that the main character itself descends down in the bottom zone. Such an example is shown in Figure (14) (pronounced as *bhavya*). At that time, the main character itself gets extracted as a subscript character. But, by checking the total height of the character extracted it can be considered as a main or subscript character. subscript character's heights are approximately $1/3^{rd}$ of main character's height.

Case (ii)

For some Kannada words even without subscripts, in the horizontal histogram, the third peak may be present due to the extension of the vowel modifier to the bottom zone as shown for a sample word in Figure (15) (pronounced as *phalavoo*). In such cases also the main characters itself get extracted as subscript character in first stage of our two-stage

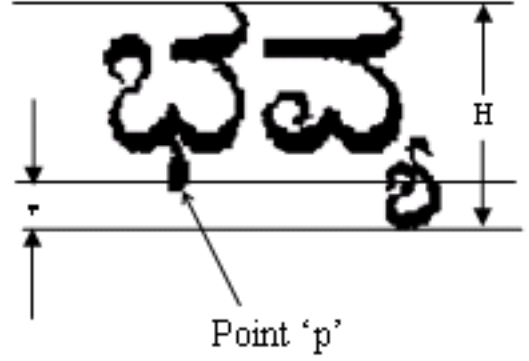


Figure 14: Kannada word having a character with its vowel modifier extended to bottom zone

process giving a wrong result. But, once again by checking the total height of the character extracted, it can be considered as a main or subscript character.



Figure 15: A sample word having characters with their vowel modifier extended to bottom zone and its Horizontal Projection Profile

Future work: In the proposed method only good quality of printed document is considered without any touching or broken characters. The proposed method can be extended to include the touching or broken characters in the document.

6 Conclusion

In this paper we have proposed a two-stage method for segmentation of Kannada characters in the presence of Vatthu characters. Here the Vatthu characters of a word are segmented first using connected component algorithm and then the remaining characters are segmented by usual vertical projection profile method. The proposed method is faster than the conventional connected component processing method. The proposed method can be applied without any modifications for segmentation of characters from a Telugu document (another South Indian language) which has its alphabet set consisting of main and Vatthu characters similar to Kannada language except for the changes in the shape of the characters.

References

- [1] K. G. Aparna and A. G. Ramakrishnan. A complete Tamil Optical Character Recognition System. In *proceedings of Document Analysis Systems V, 5th International Workshop*, pages 53–57, 2002.
- [2] T.V Ashwin and P.S. Sastry. Font and size independent OCR for printed Kannada documents using SVM classifier. *Sadhana*, 27:35–57, 2002.
- [3] Veena Bansal and R.M.K. Mishra. Segmentation of touching and fused Devanagari characters. *Pattern Recognition*, 35:875–893, 2002.
- [4] G. Richard Casey and Eric Lecolinet. A Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on PAMI*, 18(7):690–706, 1996.
- [5] M. Shridhar F. Kimura and Z. Chen. Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words. In *proceedings of ICDAR 1993*, 1993.
- [6] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison Wesley, 1993.
- [7] R. Srinivasa Rao Kunte and R.D. Sudhaker Samuel. Wavelet Features Based On-Line Character Recognition for Handwritten Kannada Characters. In *proceedings of 2000 IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, pages 605–608, 2000.
- [8] P. Nagabhushan and M. Radhika Pai. Modified region decomposition method and optimal depth decision tree in the recognition of non-uniform sized characters- An experimentation with Kannada characters. *Pattern Recognition*, 20(14):1467–1475, 1999.
- [9] B. Amara Najouna and E. Noureddine. A Robust Approach for Arabic Printed Character Segmentation. In *proceedings of ICDAR 2003*, pages 865–868, 2003.
- [10] U. Pal and Sagarika Datta. Segmentation of Bangla Unconstrained Handwritten Text. In *proceedings of ICDAR 2003*, pages 1128–1132, 2003.
- [11] Lu. Yi. Machine Printed Character Segmentation- An Overview. *Pattern Recognition*, 28(1):67–80, 1995.
- [12] Anniwear YMIN and Yoshinao AOKI. On the segmentation of Multi-font printed Uygur Scripts. In *proceedings of International Conference on Pattern Recognition*, pages 215–219, 1996.