Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents

Dr.S.Basavaraj Patil
AMC Engineering College
18th KM, Bannerghatta Road
Bangalore-560083, India.

ABSTRACT
The paper presents the Neural Network based Bilingual OCR system which can read printed document images, written in two scripts of English and Kannada languages. Such systems are highly preferred in automation of multi-script, multi lingual document processing. The developed system includes document image pre-processor, dynamic feature extractor, neural network based script classifier, Kannada character recognition system and English character recognition system. Document image pre-processor, accepts the bilingual document image and performs grey to two tone conversion, segmentation into lines and words. Dynamic feature extractor extracts distinctive equal number of features from each separated word irrespective of size of the word. These features are accepted by probabilistic neural classifier and are sorted by script, Kannada and Roman. Developed Kannada character recognition system accepts these words and further segments each word into characters and maps the recognized characters into corresponding ASCII values of the chosen Kannada font. Similarly specifically developed English character recognition system, segments English words into characters and maps to corresponding ASCII value of the specific English font. Thus recognized English and Kannada characters are written into separate ASCII files language wise. The results are exciting and proved the effectiveness of the approach.

General Terms
Pattern Recognition, Script Identification, Neural Networks, Optical Character Recognizer(OCR)

Keywords
Script Classification, Kannada Character Recognition, Bilingual OCR

1. INTRODUCTION
Document Processing in the Indian environment has special significance, since eighteen official languages are in use in the country. Throughout the country, every government office uses at least two languages, English and the official language of the corresponding state. The state of Karnataka has an official language as Kannada, however many national organizations such as Banks, use English and Kannada. Even all the documents in the government offices of Karnataka state usually appear in two languages, Kannada and English. This is the major reason for choosing these two languages for experimentation in Bilingual OCR system. The aim of the automation of document processing is to convert the scanned paper document to the machine readable codes such as ASCII. In this paper the same is experimented by taking the bilingual documents printed in English and Kannada languages. Here the bilingual document image is finally converted into two machine understandable ASCII files.

The major function of an OCR system which can read two language scripts, English and Kannada is to separate words in the document script/language wise and then feed them to the corresponding OCR systems to convert the same to machine codes. For this purpose the Kannada OCR system has been developed and used. English OCR system is implemented and both Kannada and English OCR systems are combined to form the Bilingual OCR system.

Researchers of an International community have some contributions in the document script identification[1-4] but rarely attempted to process multi-script, multi-lingual documents. It may be due to the reason that such a multi-script, multi-lingual documents appear only in Indian Society. At national level the contribution starts from from B.B.Chauduri and U.Pal. They have discussed an OCR system to read two Indian language scripts, Bangla and Devnagari (Hindi) [5]. The description of their system is as follows.

The diagrammatic structure of their OCR system is shown in figure 1. The system performs text digitization, gray-tone to two tone conversion, line and word detection, character segmentation followed by actual character recognition. In this system they have adopted manual method to switch between the Banlga and Devanagari depending upon the document content. Depending upon the position of the mechanical switch, the OCR system detects the features required for classification and then the characters are converted to suitable machine codes.

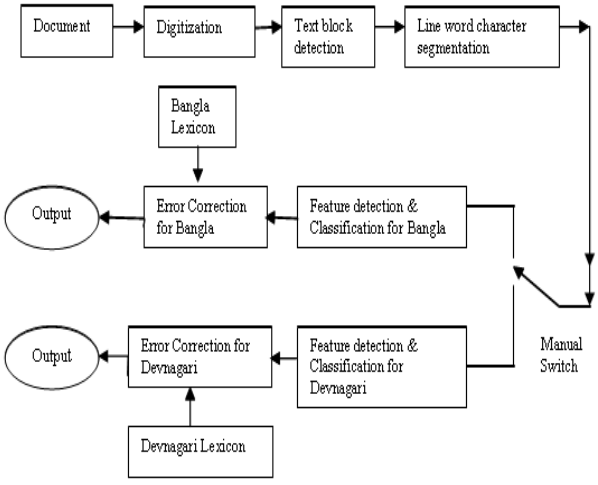


Figure 1: Bangla and Devnagari Bilingual OCR

Recently Sanghamitra Mohanty et al [6] discussed Bilingual Script Identification and Recognition Method. They differentiated script line-wise. Their method utilized the horizontal histogram for line distinction belonging to different script. Javawahar et al [7] developed Hindi-Telugu Bilingual

OCR. Their bilingual recognizer is based on principal component analysis followed by support vector classification. Sanjeev and Sudhaker [8] used Gabor filter based features for separating Kannada and English words from bilingual documents. They used neural classifiers for classification. Latest work from Padma et al [9] describes the method based on the distinct features extracted from the top and bottom profiles of the individual text lines. The method is simple, as it does not require any character or word segmentation.

The organization of this paper is as follows. In section 2, we describe developed Bilingual OCR system. In section 3, we describe Kannada OCR system and specifically developed English OCR system is discussed in section 4. In section 5, we report the experiments carried out on the Bilingual OCR system along with the results obtained. Finally the conclusions are made in section 6.

2. DEVELOPED BILINGUAL OCR SYSTEM

The following important facts are to be noted in the above discussed B.B.Chauduri's system. It requires the user to operate a manual switch to perform the OCR function. That indicates that there is no script identification technique embedded in this system. It is just a simple combination of the two character recognizers and hence may not be useful to process directly the bilingual documents. For processing bilingual documents, that is the documents printed in two languages, they first recommend to separate the words, script wise from such documents before feeding it to their OCR system, using their word script separation technique [5]. In the developed Bilingual OCR system presented here the above drawback is eliminated. In the system presented here bilingual document directly can be processed and there is no need to separate the words script wise beforehand, because automatic neural based script identification system for individual word is embedded[10].

The developed Bilingual OCR system is shown in figure 2. The system diagram of the developed Bilingual OCR system presents the complete implementation details. The system includes automatic word script identification system for two language scripts. This subsystem has already been discussed in [10]. Here we focus on Bilingual OCR systems as a whole. It includes another subsystem Kannada OCR system which will be discussed in next section. English OCR system for single font, has been developed specifically for this purpose and embedded. This system has been explained in the next subsection. The Bilingual OCR system algorithm is as follows.

Algorithm 1: Bilingual OCR system algorithm

1. Accept the bilingual document image printed in English and Kannada languages.
2. Apply segmentation algorithms. These algorithms are based on white spaces between lines and words as dealt in [10]. They return individual word document images of the input document in a words array and also total number of words.
3. Initialize two new arrays lang1_words and lang2_words
4. Make i=1
5. Select i^{th} word document image from the 'words' array.
6. Extract the features from above word document image [10].
7. Pass the above extracted features to Individual word script identification system [6]. Identify the script of the word.
8. If the script of the word is Kannada (classifier output '1') store the word in 'lang1_words' array else (classifier output '2') store the word in 'lang2_words' array.
9. Make i=i+1
10. If i is less than or equal to the total number of words (found in step 2) in the document go to step 5 else go to next step 11.
11. Pass all the words of lang1_words array to Kannada OCR system (discussed in section below). The system segments each word into characters and recognizes them by decomposing into subparts. It reconstructs decomposed parts by allotting them suitable ASCII codes of the specific font, shreelipi851. These mapped ASCII codes will be written into new text file 'Kan1.txt'.
12. Pass all the words of lang2_words array to English OCR system (from the step2 method). The system segments each word into isolated characters and extracts the features. These features are fed to the PNN classifier which has been trained on that specific font (Courier New font) with small letters a to z and capital letters A to Z . This classifier recognizes English characters and writes equivalent ASCII codes into the new text file 'Engl.txt'.

3. KANNADA OCR SYSTEM

Although a lot of research work is available, most of it is on English alphabets and numerals which can be enclosed in standard rectangular structures [11-13]. Sufficient work has been done for the recognition of Chinese characters, Korean characters, Japanese characters[14-15]. Some reports are also available on Indian character recognition such as Bangla, and Devnagari [4,5,16].

Generally a character is enclosed in a rectangular structure for the purpose of recognition. Suen [17] also proposed this type of rectangular grid structure for recognition. Since their purpose is to identify English alphabets and numerals (where most of them are of uniform sized) they preferred rectangular grid structure of the constant size. Further they recommended to divide the rectangle into 2, 4 or 6 equal parts. The character is recognized by examining all these parts individually. But such an approach is more suitable for uniform sized characters like English.

Recognition of non-uniform sized characters is rarely available in the literature. Very little work is seen in the literature on recognition of Indian languages in general [18] and particularly on Kannada characters. Recently, Nagabhushan and Pai [19] presented a region decomposition method for recognition of such characters. They used the 3x3 bricks for representation of characters. By adding one column structure of 3 bricks horizontally, they create the horizontal extension. Similarly by adding one row structure vertically they create the vertical extension. The recognition is done based on an optimal depth logical decision tree developed

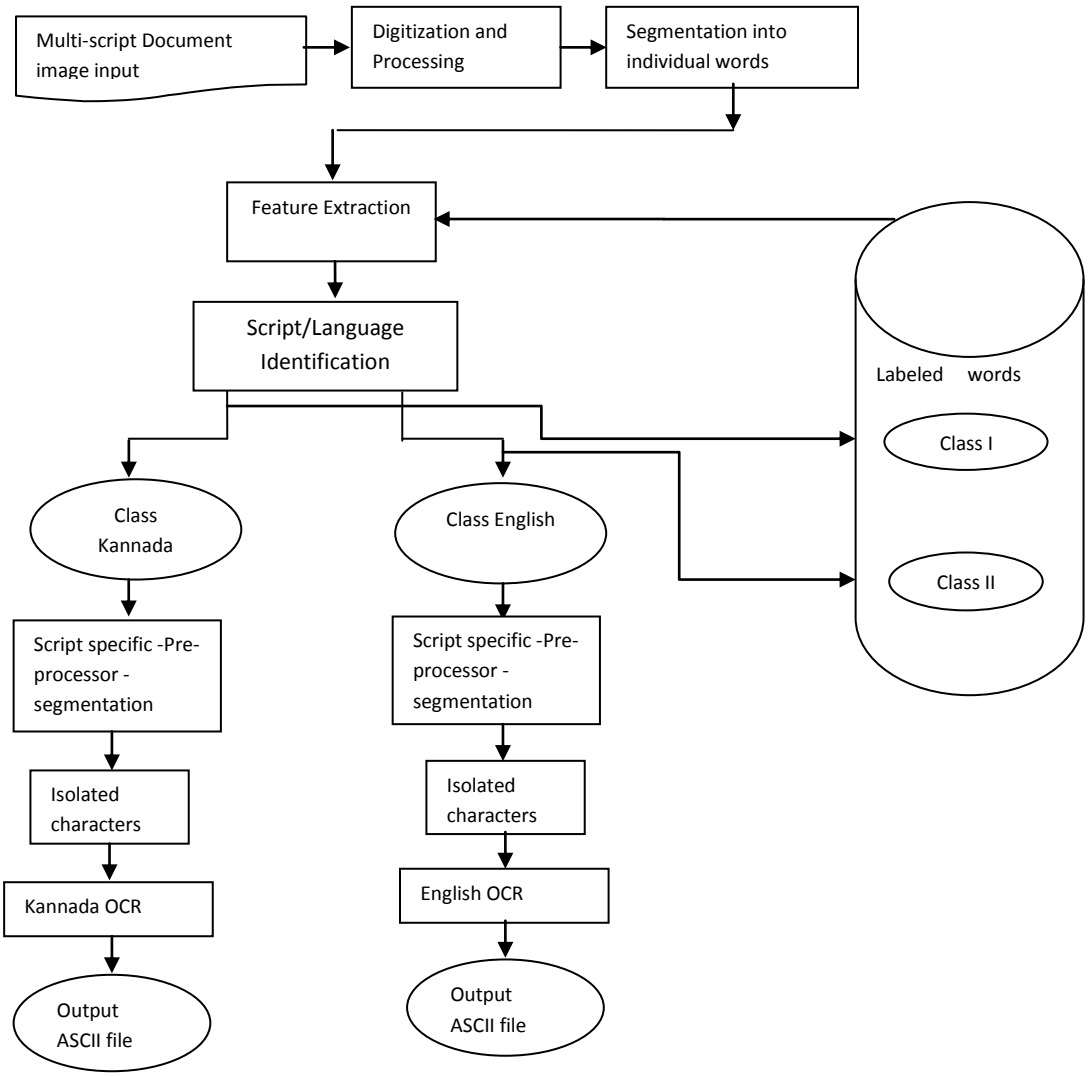


Figure 2 : Developed System

Table 1a-d: List of basic Kannada characters

Character	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಅಂ	ಅಃ
Phoneme	a	aa	e	ee	u	uu	Ru	ruu	ae	aee	i	o	oo	ow	am	aha

(a)

Character	ಕ	ಖ	ಗ	ಘ	ಙ	ಚ	ಛ	ಜ	ಝ	ಞ	ಟ	ಠ	ಡ	ಢ	ಣ	ತ
Phoneme	ka	kha	ga	gh a	nn a	Cha	cha a	ja	jha	ny a	ta	tha	da	dh a	na	ta

(b)

Character	ಥ	ದ	ಧ	ನ	ಪ	ಫ	ಬ	ಭ	ಮ	ಯ	ರ	ಲ	ವ
Phoneme	tha	da	dha	na	Pa	pha	ba	bha	ma	ya	ra	la	va

(c)

character	ಶ	ಷ	ಸ	ಹ	ಳ
Phoneme	sha	shaa	sa	ha	lha

(d)

Table 2: Modified forms of the consonant character, the ' ಕ '

Modifier	◌	◌ಾ	◌ಿ	◌ೀ	◌ು	◌ೂ	◌ೃ	◌ೄ	◌ೆ	◌ೇ	◌ೈ	◌ೊ	◌ೋ	◌ೌ	◌ಂ	◌ಃ
Consonan	ಕ್	ಕಾ	ಕಿ	ಕೀ	ಕು	ಕೂ	ಕೃ	ಕೄ	ಕೆ	ಕೇ	ಕೈ	ಕೊ	ಕೋ	ಕೌ	ಕಂ	ಕಃ
Phoneme	ka	kaa	ki	kee	ku	kuu	kru	kruu	ke	Ke	kai	ko	koo	kau	ku m	kaha

during the learning phase. They carried out an experiment to recognize the plane printed Kannada characters and not compound characters. And also they have not worked on any commercially available font but created their own for recognition conveniences. Another significant work in this connection is from Ashwin and Sastry [20]. They achieved recognition of Kannada characters by employing a number of 2 class classifiers based on support vector machine method. In the work presented here not only such compound characters are also considered but also a commercially available font is considered.

3.1 Properties of Kannada characters

Kannada character set has 50 basic characters out of which the first 16 are vowels and the last 34 are consonants. The list of basic characters is given in Table 1. A consonant combined with a vowel forms a modified character. Table 2 depicts the different forms of a consonant for the first consonant character obtained by application of the modifiers. Sometimes two or more consonants combine together to form a compound character. In these compound characters, the second consonant is written as subscripts

3.2 Character Model

In the beginning standard sized rectangular grid which is of the form figure 3 is thought of.

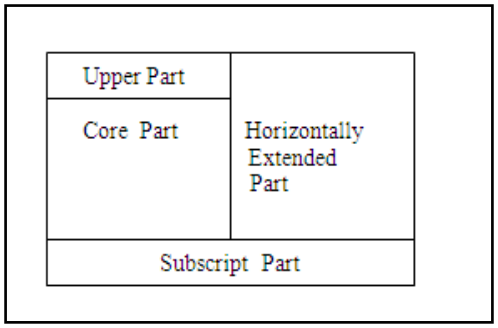


Figure 3: Rectangular Grid for characters

This rectangular grid has 4 fixed parts as upper part, core part, horizontally extended part and subscript part. Commercially available shree-lipi font has been verified and found that all the characters do not agree with this type of character model. For some characters upper part, lied with 3 rows of pixels and core part started in the 4th row but other type of characters took 4 rows for upper part and core part started with next row. If the area for upper part is fixed either for 3 rows or 4 rows in both the cases possibility of misrecognition exists. Similarly this fixed character model did not satisfy the horizontally extended parts. Even the core part, for some characters like, also ಕ್ಷ took very few columns whereas the character like ಟಾ took more number of columns than normal number of columns. Because all the above reasons, the above model has been modified as variable sized rectangular grid model.

3.3 Proposed Methodology

In this section a new method for recognition of Kannada characters, is presented. To recognize a Kannada character, decomposition is a necessity. For decomposing the character, variable sized rectangular grid model is adopted. The method of decomposing a character is as follows.

Method 1: Kannada character recognition

1> ***Decomposing the upper part***: First consider the prototype set of all upper modifier parts in the unknown character for their existence. If any of the upper modifier is matched to the top left rectangular portion of unknown character, then decompose that part. If no upper modifier matches with the top left rectangular portion, then there is no necessity to decompose the character. In such cases except subscript whole character map is considered for the recognition.

2> ***Subscript decomposition:*** In the given unknown character bitmap, count the number of rows. If it is more than the threshold limit ' Rth' (which has been assigned the value after a manual study of all possible characters in a particular font of specific size, (In the experimented case for example it is Shreelipi font of size 12) then the subscript must be present, otherwise there is no subscript. If the subscript is present decompose all the rows other than 'Rth' or fs (fs= Total number of rows- Rl) rows from the bottom of the character bit map.

3> ***Decomposition of horizontal extension:*** To extract the horizontally extended part, consider the prototype set of horizontal extensions. From the set take one at a time and verify for its existence by going from right side of the character bitmap. If any of the horizontal extensions are present then extract that part, from the character. If none of the horizontal extensions match with unknown character then decide that there is no horizontal extension.

3.4 Developed Kannada OCR system

Developed system for Kannada character recognition is shown in figure 4. The input document image is pre-processed into lines, words and individual isolated characters. Since document images used are written in MS-Paint programme segmentation has been achieved by performing simple black to white transitions. All the resultant isolated individual characters are stored. These characters processed by developed system one after the other in following steps.

1. Character is decomposed into possible number of parts (maximum 4 parts) as explained in section 3.3.

2. While decomposition some parts of the character might have been recognized If not recognized, recognize the individual part by rule based recognizers.

3. Combine all the parts and construct a character and map into character ID

4. Using the above character ID's reconstruct the word and sentence and paragraph.

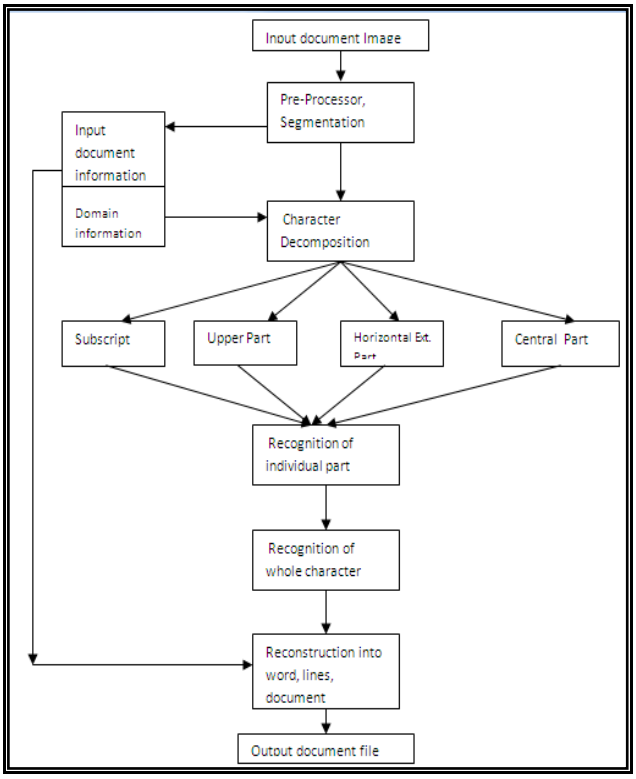


Figure 4: Developed Kannada OCR system

3.5 Experiments and results

Experiments are conducted to test the developed system. The documents are written in MS-Paint programme using the Kannada font and then saved as bit map image files. These document images are fed to the developed system for recognition. The results of the developed system are obtained in two ways. The first method is graphical display of the equivalent font and second method is display of an ASCII file. The results for a sample documents shown in figure 5 are shown in figure 6.

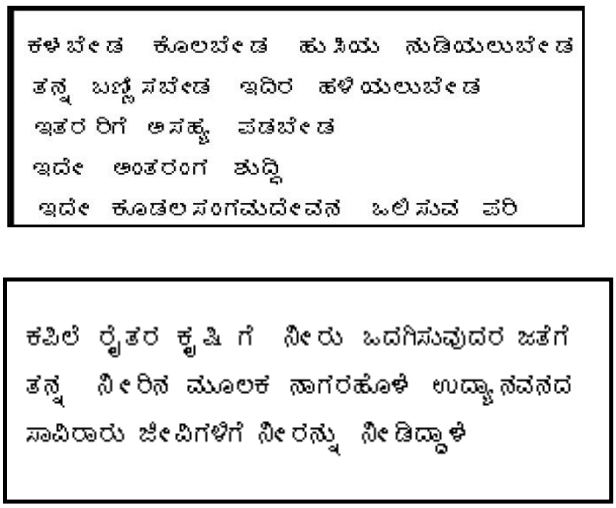


Figure 5: A Sample Kannada document images

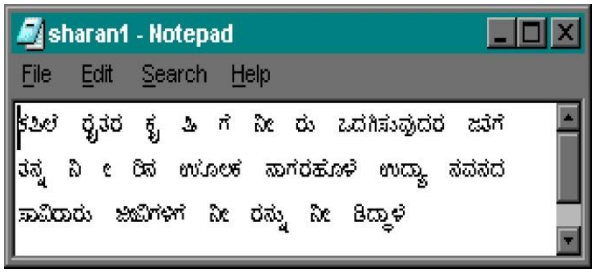


Figure 6: Results of Kannada OCR system (in ASCII format) for the sample s shown in figure 5

4.0 ENGLISH OCR SYSTEM

English OCR system specifically to include in the Bilingual OCR system has been developed. Since all the bilingual documents considered for experimentation are taken for single font, single font character recognizer serves the purpose. To develop a single font English OCR system, first the following document (shown in figure 7) is considered. In this document all the possible English capital letters, English small letters (of Courier New font) and full stop are present. If the developed English OCR system recognizes these characters in the input bilingual document, that may be sufficient presently to test the bilingual OCR system The diagrammatic representation of English OCR is depicted in figure 8.

Figure 7 : Sample English alphabet used in English OCR system

From the document shown in figure 7, all the characters are segmented and normalized to fixed size. Using these normalized characters, the probabilistic neural network based recognizer is designed. For this purpose bar mask encoder type of feature extractor is used [10]. It extracts the 50 features and provides to the neural network based system. The trained network is ready for recognizing unknown characters. All the recognized characters are given tag of ASCII codes and are written into an ASCII file. To reconstruct the sentences, the information of the input document is utilized.

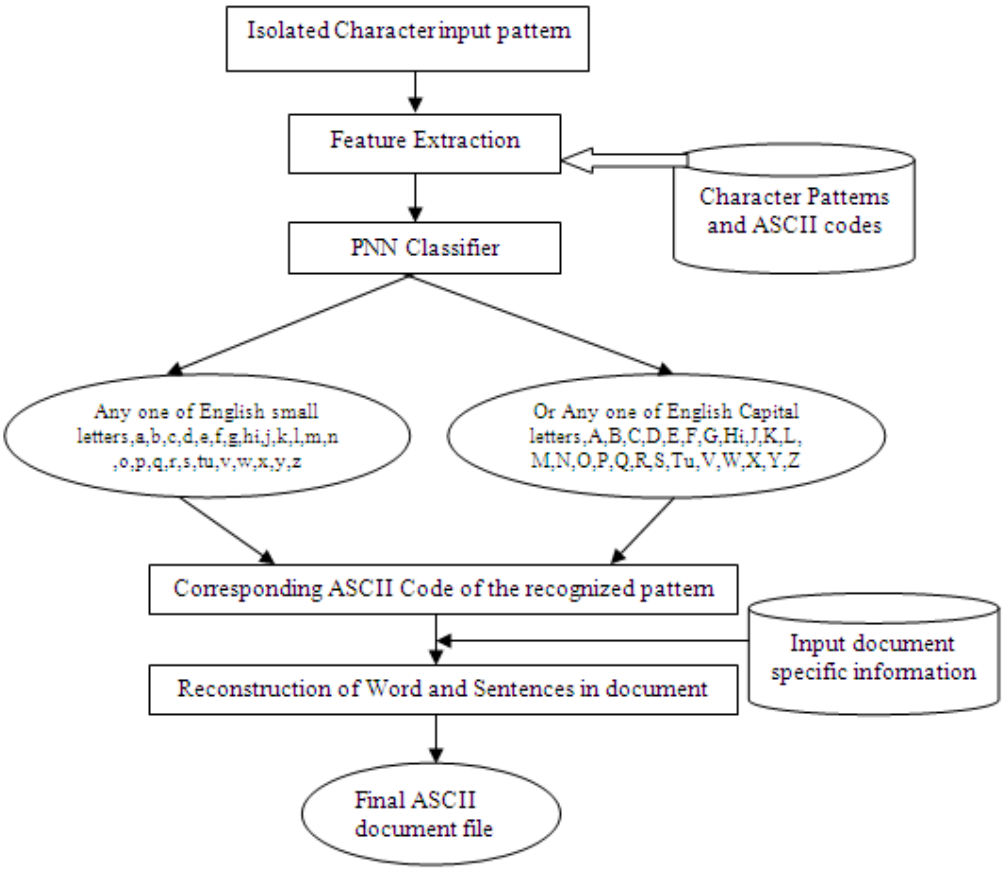


Figure 8: Specifically developed English OCR system

The algorithms used in the construction of above OCR system are as follows.

Algorithm 2: English OCR system.
1. Accept the document image (written in specific font Courier New, considered for English OCR system).
2. Segment the above document image into lines, words and characters , by segmentation procedures which are discussed in [10]. They return words cell array which contains all the segment ed words.
3. Make i=1
4. Segment i^{th} word into characters. Let total number of characters be total_char.
5. Make j=1
6. Present j^{th} character to the feature extractor. It extracts 50 features.
7. Pass the above features to the trained probabilistic neural classifier.
8. PNN classifier recognizes the character and also associated ASCII Code for that character.
9. Write the above ASCII code in the previously opened file 'Engl1.txt' .
10. Make j=i+1.
11. If j is less than or equal to total number of characters (total_char) in the word go to step 6 else go to next step.
12. Write a blank space in the opened file 'Eng1.txt', indicating end of word.
13. i=i+1
14. If i is less than or equal to total number of words go to step 4 else go to step 15.
15. Close the 'Eng1.txt' file.

The above algorithm should be run first time with document image shown in figure 7. After the PNN classifier has been trained on these alphabets, the trained classifier can be used on unknown English document images.

4.1 PNN Classifier

The probabilistic neural network is a two-layered structure. The first layer is a radial basis layer and the second is a competitive layer. The first layer computes the distances from the input vector to the training input vectors and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output, a vector of probabilities. The maximum of these probabilities is considered and the class for which it belongs is selected. The inputs to the radial basis layer are the outputs obtained from the feature extractor module. In the reported experiments, this is a vector of size 50. This layer consists of radial basis neurons equal to the number of training patterns. The weights for this layer are set to the transpose of the matrix formed from the total number of training pairs. The net input to the radial basis neurons is the vector distance between its weight vector w and the input vector p, multiplied by bias b. The output of a radial basis neuron is given by the function,

$$Y = \exp(-n^2) \qquad - (1)$$

where $n = w - p .b$ and denotes Euclidean distance. Each bias in the first layer is set to the square root $(-\log(0.5))/\text{spread}$ or $0.8326/\text{spread}$. A larger spread leads to a large area around the input vector, where the radial basis neuron with the weight vector closest to the input has a much larger output than other neurons. The network tends to respond with the target vector associated with the nearest design vector. In our experiments we use trial and error method to set the spread.

5. EXPERIMENTS AND DISCUSSIONS

Two experiments are conducted. The first one is to test the English OCR system. The second one is to test the Bilingual OCR system. To test the English OCR system documents are written in MS Paint using Courier New font of size 11. Thus created English document images (in '.bmp' format) are fed to the English OCR system. English OCR system produces the ASCII text file as an output. Sample input document image and obtained result output ASCII file are shown in figure 9.

The second experiment is conducted to test Bilingual OCR system. To test the system, bilingual document images are created as follows. Separate English and Kannada document images are created using the corresponding fonts in MS Paint package and saved in '.bmp', image format. Then manually individual lines from English document image and Kannada document image are cut and pasted in the new image file.
Thus the bilingual documents are obtained. Such bilingual documents are fed to Bilingual OCR system. Sample input bilingual document image and obtained two ASCII file are shown in figure 10.a , 10b.

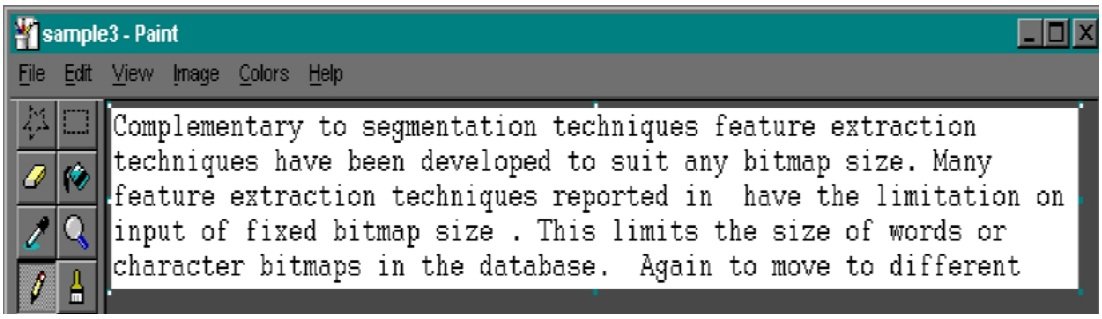


Figure 9 a: Input document image

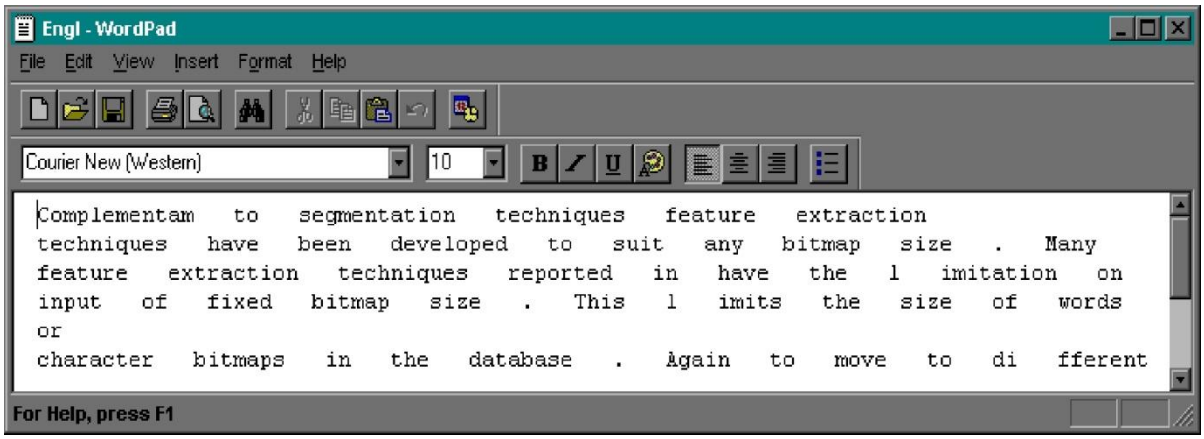


Figure 9b: (i) Sample English input document image (ii) the corresponding output file of English OCR system.

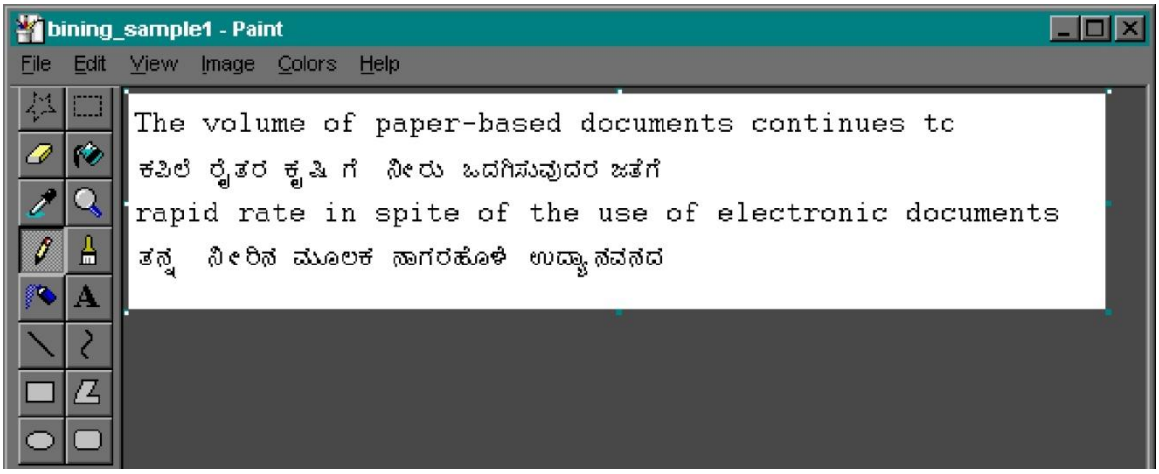
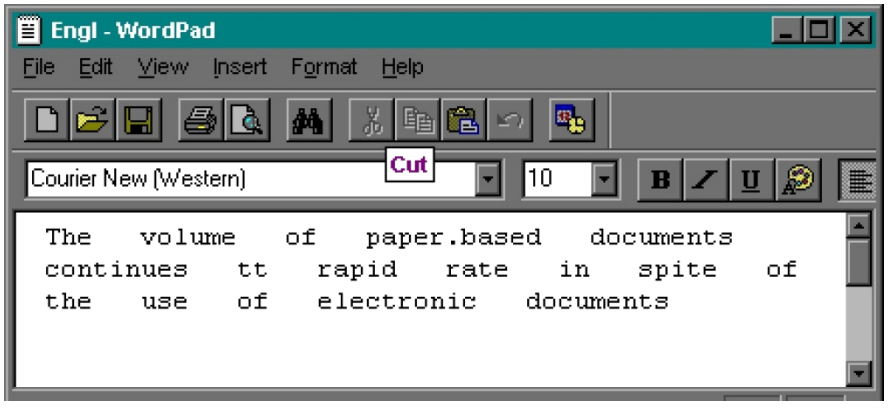


Figure10a:Sample input bilingual document image

(i)



(ii)

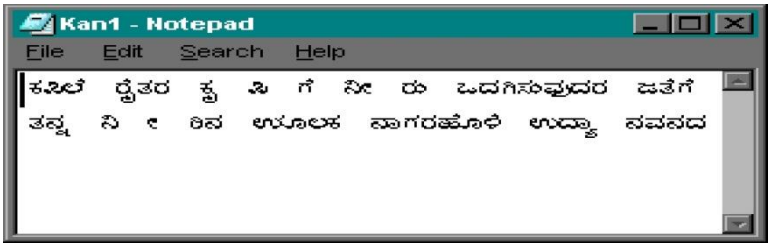


Figure 10b : Results obtained for the sample shown in figure 10a. (i)shows English text file (ii) shows Kannada text file

The above figures show that the results obtained are very good. However some of the limitations and problems are found. The first limitation is the complete system works for only one font of English and one font of Kannada. The second one is about the errors in the system. The errors of the segmentation, word script identification are very important. If a single word is labeled a wrong class, all the characters in that word will produce erroreneous outputs. To avoid such errors one more subsystem called 'character script class confirmation system' could be added to the present system. Similarly if line segmentation produces errors all the words and characters in that line produce errors in output files. All these things should he handled carefully.

6. CONCLUSIONS

This paper presents a developed Kannada OCR system and English OCR system specifically built to embed into the Bilingual OCR system for English and Kannada language scripts. The Bilingual OCR system has been built. The system includes segmentation modules, feature extraction modules, script identification modules, English OCR system and Kannada OCR system. Such experimentation with English and Kannada scripts is first of its kind, to the best of our knowledge. The results obtained are very good and proved that the approach followed is very effective.

7. ACKNOWLEDGMENTS

Author would like to acknowledge the IJCA reviewers for valuable comments. Author also would like to acknowledge Mrs.Ashvini Patil for the help in documentation of this work.

8. REFERENCES

[1] A. L. Spitz, "Determination of the Script and Language content of Document Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 3 , pp. no. 235 - 245, March 1997.

[2] T. N. Tan, "Rotation Invariant Texture Features and their use in Automatic script Identification", *IEEE Transactions on PAMI*, Vol.20, No.7, pp. no. 751 - 756, July 1998.

[3] J. Hochberg, P. Kelly, T. Thomas and Lila Keens, "Automatic Script Identification from Document Images using Cluster based Templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No.2, pp. no. 176 - 181, Feb 1997.

[4] B. B. Chaudhuri and U. Pal, "Automatic separation of machine printed and handwritten text lines", *5th Ineternational Conference on Document Analysis and and Recognition*, Vol.1, pp. no. 645 - 648, 1999.

[5] U.Pal and B.B. Chauduri, "Automatic seperation of words in multi-lingual multi- script Indian documents", *4th International Conference on Document and Recognition*, Vol.2, pp. no.576 - 579, 1997.

[6] Sanghamitra Mohanty "A Novel Approach for Bilingual (English-Oriya) Script Identification and Recognition in a printed document", *International Journal of Image Processing*, Volume 4, Issue 2, 2010.

[7] C.V.Jawahar, Pavan Kumar, S.S.Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents and its applications", Proceedings of 7th *International Conference on Document Analysis and Recognition (ICDAR)*-2003.

[8] Sanjeev Kunte and Sudhakar Samuel, "A Bilingual Machine-Interface OCR for printed Kannada and English Text Employing Wavelet Features", *10th IEEE International Conference on Information Technology,2007.*

[9] Padma and Vijaya, "Script Identification from Trilingual documents using profile based features", *International Journal of Computer Science and Applications,*

Volume 7, No.4, pp. 16-33, 2010.

[10] S.Basavaraj Patil and N V Subbareddy " Neural Network based System for Script Identification in Indian Documents", *Sadhana, Special Issue on Indian Language Document Processing",*Vol.27,part-1,2002.

[11] C. C. Tappert, C. Y. Suen and T. Wakahara, "The state of art in on-line handwriting recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence* No.12, pp. no.787 - 808, 1990.

[12] T. Y. Young and Fu, *Handbook of Pattern Recognition and Image processing*, Academic Press, New York, 1986.

[13] J. R. Ullman, *Pattern Recognition techniques*, Butterworths, London, 1973.

[14] R. H. Cheng, C. W. Lee and Z. Chen, "Pre-classification of handwritten Chinese characters based on basic stroke substructures", *Pattern Recognition Letters* Vo.16, pp. no. 1023 - 1032, 1995.

[15] C. C. Han, Tseng, Y. L. Fan, and K.C. Wang, "Coarse classification of Chinese characters via stroke clustering method", *Pattern Recognition Letters*, Vol.16, pp. no.1079 - 1089, 1995 .

[16] K. K. Biswas and S. Chatterjee, "Feature based recognition of Hindi characters", *International conference on Pattern recognition, Image processing and Computer Vision*, Kharagpur, pp. no.182 - 187, 1995.

[17] C. Y. Suen, J. Guo and Z. C. Li, "Analysis and recognition of alphanumeric handprints by parts", *IEEE Transactions on Systems, Man and Cybernetics*, Vol.24, pp. no. 614 - 631, 1994.

[18] Sameer Antani and Lalitha Agnihotri, "Gujarati character recognition", *International Conference on Document Analysis and Recognition*, pp. no.418-421, 1999.

[19] P. Nagabhushan, Radhika and M. 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 Letters*, Vol.20, pp. no.1467 - 1457, 1999.

[20] T. V. Ashwin and P. S. Sastry , "A font and size independent OCR system for printed Kannada documents using support vector machines", *SADHANA*, Vol.27, Part 1, pp. no.35 - 58 February 2002.