Feature Based Kannada Character Classification Method of Kannada Character Recognition

Sandhya.N, Krishnan.R, D.R.Ramesh Babu

Abstract- In Kannada language there are more than 510 character sets to be recognized including the vowels, consonants, consonant modified by vowel and consonant conjuncts the classification of characters becomes difficult. In this paper we present a new classification method of Kannada characters which can be used as a preliminary step for recognition. An analysis of Kannada characters was done and syntactic features were identified. Firstly the basic features present in the vowels are identified in the exact position of the character and recorded. Then by using a decision tree the characters are classified. The experimental results show the syntactic based method using basic features gives high contribution and reliability for Kannada character classification.

Index words—vowels, consonants, vowel modifiers, consonant modifiers, consonant conjuncts, decision trees

1 Introduction

Kannada script has 49 characters in its alphasyllabary and is phonemic. The number of written symbols, however, is far more than the 49 characters since different characters can be combined to form compound characters (ottaksharas). The Kannada writing system has consonants appearing with an inherent vowel.

The characters are classified into three categories: swaras (vowels), vyanjanas (consonants) and Yogavaahakas (part vowel, part consonants). The Fig 1 shows the Kannada alphabets

The partitioning of Kannada characters for recognition is difficult since there are no gaps for the consonants appearing with an inherent vowel. In Fig 2 the consonant (ka) with the vowel (uu) is a consonant modifier (kuu). Hence a syntactic approach will give better results. The features of the modifier in the character should be identified.

Here we have successfully identified the 15 vowels of Kannada alphabets using the feature specific for each vowel.

In this paper we describe a feature based method to recognize the printed Kannada characters. In Section II we explain related work. Section III explains the feature based recognition approach Section IV we explain the Pre-classification Algorithm for Kannada vowel/consonants from a printed Kannada document. Section V describes the possible algorithm for consonants, vowel modifiers and conjuncts. This is followed by Section VI conclusion and future work.

2 Related Work

Refer to the paper [1] for recognizing lines of unconstrained handwritten text. The difficulty of segmenting cursive or overlapping characters, combined with the need to exploit surrounding context, has led to low recognition rates for even the best current recognizers.

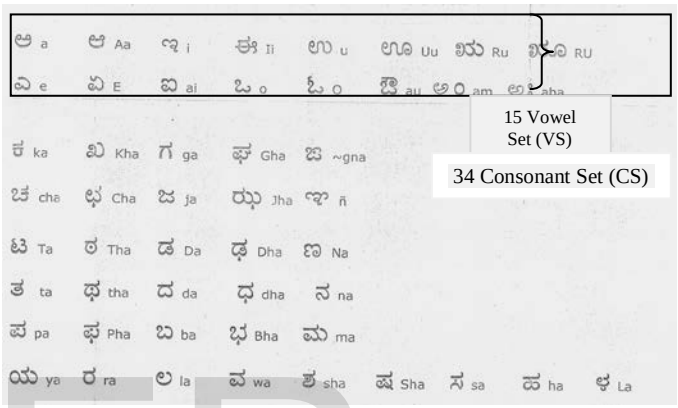

Fig1. Kannada alphasyllabary and its phonemic

(ka) (uu) (kuu)

Fig 2. Consonant Appearing with an Inherent Vowel

G. Vamvakas et al. [2] proposed a Feature Extraction and Classification Methodology for the Recognition of Historical Documents. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the image as well as on calculation of the centre of masses of each sub-image with sub-pixel accuracy.

G. Pirlo et al. [3] proposed a paper on zoning-based classification. This paper presents a new class of membership functions, which are called fuzzy-membership functions (FMFs), for zoning-based classification.

Salvador Espana Boquera et al. [4] proposes the use of hybrid Hidden Markov Model (HMM)/Artificial Neural Network (ANN) models for recognizing unconstrained offline handwritten texts.

Sang Sung Park et al. [5] constructed an OCR system that saves abstracted characters to DB automatically after extracting only equivalent and necessary characters from a large amount of documents by using BP algorithm that is one of artificial neural network.

Prachi Mukherji et al. [6] proposes Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition

3 Feature Based Recognition Approach

3.1 The Features

The features used to classify the vowels and consonants are given in Table 1 which is small parts of the characters. The decision tree in Fig 4 shows the classification of Kannada vowels/consonants based on the features identified.

TABLE 1
List of Features to Recognize Vowel and Consonant

Vowel/ Consonant Features								

3.2 Structure of Kannada Character

The vowel modifier set represented as VMS and consonant set as CS. As shown in Fig 2 all the CS will be modified by the VS. In Fig 3 we represent the vowel modifier feature in the character along with the consonant conjunct.

f:Ω → {cs x vms x cc, cs, vs}

Where f: Ω is the valid character. cs is character set, cc is the consonant conjunct, vs vowel set and vms is the vowel modifier set [7].

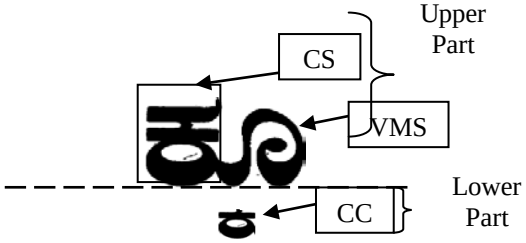

Fig 3. Structure of a Kannada Character

To identify a vowel/consonant the feature set (FS) is used. The CFS and VMS are used for the recognition of consonant modifier.

4 The Process of Classification

The algorithm aims at identifying vowels/consonants in the printed noisy Kannada document. The document has to be preprocessed and binarized. The steps in the algorithm are shown in Fig 5.

Algorithm for Classification of vowels/consonants
1. Pre-processing the scanned document.
2. Removal of noise types such as skew & background noise present in the document.
3. Segmentation of characters in the document.
4. Hypothesize characters based on features and decision tree.
5. Verification of characters hypothesized.

Fig 5. Algorithm for Classification of vowels/consonants

1. Preprocessing

The salt and pepper noise present in the document image has to be removed by using adaptive median filter [11].

2. Removal of noise types

The prominent noise [8] skew and background noise is removed. Skew is removed using base point algorithm 10]. Background noise is removed by using morphological operations dilation and erosion [12].

3. Segmentation of characters

The segmentation algorithm in the paper [9] is used to separate out characters have the following steps:

- Labelling of the connected components
- Measure the image properties needed for each character
- Based on the properties, plot the bounding box.
- From the bounded box, the object segments are extracted from the image.

Fig 6 shows output of the segmented character from a document

4. Hypothesize characters based on the decision tree

- Size of the segmented character is scaled according to standard size of the character.
- Check if a particular feature in the set of feature (FS) is present in the segmented character.
- Depending on the features present and the decision tree shown in Fig 4 we can hypothesize the character.

Fig 7 shows the output of segmented character recognized based on features.

5. Verification of Characters Hypothesized

- The characters hypothesized are verified by scaling and correlation.

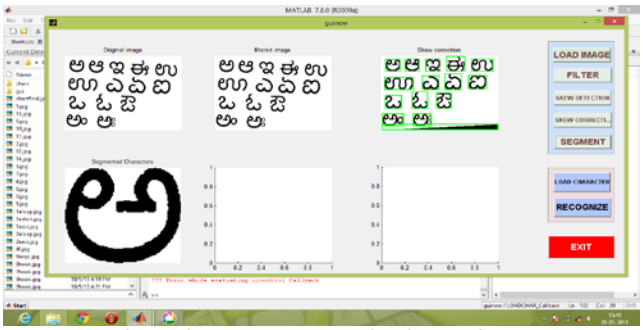

Fig 6. Character segmentation from a document

Fig 7. Segmented Character recognized based on features

5 Experimental Results

The method is implemented using MATLAB 9.0. The experiment was conducted on multiple font sizes and the results were satisfactory.

6 Proposed Algorithm for Consonant Modifier Features

6.1 The Consonant Features

In Fig 4 the consonant with vowels are classified using a decision tree and features. There are only 26 features for the vowels/consonant features.

6.2 The Consonant Modifier Features

The consonant modifier is a set of 476 characters 34(consonants) modified by 14 (vowels). As shown in Table II by using 9 vowel modifier features we can identify 14 vowel modifier as 1st is the consonant itself. Five vowel/consonant features are repeated in vowel modifier features. The entire 476 character set can be hypothesized by 26 features (vowel/consonant) + 9 features (vowel modifier) ie 35 features.

6.3 The Consonant Modified by Vowels

In Table 2 the vowel modifier features are given. These features will be repeated for all the consonants. Some of the features are also repeated from the vowel features. Hence there are only 9 features which are new for the vowel modifier feature.

6.4 The Consonant Conjuncts

There are 34 consonant conjuncts for each consonant are listed in the Table 3. These will be present in the lower part of the character as shown in Fig 3.

TABLE 2
Features for Vowel Modifier Recognition

Consonant	Vowel	Vowel Modifier Feature
ಕ	ಅ	ಕ
ಕಾ	ಆ	ಾ
ಕಿ	ಇ	ಿ
ಕೀ	ಈ	ೀ
ಕು	ಉ	ು
ಕೂ	ಊ	ೂ
ಕೃ	ಋ	ೃ
ಕೆ	ಎ	ೆ
ಕೇ	ಏ	ೇ
ಕೈ	ಐ	ೈ
ಕೊ	ಒ	ೊ
ಕೋ	ಓ	ೋ
ಕೌ	ಔ	ೌ
ಕಂ	ಅಂ	ಂ
ಕಃ	ಅಃ	ಃ

TABLE 3
Consonant Conjuncts

Conjunct No.	Consonant	Conjunct
1	ಕ	ಕ್
2	ಖ	ಖ್
3	ಗ	ಗ್
4	ಘ	ಘ್
5	ಙ	ಙ್
6	ಚ	ಚ್
7	ಛ	ಛ್
8	ಜ	ಜ್
9	ಝ	ಝ್
10	ಞ	ಞ್
11	ಟ	ಟ್
12	ಠ	ಠ್
13	ಡ	ಡ್
14	ಢ	ಢ್
15	ಣ	ಣ್
16	ತ	ತ್

17	ಥ	ಧ
18	ದ	ಂ
19	ಧ	ಃ
20	ನ	಄
21	ಪ	ಅ
22	ಫ	ಆ
23	ಬ	ಇ
24	ಭ	ಈ
25	ಮ	ಉ
26	ಯ	ಊ
27	ರ	ಋ
28	ಲ	ಌ
29	ವ	ಎ
30	ಶ	ಏ
31	ಷ	ಐ
32	ಸ	ಒ
33	ಹ	ಓ
34	ಳ	ಔ

6.5 Algorithm for Consonant Modifiers and Conjuncts

The Kannada characters which are segmented will have four forms such as: vowel, consonant, consonant modified by a vowel and consonant with consonant conjuncts. The algorithm given in Fig 5 hypothesizes vowels/consonants. The algorithm given in Fig 8 is used for complete Kannada character set including consonant modified by a vowel and consonant with consonant conjuncts.

VI Conclusion and Future Work

It is shown that the set of feature of vowels, consonants, vowel modifiers and consonant conjuncts (13, 13, 9 and 35) total 70 features can be used to recognize 574 printed Kannada alphasyllabary using a scaling decision tree and feature based method. The vowels are recognized for one font style. The features can be modified with minor shearing for other font styles. In future we plan to use machine learning techniques for the features identified in this paper.

References

[1] Alex Graves, Marcus Liwicki, Santiago Fernandez, Roman Bertolami, Horst Bunke, and Jurgen Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition",IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 5, May 2009

[2] G. Vamvaka, S B. Gatos and S. J. Perantonis, "A Novel Feature Extraction and Classification Methodology for the Recognition of Historical Documents", 10th International Conference on Document Analysis and Recognition, 2009.

[3] G. Pirlo and D. Impedovo, "Fuzzy-Zoning-Based Classification for Handwritten Characters", IEEE transactions on fuzzy systems, vol. 19, no. 4, August 2011

[4] Salvador Espana-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez, Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models, IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 4, April 2011

[5] Sang Sung Park, Won Gyo Jung, Young Geun Shin, Dong-Sik Jang, "Optical Character Recognition System Using BP Algorithm" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008

[6] Prachi Mukherji and Priti P. Rege, "Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition", Journal of Pattern Recognition Research4 (2009) 52-68

[7] Mingji Piao, Sejin Kim, Rongyi Cui, "Structure Based Modern Korean Character Set Partitioning and Pre-Classification Method of Korean Character Recognition", International Conference on computer Science and Information Processing(CSIP), 2012

[8] Sandhya.N, Krishnan's, D.R.Ramesh Babu, "A language independent characterization of document image noises in historical scripts, International Journal of Computer Applications, (0975 – 8887), Volume 50-No9, 2012

[9] M Swamy Das, Dr. CRK Reddy, Dr. A Govardhan, G. Saikrishna, International Journal of Engineering Science and Technology, Vol. 2(11), 2010, 6606-6610.

[10] Sharda, Varun and Kishan, Avinash C (2009) Skew Detection and Correction in Scanned Document Images. BTech thesis.

[11] http://en.wikipedia.org/wiki/Adaptive_filter.

[12] B Gangamma and Srikanta Murthy K. Article: Enhancement of Degraded Historical Kannada Documents. International Journal of Computer Applications 29(11):1-6, September 2011. Published by Foundation of Computer Science, New York, USA

Algorithm including consonant modifiers and conjuncts

1 A vector array is maintained for each character segmented.
2 Each character segmented has a size and if the height is exceeding then the conjunct will be hypothesized.
3 Each vector has 3 components vowel/consonant, vowel modifier and conjuncts.
4 If a character is hypothesized as a vowel/consonant using Fig 4 then the first bit is given the number represented for the vowel/consonant.
5 Then a modifier is hypothesized using Table III. If a modifier is found the second bit is given the number represented for the vowel modifier.
6 The lower Part of the character is considered. If the conjunct is found as per the features given in table III then the conjunct number is given in the third component.
7 Based on the number present in the 3 components the

Fig 8 Algorithm including consonant modifiers and conjuncts

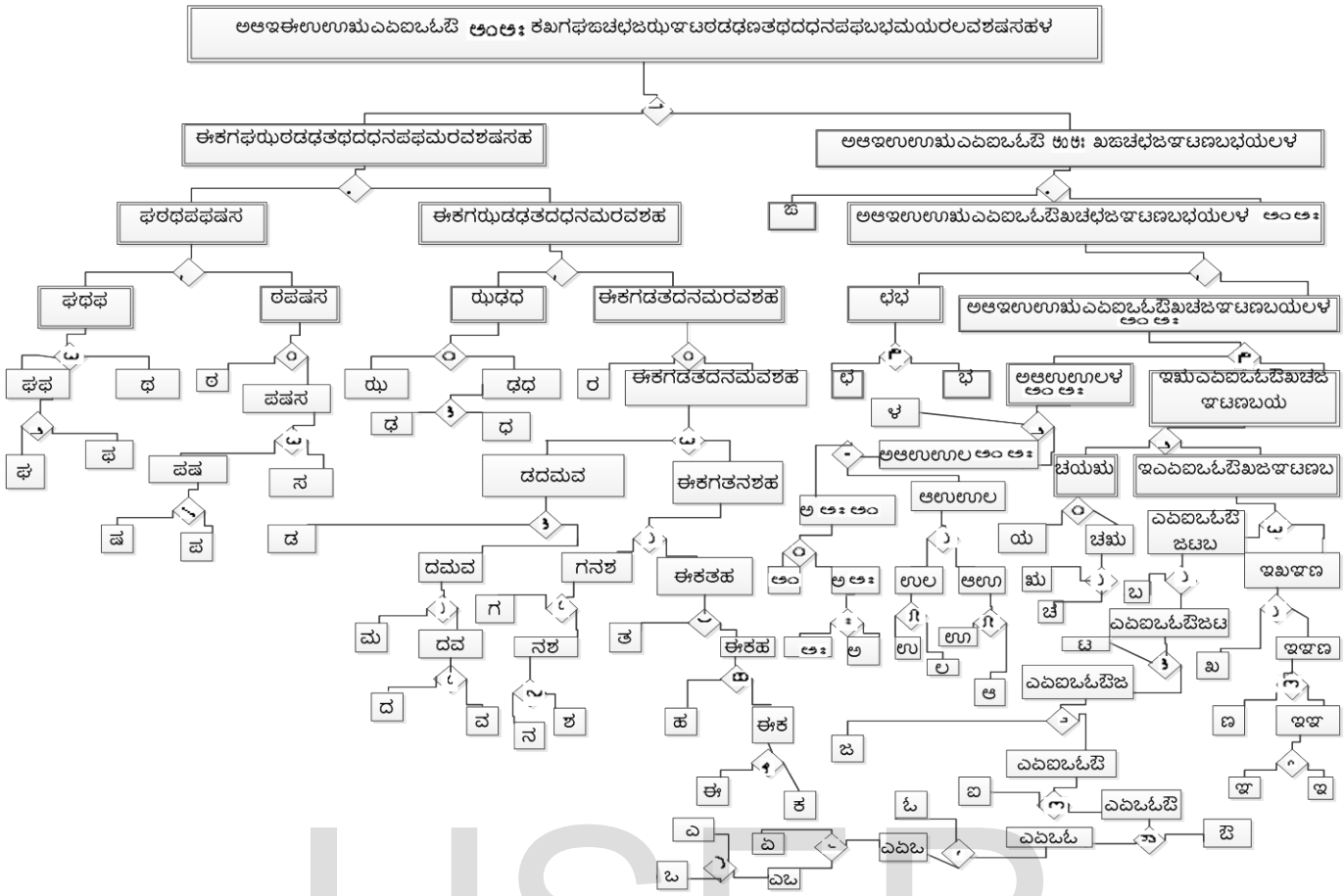

Fig 4. Decision Tree based on the features for vowel/consonant recognition