International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 333-335

CLASSIFICATION OF KANNADA FONT USING HILBERT-HUANG TRANSFORM METHOD

K. Karunakara[1] & B. P Mallikarjunaswamy[2]

Character recognition by machine opens many challenges in the development of Optical Character Recognition system. Many methods have been tried for recognition of characters of Indian languages. This paper discusses a novel method of classifying Kannada character using Hilbert Transform and Empirical Mode Decomposition method. Text images are normalized, stroke features are extracted, then it is decomposed using Empirical Mode Decomposition. With this method any complicated dataset can be decomposed into finite number of Intrinsic Mode Functions which admit well behaved Hilbert transforms. The first two of these intrinsic functions are used to compute high frequency energies, which is the average of two intrinsic mode functions. It is then combined with the averages of five residues and a feature vector is formed. This can be used to identify embedded structures. The result so obtained is compared with the sample database and most nearest matching is used to recognize the fonts. The performance of this method is tested on a variety of documents.

Keywords: Hilbert-Haung Transform(HHT) Emperical Mode Decomposition (EMD), Intrensic Mode Function(IMF), Optical Character Recognition (OCR)

1. Introduction

Automatic document analysis and processing has gained significant importance as there is large amount of data and information available since the time immemorial. These documents need to be computerized. Of late, optical character recognition has been greatly developed due to influence of Internet and multimedia tools. Some tools are now commercially available for printed character recognition. Extensive studies have been carried out on recognition of characters in languages like English, Chinese, Arabic etc…[Li et.al, 2004] [Jung et.al,1999][Yong et.al,2001][Zhihua yang et.al,2006][Zhouchen et.al,2007]. In India, a great deal of information and knowledge is available in more than two thousand Indian languages. Some major works have been carried out in Sanskrit, Bengali, Tamil etc…[K P Pramod et.al,2004][U Pal et.al, 1997][Venna Bansali et.al, 1999] to develop tools for automatic character recognition. Many attempts are also made in Kannada and Telugu languages [R. Snjeev kunte et al, 2004] [Ashwini T V et al, 2004]. However, much work has not been carried out to extract attributes of a character such as font- type, the size and so on. It is very essential to recognize these attributes as it improves the process of character recognition.

Kannada language is one of the oldest languages in India and more than 60 million people use this language for their day to day activities. Basic alphabet set of Kannada language consists of 15 vowels and 34 consonants. One of the complexities of Kannada language is consonant vowels and conjunct consonants.

This paper discusses a novel method to recognize Kannada fonts and its style used in a document. Recognition of these attributes of the fonts will help us in producing proper re-editable text. This also helps to develop sophisticated OCR systems which identify the fonts with greater accuracy. Fonts are normally classified based on (a) local attributes such as serifness, boldness (b) on local or global typographical features or (c) texture analysis.

To develop an Automatic character recognition system demands solutions to the many challenging situations such as font type, style, texture etc. Many methods have been tried for recognition of these attributes for Indian languages. Initially text images are normalized, it is then decomposed using Empirical Mode Decomposition to find finite number of intrinsic mode functions which admit well behaved Hilbert transforms. The first two of these intrinsic functions are of the highest frequencies, and are used to produce high frequency energies, then it is added with the averages of residues to find feature vector. This vector can be used to for sharp identification of embedded structures.

The main objective of this paper is to:

- Find the attributes of Kannada characters using Hilbert transform and Empirical Mode Decomposition method.

[1]Research Scholar, Kuvempu University, Gouthama Research Center, Sri Siddhartha Institute of Technology, Tumkur

[2]Professor, Deptt of Comp. Sc., Sri Siddhartha Institute of Technology, Tumkur

Email: karunakara_k@rediffmail.com

2. CHARACTER FEATURE EXTRACTION

Given many text images, the sizes of the characters in the text images normally varies so also spaces between the characters, and or between lines of text. These uneven format results in poor recognition of characters. Therefore normalization is done before character recognition. This can be done by normalizing

- Spaces between each pair of words and text lines.
- Size of the character.
- Fill the unexpected blank spaces with text padding which can be extracted randomly from the part of non-blank spaces of the text block.

Each font in Kannada has its own features and obviously has different distribution of energies. The feature of the character like Na and Sa or Ma and Va which appears to be same but variations in the energy levels. The high frequency energies of the two set of fonts extracted from the intrinsic mode function differs in energy levels. Since every characters have some geometric structures. Using EMD its six IMFs and the residues are computed and feature vector is calculated.

3. EMPIRICAL MODE DECOMPOSITION

Hilbert_Haung Transform (HHT) is a novel analysis method for non-linear and non-stationary data [Haung et al, 1998]. Empirical mode decomposition is the key part of the HHT method, with which any complicated data set can be decomposed into finite number of intrinsic mode functions (IMFs) which admit well behaved Hilbert transforms. An IMF is defined as a function which satisfies the following conditions:

- The number of extrema and the number of zero crossing must be either equal or differ at most by one.
- At each point of the data, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

With Hilbert transform, the IMFs yield instantaneous frequencies as function of time, that gives sharp identifications of embedded structures. The final presentation of the results is a time-frequency-energy distributions, designated as the Hilbert spectrum. EMD depending on the signal brings not only high decomposition efficiency but also sharp frequency and time localization. Signal analysis based on the HHT is widely used in different areas for its physical significance. Empirical Mode Decomposition [Haung et.al,1998] decomposes a complex signal into Intrinsic Mode Functions.

4. IMPLEMENTATION

Features of the character block are extracted from the text images using following algorithms.

Algorithm - 1

Let X(t) be a signal.

Step 1: Initialize :Let r0 (t)=X(t) and i=1;

Step 2: Extract the ith IMF as follows:

(a) Initialization: Let h0 (t)=ri – 1 (t) and j = 1;

(b) Extract the local minima and maxima of hj – 1(t);

(c) Interpolate the local maxima and the local minima by cubic splines t form uj – 1 (t) and lj – 1(t) as the upper and lower envelops of h j – 1(t), respectively;

(d) Calculate mj – 1(t) = [u j – 1(t) + l j – 1 (t)]/2 as an approximation of the local mean of

hj – 1(t) at t;

(e) Let hj(t) = hj – 1(t) – mj – 1(t);

(f) If the stopping criteria is satisfied, i.e hj(t) is an IMF, set imfi(t) = hj(t); else go to (b) with j = j + 1

Step 3: Let ri(t) = ri – 1 – imfi(t);

Step4: If ri(t) still has at least two extrema, go to.

Step 2 with i = i + 1; Else the decomposition is finished and ri(t) is the residue.

The above algorithm is used to extract the features of each font style. For each font style its IMFs and residues are computed and stored. This is used for identification of characters. Because of similarity between many fonts the residues are also utilized to differentiate between the fonts along with first two IMFs of character feature.

Algorithm-2: For each character block of gray scale 256, the feature vector is computed. This feature vector is the basis for identification of the font type and its style in the text.

Step1: Conduct the preprocessing of text and perform normalization.

Step 2: Calculate the energy functions of the each character. Calculate high frequency energy ex and residue energy rx.

Step 3: Let V = [ex, rx], which is the feature vector for the character.

5. RECOGNITION OF FONT

Once the feature vector is obtained, the next step is to identify the font type of the character. This is implemented by designing a Weighted Euclidean distance classifier which works on the basis of statistical decision theory, which in

turn depends upon the statistical distribution of large number of samples. For the different character blocks, mean vector V_{mean} and its variance V_{var} is computed based on the training samples. Similarly, for unknown font, feature vector is computed and the result compared and character features are identified.

6. EXPERIMENT RESULTS

Experiment have been conducted for four popular font cases in kannada. Gray scale images are taken for computation as shown in fig 1. Image size used is 128 x 100 pixels.

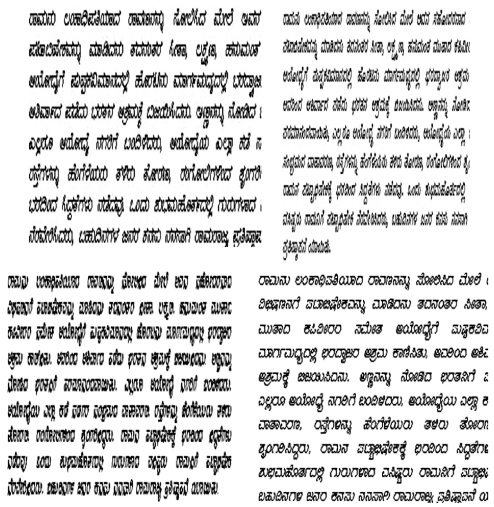

Fig. 1: Few Samples of the Text Blocks used for the Experiment Purpose. From Top to Bottom (Clockwise) Nudiakshara1, Baraha Regular, Nudiakshara3, Nudiakshara2.

As some fonts and styles have similar shapes and their IMFs are more similar and are susceptible for confusion. The result is computed based on 30 samples for each font type. Font recognition rate is tabulated in Table 1

Table 1
Font Recognition Rate

	Nudi1	Baraha	Nudi2	Nudi3
Recognition rate	93.3	89.4	92.2	87.6

7. CONCLUSION

This paper describes a method to recognize features of Kannada character based on Empirical Mode Decomposition. Unique features of each font are obtained using EMD. High frequency energies and residual energies act as a base for character recognition. The advantageous of this method is better recognition of fonts. The sizes of each testing sample and training sample may be different. HHT algorithm is used here to identify features of Kannada Characters. The result obtained using HHT transform is encouraging and there is a scope for further improvement of the result. This concept can be further extended to recognize the hand written Kannada characters.

REFERENCES

[1] Ashwini T. V l "Font and Size Independent OCR Printed Kannada Documents using SVM Classifier"-2000.

[2] Huang, N E Shen, Z, Long, S.R et.al,1998. The Emperical Mode Decomposition and the Hilbert Sprectrum for Nonlinear and Non Stationary Time Series Analysis. Pro. Roy. Soc. Lond. A 454, 903-995.

[3] Jung, Minchul, Shin, YongChul, Srihari, S N., 1999. "Multifont Classification using Typographicl Attributes". IEEE Computer Society Press, Bangalore, India, pp353-356.

[4] K. P Pramod, Raghavendra, R.Gopal, B P Vijaya Kumar, S M Dilip Kumar "Enhancement of Character Recognition through Intelligent Competent Technique".

[5] Li. Chen, Ding, Xiaoqing, 2004. Font Recognition of Single Chinese Character based on Wavelet Feature. Acta Electron. Sin32(2). 177-180.

[6] R. Sanjeev Kunte, R D Sudhakara Samuel, R Srinivasa Rao Kunte, " Segmentation of Characters from Machine Printed Kannada Characters".

[7] R. Srinivasa Rao Kunte et al "On-line Recognition of Handwritten Kannada Characters Employing Wavelet Features", Conference on Communication, Control and Signal Processing.

[8] U Pal, B B Choudary, Printed Devanagari Script OCR System, Vivek, 10, Pp 12-24, 1997.

[9] Veena Bansali, R M K Sinha "Segmentation of Touching Character in Devanagari, Computer Vision Graphics and Image Processing, Recent Advances", Viva Books pvt ltd., Pp 371-401,1999.

[10] Yong, Zhu,Tan,Tieniu, Wang, Yunhong, 2001. "Font Recognition based on Global Texture Analysis". IEEE Trans. Pattern Analysis and Machine Intelligence, 23(10), 1192-1200.

[11] Zhouchen Lin, Liang Wan, 2007. Style-preserving English Handwriting Synthesis, Pattern Recogniton 2097-2109.

[12] Zhihua Yang, Lihua Yang, Dongxu Qi, Ching Y Suen, 2006. An Empericl Mode Decomposition based Recognition method for Chinese Fonts and Styles.