Pattern Recognition Letters 84 (2016) 29–34

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Modeling of palm leaf character recognition system using transform based techniques☆

Narahari Sastry Panyam[a,*], Vijaya Lakshmi T.R.[b], RamaKrishnan Krishnan[c], Koteswara Rao N.V.[a]

[a] CBIT, Hyderabad 500075, India
[b] MGIT, Hyderabad 500075, India
[c] Department of Space, ISRO, Coimbatore 641046, India

ARTICLE INFO

Article history:
Received 2 November 2015
Available online 2 August 2016

Keywords:
Palm leaf character recognition (PLCR)
3D feature
Discrete wavelet transform (DWT)
Discrete cosine transform (DCT)
Fast Fourier transform (FFT)

ABSTRACT

Optical character recognition (OCR) has been a well-known area of research for last five decades. This is an important application of pattern recognition in image processing. Automatic mail sorting generated interest in the handwritten character recognition (HCR) over a period of time. Palm leaf manuscripts which are very fragile and susceptible to damage caused by insects, contain huge amount of information relating to music, astrology, astronomy etc. Hence it becomes necessary for these manuscripts digitized and stored. These palm leaf manuscripts created interest for the young generation researchers since the last decade. This work exploits a special 3D feature (depth of indentation) which is proportional to the pressure applied by the scriber at that point. This 3D feature is obtained at each of the pixel point of a Telugu palm leaf character. In this work two dimensional Discrete wavelet transform (2-D DWT), two dimensional fast Fourier transform (2-D FFT) and two dimensional discrete cosine transform (2-D DCT) are used for feature extraction. The 3D feature along with the proposed two level transform based technique helps to obtain better recognition accuracy. The best recognition accuracy obtained in this model is 96.4%.

1. Introduction

Handwritten character recognition is a computer enabled process that facilitates interpreting and reading legible handwritten inputs provided on old manuscripts like palm leaves, stone inscriptions, etc. [1]. This involves the classification of handwritten characters into appropriate and recognized class [2]. It involves the minute reading of features which define each character. Optical scanning can read the image of a written text from a piece of paper. The most important principle involved in the recognition of handwriting is optical character recognition (OCR). This may be augmented with the help of formatting which will entail correct segmentation into characters and recognize to read the plausible words.

Since the advent of digital computers it has become difficult to conduct research on machine simulation for human functions. It is unfortunate that the digital word has yet to devise a powerful computer that can substitute for the functions of optical senses of the humans. In recent times research focused on machine simulation that will mimic reading as the human being does. However, the recognition accuracy is very low for handwritten characters [3].

The process of character recognition can be structured into two broad categories Offline and Online character recognitions [4]. The Offline character recognition process involves the scanning of the document, digitalization of the data and storing it in a computer for the process of character recognition. Contrary to this, characters are first extracted simultaneously while they are created for online character recognition systems. Obviously, Offline character recognition is not inhibited by external factors like speed and stroke in the movement of writing. The Online system of recognition will be impacted by these features.

Pattern recognition techniques till recent times resorted to template or feature based approach [4]. In the template based recognition, the desired pattern for recognition is superposed upon a template pattern that is considered ideal. The extent of correlation between the pattern for recognition and ideal pattern help to decide the degree of match. This was the approach in the earlier OCR systems. In recent times, the feature based approach is clubbed with the template approach for better results [4]. For example, the

☆ This paper has been recommended for acceptance by Prof. L. Heutte.

* Corresponding author.
E-mail address: ananditahari@gmail.com (N.S. Panyam).

http://dx.doi.org/10.1016/j.patrec.2016.07.020
0167-8655/

Bangla OCR system incorporated feature based approach for basic characters and template based approach for compound characters [4].

The more refined feature based approach extracts special and specific properties from the test pattern and implies them in a more sophisticated model. This can be two types, [4,5] spatial domain and transform domain. The spatial domain approach depends on deriving features directly from the pixel image representation individual to a pattern. The transform domain is more elaborate. It involves the transformation of image into another space using any one of the transform techniques. Today a modern set of techniques are available to derive any feature from any pattern to a degree of clarity. Nevertheless a prominent graph based work for popular Indian language script is not reported in the literature [4]. Though Sinha and Mahabala [4,5] employed embedded picture language for Devanagari (a script similar to Hindi) OCR studies.

India is a diverse nation and rich in literature. As of today there are 33 languages and 2000 dialects of which 22 are recognized under the constitution. Thus, it is a multilingual nation – Assamese, Hindi, Kannada, Bengali, Konkani, Malayalam, Manipuri, Tamil, Telugu, Punjabi, Sanskrit and Urdu [6] are the languages officially recognized. Most of these use 12 scripts in various forms. Hindi, Konkani and Sanskrit languages use Devanagari script, Bangla, Assamese and Manipuri use Bengali script and Punjabi language use the Gurumukhi script. A common trait of these languages is that there is no upper case and lower case letters and they are all descendants of the Brahmi script [7,8]. All of them are phonetic i.e., the symbols used in the script relate to a phoneme or the sound of the language. Another prominent feature of these languages is the vowel is not explicitly written when it follows a consonant in a word, which leads to writing a composite character. Consonants sometimes combine with the vowels or another consonant to form a complex character [7,8]. The most common features of these languages is that they are scripted from left to right except for Urdu which is written from right to left [8]. Chaudhuri and co-workers reported that there is no standard database for Indian languages and the research contributed in this area is low [3].

The ancient literature of Southeast Asia including India is preserved as written on palm leaves [9]. This was the main writing material for many centuries. Popular literature, scientific treaties and history are found since fifth century B.C. [9,10] on palm leaves. One of the best preserved oldest existing documents is recognized to be recorded in the second century A.D. Palm leaves were used as writing material to record art, medicine, astronomy, etc., and were preserved and passed through generations [9,11–15].

Telugu script which is an offshoot of Brahmi script has complex structural characteristics, which are difficult for character recognition [16]. It has 16 vowels and 36 consonants [14]. The challenges for Telugu character recognition are as follows.

- Compared to English, Indian languages have more number of basic and composite characters.
- With smaller number of users, languages like Telugu have not attracted equivalent efforts for character recognition.
- Grammatical operation of sandhi, which literally means “junction” or “union”.

2. Related work

Sastry et al. [11,12] have reported that Principal Component Analysis (PCA) using nearest neighborhood classifier (NNC) classification method got very low recognition accuracy in all the three planes “XY”, “YZ” and “XZ” ranging between 37% and 40%. They further published that using 2-D correlation features and NNC classification method the recognition accuracy obtained was 90% in “YZ” plane of projection. The confusing characters like “Ya”, “Ma”, “Na”, and “Va” in “XY” plane, could give higher recognition accuracy in “YZ” plane of projection using their algorithm.

Pal and Chaudhuri [4] devised the completely capable system for the first time to perform OCR of the printed Bangla scripts. Pre-processing in this system starts with skew correction, noise reduction is next and gradation of the images (lines, zones and characters) follow suit. The blend of salient features and pattern similarity is engaged for identification using the combination of feature based and template based approaches. For recognizing the basic characters, feature-based classifier was employed whereas the compound characters, a combination of two or more basic characters, were recognized by employing run-based template approach preceded by grouping the characters. The highest score reported was 96% for printed Bangla scripts.

Rajasekaran and Deekshatulu [17] employed a two-stage classification system for printed Telugu characters. The curve tracing approach was used to get the primitive shapes in the first stage. These primitive shapes were further used to obtain the Telugu characters. Patterns were obtained by tracing the distinctive points, after removing the primitive shapes from the character, in the second stage. They reported the recognition accuracy as 95.34% for printed Telugu characters.

Salimi and Giveki [18] presented recognition system for Arabic handwritten numerals using singular value decomposition (SVD) classifiers and multi-phase particle swarm optimization (PSO). By partitioning the image horizontally and vertically, features were extracted using 2D PCA algorithm. The partitioned images were classified using two simple PSO rules. The combined PSO rule was used to minimize the fitness function. The handwritten numerals were classified using several classifiers such as multi-layer perceptron (MLP), radial basis function (RBF), SVD and artificial neural fuzzy inference system (ANFIS) in their work. They reported recognition accuracies ranging from 63% to 91.7% using various classifiers mentioned.

Murthy and Ramakrishnan [19] proposed a hierarchical approach for Online handwritten Tamil and Kannada datasets. In first stage of classification the non-parametric classifier, NNC, with PCA was used to classify the characters. This stage reduces the number of classes and these reduced classes were fed to multiple classifiers at second stage. In second stage using different features separately, with a dynamic time warping (DTW) classifier, the reduced set of confusion classes were classified. The different features on which the work was carried out were quantized slope, dominant points and quartile features. The reported recognition accuracy for Tamil dataset was 81.1% and 90.2%, using single and multistage classifiers respectively. The recognition accuracy (RA) for Kannada basic characters was reported as 76.5%. The RA improved to 92.2% using multistage classifiers.

Aradhya et al. [20,21] proposed character identification using combination of FT and PCA. Aradhya et al. [22] presented recognition of numerals and multilingual text using wavelet transforms and wavelet entropy, respectively.

Abaynarh et al. [23] contributed work to recognize Amazigh handwritten characters by extracting Legendre moment based features. They reported recognition rate using neural network classifier as 97.46%.

Malik and Dixit [24] adopted wavelet transform and Hop-field network to recognize English alphabetical handwritten characters up to a distortion level of 30%. Further they reported that for a few characters having distortion level between 30% and 40% could be recognized. The characters having distortion level above 40% could not be recognized by their method.

Das et al. [25] contributed work on recognizing Bangla isolated handwritten compound characters. Using the combination of topological components like longest run and convex hull based