A Holistic Approach for Handwritten Hindi Word Recognition

Samir Malakar, Department of Computer Science, Asutosh College, Kolkata, India
Pankaj Sharma, ATOS India, Kolkata, India
Pankaj Kumar Singh, Wipro Limited, Kolkata, India
Maitrayee Das, Ensim India Pvt. Ltd., Kolkata, India
Ram Sarkar, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
Mita Nasipuri, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

ABSTRACT

Holistic word recognition attempts to recognize the entire word image as a single pattern. In general, it performs better than segmentation based word recognition model for known, fixed and small sized lexicon. The present work deals with recognition of handwritten words in Hindi in holistic way. Features like area, aspect ratio, density, pixel ratio, longest run, centroid and projection length are extracted either from entire word image or from the hypothetically generated sub-images of the same. An 89-elements feature vector has been designed to represent each word in the feature space and five different classifiers have been used for measuring recognition performances. Considering the complexities of Hindi characters, the technique shows an impressive result using a Multilayer Perceptron (MLP) based classifier. Moreover, the technique shows scale and rotation invariant nature to a significant extent.

KEYWORDS

Devanagari Script, Handwritten Word, Hindi, Holistic Word Recognition, OCR

INTRODUCTION

Technological advancements open the door for using digital media for communication which leads to a paperless society. The first and foremost requirement to achieve this goal is to digitize the huge existing paper documents, either printed or handwritten form. Optical Character Recognition (OCR) would be a good choice to accomplish this as OCRed data are manipulated and maintained very easily by the digital devices. OCR systems which are described in the works Chaudhuri & Pal (1998) and Natarajan et al. (2001) perform well for printed documents. But no such good OCR system is found in literature for handwritten documents till date. The poor performance of the system on handwritten documents is mainly due to the lack of efficient word extraction module, described in Sarkar et al. (2011), and word recognition module, described in Basu et al. (2009), Bhowmik et al. (2015). In the literature, two major approaches for word recognition are found viz., segmentation based word recognition (Basu et al., 2009) and holistic word recognition (Bhowmik et al., 2015).

DOI: 10.4018/IJCVIP.2017010104

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
In the first category of approaches, word images are first segmented into constituent characters and then each character is recognized to get represented with some standard codes for storing them in a digital format which can be editable by standard word processing software. The most advantageous point about segmentation based approach is that it is lexicon free. But in most of the cases, it performs poorly for handwritten documents due to the ambiguity of the character segmentation (Edelman et al., 1990). The low segmentation accuracy indicates the failure of segmentation algorithm to yield the ideal segmentation points on the handwritten word images resulting in under/over segmentation (Malakar et al., 2011). As a result, the OCR systems are forced to work with various combinations of erroneous characters or their components. Deciding the possible character subparts from over-segmented characters and all possible combinations of valid characters or character subparts lead to large set of patterns to recognize. Thus, designing such systems to recognize word(s) for unbounded lexicon is perhaps impossible. The second approach i.e., the holistic approach has been introduced by the researchers as an alternative Handwritten Word Recognition (HWR) mechanism. This mechanism tries to recognize the whole words as a single pattern. It provides acceptable recognition accuracy for small and pre-defined lexicons. In other words, this approach is not suitable for unknown and large lexicon.

Therefore, it is a wise decision to apply the holistic word recognition methodology when the words to be recognized belong to a small lexicon and pre-defined. Therefore, while recognizing the names of months, week days, cities, countries, states, capitals, user defined keywords/key phrases etc. holistic approach would be the better option.

Characteristics of Hindi Language and Its Script

Hindi is a widely-used language in India. Based on the information published in the article “Wikipedia: List of languages by number of native speakers”, it ranks 1st in India and 4th in the world. According to the information published in Ethnologue website (19th ed., 2016), around 260 million of people are native Hindi speakers. Government of India uses Hindi as official language along with the English language. It is also the official language of the Indian states like Bihar, Chhattisgarh, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh, and Uttarakhand and union territories like Andaman & Nicobar Islands, Chandigarh, Dadra & Nagar Haveli, Daman & Diu and Delhi.

Hindi language is written using Devanagari script which is an Indic script used to write various languages of Indian subcontinent. Unlike Roman script, Devanagari character-set consists of a much larger number of characters. In Devanagari script, there are 48 basic characters, 11 modified shapes and around 260 compound characters. Among these basic characters, 12 are vowels and 36 are consonants. Figure 1, Figure 2, and Figure 3 show the handwritten shapes of Devanagari vowels, consonants and modified shapes respectively.

Devanagari script is written horizontally from left to right and it is a Matra based script (Sarkar et al., 2010). Matra is a distinctive horizontal line running along the top of the characters of the scripts like Bangla, Gurumukhi, Sylheti Nagar etc. Matra covers partly or fully the entire character width depending on the character. The consecutive characters, which have Matra, are joined to form common Matra (Figure 4). Word written in Devanagari script can be partitioned horizontally into three adjacent zones (Sarkar et al., 2010) namely Upper Zone, Middle Zone and Lower Zone (Figure

![Vowels in Devanagari script](image-url)
A Key Point Selection Shape Technique for Content Based Image Retrieval System
www.igi-global.com/article/a-key-point-selection-shape-technique-for-content-based-image-retrieval-system/171131?camid=4v1a

Fast and High Capacity Digital Image Watermarking Technique Based on Phase of Zernike Moments
www.igi-global.com/chapter/fast-high-capacity-digital-image/77547?camid=4v1a