Bangla and Oriya Script Lines Identification from Handwritten Document Images in Tri-script Scenario

Sk Md Obaidullah, Aliah University, Kolkata, India
Chayan Halder, West Bengal State University, Kolkata, India
Nibaran Das, Jadavpur University, Kolkata, India
Kaushik Roy, West Bengal State University, Kolkata, India

ABSTRACT

In this paper, two popular eastern Indian scripts namely Bangla and Oriya are considered for Line-level script identification considering two Tri-script groups where Devnagari and Roman are kept common in each group. A 27 dimensional feature vector has been constructed using FD (Fractal Dimension) and IMT (Interpolated Morphological Transform). 600 Line-level handwritten document images of each Tri-script groups have been considered for experimentation. Promising results has been found using multiple classifiers where MLP (Multi-Layer Perceptron) Neural Network and LMT (Logistic Model Tree) perform best for BDR (Bangla-Devnagari-Roman) combinations with 97% accuracy and LMT outperforms others for ODR (Oriya-Devnagari-Roman) combinations with 97.7% accuracy. Bi-script performance analysis has also been made where combinations BR (Bangla-Roman) and BD (Bangla-Devnagari) results with accuracy of 98% and 97.5% respectively for the first group. Whereas for the second group OD (Oriya-Devnagari) and OR (Oriya-Roman) shows an accuracy of 98.25% and 98% respectively.

KEYWORDS

Average Accuracy Rate, Classifier Comparison, Fractal Dimension, Handwritten Script Identification, Interpolated Morphological Transform, Line-Level Document, MLP Neural Network, Model Building Time, OCR, Performance Analysis

1. INTRODUCTION

India is a multi-lingual/multi-script country with 23 official languages (including English) and 13 different scripts (including Roman) are used to write those languages (Obaidullah et al. 2013; Ghosh et al. 2010). Multi-script documents are very common scenario in our country where a single document may be written by using more than one script types. In our daily life we come across various such multi-script documents like Postal Documents (Roy et al. 2004; Radha et al. 2014), Govt. Application Forms, Railway Reservation Forms etc. Another situation exists where collection of different documents written by different scripts need to be handle. Postal document sorting is one such example. To address these document processing problems there is a pressing need of development of sophisticated and smart techniques. Optical character recognition or in short OCR is an intelligent technique to convert image into its textual version. Normally a particular OCR is applicable for a particular script for which it was designed. As an example, Bangla OCR can process document images

DOI: 10.4018/IJSSMET.2016010103

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
having Bangla/Asamese/Manipuri characters, Devnagari OCR can process Hindi/Marathi/Sanskrit etc. document images. But when it comes about multi-script document containing more than one scripts say Bangla and Devnagari both then a single OCR fails to process such document images. Not only that, if a single document is written by single script then manual intervention is required to choose a specific OCR for a specific document image. This manual intervention leads to huge manpower and resource loss at current scenario. To overcome the problem one feasible solution is to develop an automatic script identification system for all official Indic scripts first then choosing appropriate script specific OCR to process document image written by a specific script. Present work emphasizes to develop an approach for Line-level script identification technique observing its usefulness in many real life document images where script differentiation is present at line wise. Four eastern Indian scripts namely Bangla, Oriya, Devnagari and Roman are considered for the present work. Brief discussions about these four scripts are presented in the following section.

**Bengali Script:** Bengali, originated from the Indo-European language family (Obaidullah et al. 2014), is the state language of West Bengal and Tripura. People living in Bihar, Dhanbad, Manbhum, Singhbhum, Santal Parganas of Jharkhand, Garo Hills of Meghalaya, Goalpara District of Assam, Nagaland and Mizoram also use this language. It is spoken by 181 million population of India. Assamese can be classified under the Indo-European group of languages. It is the state language of Assam. About 16.8 million of the Indian population living in Arunachal Pradesh, Meghalaya and West Bengal use this language. Manipuri, originated from the Sino-Tibetan language group. People of Manipur mostly use the language. It is also used in Karimganj and Cachar of Assam, West and North Tripura districts, Nagaland, West Bengal and Uttar Pradesh. About 13.7 million people use this language. Oriya Script: Oriya, originated from the Indo-European language family (Obaidullah et al. 2014). It is state language of Orissa. About 31.7 million population living in Assam, Singhbhum and Ranchi districts of Jharkhand, Raipur, Raigarh and Bastar districts of Chhattisgarh, Midnapore district of West Bengal and Vishakhapatnam District of Andhra Pradesh. Devnagari Script: Hindi, one of the most popular languages in India uses this script (Obaidullah et al. 2014). Under Indo-European language family, Hindi is predominantly used by about 182 million people residing in the northern parts of India namely in Uttar Pradesh, Himachal Pradesh, Uttarakanchal, Delhi, Rajasthan, Punjab, Madhya Pradesh and northern Bihar. Marathi, a member of the Indo-European language family was given the status of the State language of Maharashtra. People dwelling in the neighboring states also use this language. Marathi is used by 68.1 million people. Bodo, a member of the Sino-Tibetan group of language used by the people living in South Bank of Assam, Chandel District of Manipur, West Garo Hills District of Meghalaya and Darjeeling, Jalpaiguri and Coch-Behar of West Bengal. About 0.5 million people use this language. Konkani belongs to the Indo-European language family. The language is used by inhabitants of Karnataka, Kerala, Dadra and Nagar Haveli and the north and central coastal strip of Maharashtra. It is the official language of Goa. About 7.6 million people use this language. Sanskrit is under the family of Indo-European group of language, it is mainly used as the liturgical language. About 0.03 million people use this language. The origin of the Sindhi language can be traced to the Indo-European language family. One of the official languages of India, Sindhi is used by many people residing in Uttar Pradesh, Delhi, Rajasthan, Madhya Pradesh, Gujarat, Maharashtra, Andhra Pradesh, Tamil Nadu, Orissa and Bihar. About 21.4 million people use this language. Nepali language belongs to the Indo-European language group. About 13.9 million people living in Darjeeling area of West Bengal, Sikkim, Arunachal Pradesh, Manipur, Nagaland, Meghalaya, Tripura, Mizoram, Assam, Bihar, Himachal Pradesh, Uttarakanchal, Uttar Pradesh and Haryana use this language.
Changes in CRM Approach: Refined Functional Blocks for Customer Creative Engagement in Services
www.igi-global.com/article/changes-in-crm-approach/142871?camid=4v1a

Online Services Delivered by NTO Portals: A Cross-Country Examination
www.igi-global.com/article/online-services-delivered-nto-portals/4022?camid=4v1a