Deep Learning for Trilingual Character Recognition

M. Yashodha, Dept. of ISE, Bahubali College of Engineering, India
SK Niranjan, JSS Science and Technology University, Mysore, India
V. N. Manjunath Aradhya, Sri Jayachamarajendra College of Engineering, Mysore, India

ABSTRACT

As India is a multilingual country, in which the national language is Hindi, regional languages still exist in each of the corresponding states. In government offices, for the purpose of communication and maintenance of files and ledgers, the languages preferred are the regional languages and Hindi. As corporate offices and private organizations also exist in the country, these bodies mainly prefer the English language with the regional language in recording documents and ledgers. So, in this regard, in India a document contains multilingual texts, and there is a need of a multilingual OCR system. In this article, a trilingual OCR system is developed using deep learning for supporting English, Hindi and Kannada languages, the regional language of the state Karnataka.

KEYWORDS

Artificial Neural Networks, Bilingual, Deep Neural Networks, OCR, Trilingual

INTRODUCTION

One of the human functioning is reading, that has been evolved with the computer by researchers to implement optical character recognition (OCR). Now OCR has become one of the effectual applications of technology in the area of Image Processing and Pattern Recognition. The main aim of OCR is to classify optical patterns of alphabetic characters and numerals that are present in digital image. OCR technology allows the system as a powerful recognition software which can recognizes both printed and handwritten texts of varied fonts. Optical character recognition is an approach towards translating images of printed or handwritten text into machine editable text. The task of character recognition is achieved through segmentation, feature extraction and classification stages. Optical character recognition can be performed offline once the writing or printing has been completed whereas the online character recognition is attained by using computer in which it recognizes the characters as they are written / printed at that moment. Both handwritten and printed characters may be recognized but the accomplishment is directly proportional to the quality of input documents. Towards achieving accurate recognition of text under all possible conditions, OCR still remains a highly challenging task. Currently, efficient OCR packages are accessible for recognition of printed documents of languages such as English, Chinese etc. In concern to Handwritten Character Recognition (HCR) facilities are extensive in English and Chinese languages, where as in Indian languages many research works are employed to develop the OCR system for Hindi, Devanagari, Oriya, Tamil, Telugu and Kannada, etc. In this area more research and development are still happening with continuous effort.

DOI: 10.4018/IJNCR.2019010104

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
RELATED WORKS

A document with more than one language shall be considered as a multilingual document. We find the multilingual documents in the form of application forms, like railway reservation forms, bank challans, job application forms, etc., where the information has to be provided using more languages. Our nation is a multilingual country where many official documents may contain at least three languages, one which is official language of the local state, and the other languages are Hindi and the language which is common for the purpose of official communication that is English. Thus, documents of multilingual nature are a challenge for document analysis tasks and further recognition tasks when compared to documents with single scripts which are known as the monolingual documents. The processing of multilingual documents is carried out by using script identification algorithms for any particular language. This avenue of multilingual document analysis shall be categorized into two types: viz., printed and handwritten documents. There are many attempts in the literature towards the analysis of multilingual document in printed nature. But there is a huge scope towards the hand printed multilingual document analysis.

Few attempts are made for the recognition of English, Chinese, Korean etc., A very few attempts have been carried out with respect to Indian scenario. In Aradhya (2005), a method for determining the up/down orientation of text in a scanned document of unknown orientation has been proposed. It establishes the direction in which the text as a whole is oriented, which finds its application in automated mail processing. In Basu et al. (2010), proposed a method to recognize the postal codes written in any of the four popular scripts viz Latin, Bangla, Devanagari and Urdu. First, they implemented a Hough transformation based technique to localize the postal-code blocks from structured postal documents with defined address block region. A script independent Unified pattern Classifier is used to classify these clusters. In Wen and Ding (2004), proposed a general framework for character segmentation in complex Multilingual documents. In Choudhary et al. (2013), proposed an approach to segment character images from the text containing images and computer printed or handwritten words. The character extraction technique is applied for the segmentation of untouched characters from machine printed or handwritten words on a noisy background having some images. In Aradhya et al. (2007), proposed a multilingual character recognition system for printed South Indian scripts and English documents. The approach is based on Fourier transform and Principal Component Analysis (PCA). In Pujari et al. (2004)), proposed a method where wavelet Multi-Resolution analysis for the purpose of extracting features and associative memory model to accomplish the recognition tasks. System has a Hopfield-based Dynamic Neural Network for the purpose of learning and recognition, which overcomes the inherent difficulties of memory limitation and spurious states in the Hopfield Network.

In Naveena et al (2013), during character segmentation process in Kannada word is a crucial task due to the presence of bottom extension characters and Modifiers. In this regard, they proposed enhanced character segmentation (eCS) for an unconstrained handwritten Kannada scripts. A cluster mean points are used to estimate the direction and branch point for segmenting characters. Few interesting works on handwritten character recognition and ANN can be in (Pal, 2002; Pal, 2004; Zhu, 2009; Pal, 2010; Saeeda, 2014; Impedovo, 2014; Alireza, 2011; Singh and Pawan Kumar, 2018; Gautam, 2018; Le, 2017; Tiwari, 2018; Bhatija, 2018).

PROPOSED METHODOLOGY

The proposed model consists of two stages viz., training and testing. The training stage mainly involves preprocessing, model creation and model training. In testing stage, we have preprocessing and finding out the decision of the given test sample as shown in Figure 1.

The following subsection describes detail of each component of the model as follows.
Solving Uncapacitated Facility Location Problem Using Heuristic Algorithms
*International Journal of Natural Computing Research* (pp. 18-50).  
[www.igi-global.com/article/solving-uncapacitated-facility-location-problem-using-heuristic-algorithms/225822?camid=4v1a](www.igi-global.com/article/solving-uncapacitated-facility-location-problem-using-heuristic-algorithms/225822?camid=4v1a)

Super-Efficiency DEA Approach for Optimizing Multiple Quality Characteristics in Parameter Design
[www.igi-global.com/article/super-efficiency-dea-approach-optimizing/44671?camid=4v1a](www.igi-global.com/article/super-efficiency-dea-approach-optimizing/44671?camid=4v1a)