Hindi Text Document Classification System Using SVM and Fuzzy: A Survey

Shalini Puri, Birla Institute of Technology, Ranchi, India
Satya Prakash Singh, Birla Institute of Technology, Ranchi, India

ABSTRACT

In recent years, many information retrieval, character recognition, and feature extraction methodologies in Devanagari and especially in Hindi have been proposed for different domain areas. Due to enormous scanned data availability and to provide an advanced improvement of existing Hindi automated systems beyond optical character recognition, a new idea of Hindi printed and handwritten document classification system using support vector machine and fuzzy logic is introduced. This first pre-processes and then classifies textual imaged documents into predefined categories. With this concept, this article depicts a feasibility study of such systems with the relevance of Hindi, a survey report of statistical measurements of Hindi keywords obtained from different sources, and the inherent challenges found in printed and handwritten documents. The technical reviews are provided and graphically represented to compare many parameters and estimate contents, forms and classifiers used in various existing techniques.

KEYWORDS

Character Recognition, Fuzzy Logic, Handwritten Documents, Hindi Document Classification, Image Text Processing, Printed Documents, Statistical Analysis, SVM, Word Recognition

INTRODUCTION

Hindi Document Mining (HDM) refers to the search and retrieval of important relevant information from Hindi documents, while ignoring the irrelevant ones, so that this information can be used to classify the unknown text documents into different groups. Hindi Text Document Classification (HTDC), an extension of HDM, is used to classify the documents into pre-specified and labeled, mutually exclusive categories by first providing them training and then performing the system testing and classification on them (Puri, 2011; Puri & Kaushik, 2012). The complexity of such automatic and intelligent Hindi textual classification systems increases when they need to classify the scanned printed and handwritten text documents into pre-defined classes. The training, testing and classification of these documents are very much crucial and challenging, and therefore, require major attention. The document type ranges from simple Hindi text to the scanned versions of printed and handwritten

DOI: 10.4018/IJRSDA.2018100101

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
text documents (Bag & Harit, 2013; Bagadkar & Malik, 2014; Chaudhari & Pal, 1997; Chau & Yeh, 2002; Hole & Ragha, 2011; Jayadevan, Kolhe, Patil, & Pal, 2011b; Kumar, Holambe, Thool, & Jagade, 2012; Nevetha & Baskar, 2015; Puri & Singh, 2016; Pal, Wakabayashi, & Kimura, 2009). Looking insight the application domains and to analyze today’s scenario of current existing Devanagari and Hindi based systems, the scope of such systems is found limited only to script recognition and discrimination (Hassan, Garg, Chaudhury, & Gopal, 2011; Kumar et al., 2012); identification, recognition and Shirorekha removal at character and word (keyword) levels (Shinde, & Dandawate, 2014); text summarization; recognition and separation of objects of multi – color, – font, – orientation and – size (Singh, Mittal, & Ghosh, 2014); and separation of images from non – images.

Going into much detail of Hindi character and word extraction, it leads towards another dimension for Hindi text information retrieval (Sinha, 2009) which is related to the shirorekha removal, Connected Component (CC) separation, basic and modified character recognition, matching and classification etc. Although the spectrum of these application areas covers tremendous research work performed in recent years, yet it opens a new way of designing these automated systems for imaged document classification through extracting, analyzing, and verifying the textual image contents, then gathering and reorganizing the extracted components, and finally training the classifier to be intelligent enough which can further categorize new images into pre-defined classes. Therefore, the proposed work is to introduce a new and innovative idea of Hindi image document classification, which accepts printed and/or handwritten Hindi text images which are classified into pre-defined categories. Correlating the proposed idea with the scanned versions of HTDC, the HTDC is said to be broadly divided into three types – Hindi printed document classification (HPDC), Hindi handwritten document classification (HHDC), and Hindi mixed document classification (HMDC), where HPDC, HHDC and HMDC accept printed, handwritten and mixed printed-handwritten Hindi images, respectively. All classification systems require multi-stage categorization, in which optical character recognition (OCR) is only an integral part of this system, whose results finally become the base for whole document classification. Such Hindi document classification needs to be designed beyond the OCR, which can lead towards the new dimensional area in Hindi printed and/or handwritten document classification. The underlying concepts of HPDC and HHDC are discussed here. HPDC (or HHDC) system acquires the scanned printed (or handwritten) images, which are preprocessed to extract lines, words, and shirorekha – less characters, which are further recognized and classified by the classifier. After this, these shirorekha – less characters are joined to get shirorekha – less words, called as Recognized Keywords (RK), which are further matched with the stored keywords, so that the document is categorized into correct pre – specified category.

To evaluate the applicability and availability of HTDC and its variations with the existing techniques, it was found that current available research contributions are primarily concerned towards the OCRing of simple, complex and degraded documents, and many applications such as, summarization, signature verification, bank cheque processing, criminal handwriting matching etc. further use and process the OCR outputs. So, in general terms, an OCR is meant for character categorization. Bansal and Sinha (2001), Gohil et al. (2012), Ma and Doermann (2003), Mishra and Agrawal (2005), Murthy et al. (2013), Raj and Ghosh (2014), and Sihag et al. (2009) proposed their work for Hindi (monolingual) printed character categorization through the steps of preprocessing, segmentation, feature extraction and character classification, whereas, Jain et al. (2011) proposed word based processing. Raman and Sahu (2015) presented script recognition method, and Dongre and Mankar (2011) processed the Devanagari document without Shirorekha removal. P. Agrawal et al. (2009), Arora et al. (2007), Arora et al. (2009), Chaudhuri et al. (2016), Desai and Malik (2011), Deshpande et al. (2006), (2007a), (2007b), Gaur and Yadav (2015), Hanmandlu et al. (2007), Jangid (2011), Kekre et al. (2013), Mukherji and Rege (2009), Pal et al. (2007), Rojatkar et al. (2013), Sahu and Raman (2013), Sharma et al. (2006), Shelke and Apte (2015), Singh et al. (2015), Singh et al. (2011), Thakral and Kumar (2014), and Verma (1995) presented Hindi handwritten character categorization systems, i. e., OCR, which followed very similar steps as of Hindi printed character