A Fuzzy Matching based Image Classification System for Printed and Handwritten Text Documents

Shalini Puri, Birla Institute of Technology, India
https://orcid.org/0000-0002-1700-1865
Satya Prakash Singh, Birla Institute of Technology, India

ABSTRACT

This article proposes a bi-leveled image classification system to classify printed and handwritten English documents into mutually exclusive predefined categories. The proposed system follows the steps of preprocessing, segmentation, feature extraction, and SVM based character classification at level 1, and word association and fuzzy matching based document classification at level 2. The system architecture and its modular structure discuss various task stages and their functionalities. Further, a case study on document classification is discussed to show the internal score computations of words and keywords with fuzzy matching. The experiments on proposed system illustrate that the system achieves promising results in the time-efficient manner and achieves better accuracy with less computation time for printed documents than handwritten ones. Finally, the performance of the proposed system is compared with the existing systems and it is observed that proposed system performs better than many other systems.

KEYWORDS


INTRODUCTION

The automated English Document Image Classification System (EDC) is designed to categorize the scanned pure printed or pure Hand-Written (HW) English documents into mutually exclusive predefined classes. This system is defined in 2 ways, which are English Printed Document Classification (EPDC) system and English Handwritten Document Classification (EHDC) system. Both EPDC and EHDC systems use the underlying concepts of pattern recognition, artificial intelligence, machine learning, text mining, and image mining fields. Over the last three decades, many researchers have successfully implemented various automated text mining and classification systems. These researchers have tested their systems on mono, bi, tri, and multi-lingual real and synthetic documents by using various classifiers and variations. Some other researchers have also provided good solutions for the problems of feature reduction, feature selection, and data and curse of dimensionality reductions. On the other side, in recent years, many automated image recognition, identification, and mining systems have also been developed. These systems were designed primarily for the categorization of maps, geographical areas, drawings, and graphical and pictorial designs. Nowadays, many other image...
mining systems have also come into existence, which extract and process text characters, words, and lines from the heterogeneous set of multi-font, multi-size, multi-oriented, multi-colored, multi-lingual and multi-script documents. The fields of printed character recognition and script discrimination for non-Indic, such as, Latin, Chinese, Japanese and Korean scripts are already mature. On the other side, many printed text recognizers and processors also exist for Indic scripts, such as, Devanagari, Bengali, Gujarati, and Gurumukhi etc. The printed text processing systems are always found simpler than the handwritten ones. The reason behind the complexity of handwritten text processing primarily lies in the cursive writing style, overlapped and touched characters, and uneven height, size and gaps among the characters and words. Secondly, it also depends on the writer how smoothly and clearly he writes the text. Many Indian scripts also use a head line on the top of the characters, which also increase the segmentation issues. All these conceptual illustrations of text classification systems and image mining systems have motivated the authors to propose an integrated single and multi-script document image classification system, which accepts the text document images and categorizes them into predefined classes. In this way, the area of document classification coexists with the image content retrieval and recognition paradigm.

These new dimensions of text document image processing include the major steps of preprocessing, character recognition, word recognition, and document classification. Nowadays, many researchers are paying attention to it. Puri and Singh (2018) provided a survey on Devanagari scripted Hindi text document classification system by using Support Vector Machine (SVM) and fuzzy. This survey primarily focused upon Hindi basics, importance, survival, and differentiation between Hindi and other scripts, and then it provided detailed discussions on existing research contributions from 1990 to till date. Another research contribution is a tri-layered segmentation and bi-leveled classifier based advanced, robust, fast Hindi Printed Document Classification using SVM and Fuzzy (HPDC-SF), which discussed detailed algorithmic procedures for document classification (Puri & Singh, 2019). The HPDC-SF system was designed to categorize unknown documents into predefined Hindi classes through the critical Task Stages (TS) of segmentation, Shirerekha-Less (SL) character extraction, SL word association, fuzzy matching, and classification. This system used Predefined Keywords (PK) in Romanized form of Hindi characters.

Based upon the HPDC-SF method and terminologies, a new and advanced bi-leveled Latin scripted English Document Classification system using SVM and Fuzzy (EDC-SF) is proposed, which accepts the pure text images, extracts and recognizes the characters, classifies isolated characters through SVM, calculates their confidence (score) values, joins these isolated characters to obtain the associated words (Detected Words), matches such Detected Words (DW) with PKs, calculates PK confidence values, and finally, classifies the imaged documents into classes by using fuzzy. The proposed EDC-SF system is generic, robust and fast, which can accept any number of documents. Its design implements English Printed Document Classification system using SVM and Fuzzy (EPDC-SF) and English Handwritten Document Classification system using SVM and Fuzzy (EHDC-SF) for pure printed and pure handwritten offline scanned documents, respectively, and shows high performance in both cases. To test the performance of EDC-SF, a large set of indigenous English printed and handwritten documents are collected and generated. The printed documents have been obtained from online websites, Government sites, portals, magazines, newspapers, and blogs, and the handwritten documents have been written by five different writers of the age group 15 years to 36 years.

There are many differences between EPDC-SF and EHDC-SF designs and HPDC-SF design, which are discussed here. The 3-layered segmentation of HPDC-SF is replaced by only one-layered segmentation TS in EDC-SF. The HPDC-SF segmentation stages segmented the documents into the lines, words, and SL characters. The reason behind such differences is that EDC-SF design does not require SL characters, because English alphabets and words are not bound to shirerekha (headline), and therefore, do not constitute SL characters and SL words. The second difference is that English letters are case-sensitive in nature as compared to Hindi, and English capital and small letters create differences for exact matching in Optical Character Recognition (OCR) and in DW–PK matching.
Offshoring and Transfer of Intellectual Property
[www.igi-global.com/article/offshoring-transfer-intellectual-property/38911?camid=4v1a](www.igi-global.com/article/offshoring-transfer-intellectual-property/38911?camid=4v1a)

Standards for Web-Based Integration Adapters
[www.igi-global.com/chapter/standards-web-based-integration-adapters/14660?camid=4v1a](www.igi-global.com/chapter/standards-web-based-integration-adapters/14660?camid=4v1a)