Word-Level Script Identification Using Texture Based Features

Pawan Kumar Singh, Department of Computer Science and Engineering, Jadavpur University, 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

Script identification is an appealing research interest in the field of document image analysis during the last few decades. The accurate recognition of the script is paramount to many post-processing steps such as automated document sorting, machine translation and searching of text written in a particular script in multilingual environment. For automatic processing of such documents through Optical Character Recognition (OCR) software, it is necessary to identify different script words of the documents before feeding them to the OCR of individual scripts. In this paper, a robust word-level handwritten script identification technique has been proposed using texture based features to identify the words written in any of the seven popular scripts namely, Bangla, Devanagari, Gurumukhi, Malayalam, Oriya, Telugu, and Roman. The texture based features comprise of a combination of Histograms of Oriented Gradients (HOG) and Moment invariants. The technique has been tested on 7000 handwritten text words in which each script contributes 1000 words. Based on the identification accuracies and statistical significance testing of seven well-known classifiers, Multi-Layer Perceptron (MLP) has been chosen as the final classifier which is then tested comprehensively using different folds and with different epoch sizes. The overall accuracy of the system is found to be 94.7% using 5-fold cross validation scheme, which is quite impressive considering the complexities and shape variations of the said scripts. This is an extended version of the paper described in (Singh et al., 2014).

Keywords: Handwritten Words, Histograms of Oriented Gradients (HOG), Indic Scripts, Moment Invariant Features, Optical Character Recognition, Script Identification, Statistical Significance Testing

1. INTRODUCTION

Optical Character Recognition (OCR) is the conversion of scanned images of handwritten, typewritten or printed text into machine-encoded format. In any multilingual and multi-script world, OCR systems need to be capable of recognizing characters irrespective of the script in which they are written. In general, recognition of characters written in different scripts by a

DOI: 10.4018/ijsda.2015040105
single OCR module is next to impossible. This is because of features which are necessary for character recognition depend on the structural properties, style and nature of writing which mainly varies from one script to another. For example, features used for identification of Roman script might not be useful for identifying other scripts. This could possibly be solved by using a bank of OCRs (different OCRs for different scripts) corresponding to different scripts. The text in an input document can then be recognized reliably by selecting the appropriate OCR system from the OCR repository. However, it requires a priori knowledge of the script in which the document is written. But, manual identification of the documents’ scripts may be monotonous and time consuming (P. K. Singh et al., 2014).

India is a multilingual country where people residing at different sections use different languages/scripts. However, Roman script is frequently used in conjunction with different Indic scripts in their daily life/regular work. Therefore, in this multilingual environment, to develop a successful OCR system for any script, separation or identification of different scripts is of utmost important. In an automated multilingual environment, such document processing systems relying on OCR would clearly need human intervention to select the appropriate OCR package, which is certainly inefficient, undesirable and impractical. It is difficult to feed a document as an input to OCR unless the script/language type of the text in it is pre-determined since a single OCR cannot recognize multiple scripts. The solution of this problem is to develop an automatic script identification system. Script identification facilitates many important applications such as sorting the document images, selecting appropriate script specific text understanding system and searching online archives of document images containing a particular script, etc.

Difficulties inherent in recognizing handwritten text due to the large variations in handwriting styles pose huge challenges. Due to varied writing styles, resemblances among different scripts are more feasible for handwritten documents rather than the printed ones. Cultural/individual differences, and even differences in the way that people write at different times due to diversified cultures across the globe, enlarge the inventory of possible word shapes seen in handwritten documents. Also, problems typically pertaining in preprocessing, such as ruling lines, word fragmentation, noise, skew, etc. are common in handwritten documents. Since, in a multilingual document, the script may vary from word to word, and not from character to character, so the identification of the scripts at word-level are more preferable than at character or line level. Performing script identification at word-level is much more challenging than at text-line and page-level because the information gathered from few characters present in a single word may not be adequate for the script recognition purpose.

The rest of the paper is organized as follows: a brief survey related to script identification is described in Section 2 and some basic information related to scripts used in the present work is illustrated in Section 3. The proposed technique based on texture based features is presented in Section 4 whereas Section 5 describes the selection of some well-known classifiers used in the present work. The experimental results and discussions are given in Section 6. Section 7 concludes the work and lists future directions of the work.

2. RELATED WORKS

Ample research work has been reported in literature for script identification of printed text for both Indic and non-Indic scripts. Wood et al. (1995) described vertical and horizontal projection profile method for determining Roman, Russian, Arabic, Korean and Chinese scripts. They argued that the projection profiles of document images are sufficient to characterize different scripts. However, the authors did not suggest how these projection profiles can be analyzed automati-
A Novel Time Series Forecasting Approach Considering Data Characteristics

www.igi-global.com/article/a-novel-time-series-forecasting-approach-considering-data-characteristics/120580?camid=4v1a