Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cosrev

Survey

Offline Script Identification from multilingual *Indic*-script documents: A state-of-the-art

*Pawan Kumar Singh**, *Ram Sarkar*, *Mita Nasipuri*

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

ARTICLE INFO

Article history:
Received 18 April 2014
Received in revised form
20 October 2014
Accepted 3 December 2014
Published online 27 December 2014

Keywords:
Multilingual document
Offline Script Identification
Indic scripts
Optical Character Recognition
Document image analysis

ABSTRACT

Offline Script Identification (OSI) facilitates many important applications such as automatic archiving of multilingual documents, searching online/offline archives of document images and for the selection of script specific Optical Character Recognition (OCR) in a multilingual environment. In a multilingual country like India, a document containing text words in more than one language is a common scenario. A state-of-the-art survey about the techniques available in the area of OSI for *Indic* scripts would be of a great aid to the researchers. Hence, a sincere attempt is made in this article to discuss the advancements reported in the literature during the last few decades. Various feature extraction and classification techniques associated with the OSI of the *Indic* scripts are discussed in this survey. We hope that this survey will serve as a compendium not only for researchers in India, but also for policymakers and practitioners in India. It will also help to accomplish a target of bringing the researchers working on different *Indic* scripts together. Taking the recent developments in OSI of Indian regional scripts into consideration, this article will provide a better platform for future research activities.

Contents

* *Corresponding author.*
E-mail addresses: pawansingh.ju@gmail.com (P.K. Singh), raamsarkar@gmail.com (R. Sarkar), mitanasipuri@gmail.com (M. Nasipuri).

http://dx.doi.org/10.1016/j.cosrev.2014.12.001
1574-0137/

1. Introduction

With the advancement in computer technology and the availability of low cost high capacity storage devices, storing of documents in electronic form has become a common practice. A document either in handwritten or printed form may contain writings in different scripts, graphics and images. For example, museum archives contain old fragile documents having scientific or historical or artistic value and written in different scripts with many graphic illustrations. OCR is a type of software designed to translate images of text into machine editable text. However, most OCR systems are script-specific in the sense that they can read characters written in a particular script only. Script is defined as the graphic form of the writing system. A script class refers to a particular style of writing and the set of characters used in it. Languages throughout this world are typeset in many different scripts. A script may be used by only one language or may be shared by many languages, sometimes with slight variations from one language to other. For example, *Devnagari* script is used for writing a number of *Indic* languages like *Sanskrit*, *Hindi*, *Konkani*, *Marathi*, *Nepali*, etc. It is perhaps impossible to design a single recognizer which can identify a large number of scripts/languages. So, for a multilingual text document, identification of different scripts and extraction of portions written in the same script is a pressing need so that script-specific OCR systems can be employed. However, manual identification of the mixed-script document may be too monotonous and imperceptible. So, automatic OSI techniques are necessary to identify the scripts in the given input document which can further be sent to their corresponding OCR engines. Fig. 1 shows some examples of mixed-script documents.

Script identification is a key step in document image analysis especially when the environment is multi-script and multilingual. It also serves as an essential precursor for recognizing the language in which a document is written. This is necessary for further processing of the document, such as searching, indexing or translation. For scripts used by only one language, script identification itself accomplishes language identification. For scripts shared by many languages, script recognition acts as the first level of classification followed by language identification within the script.

India is a highly multilingual country with 22 constitutionally recognized languages. Besides these, hundreds of other languages are used in India, each one with a number of dialects. The officially recognized languages are *Hindi*, *Bengali*, *Punjabi*, *Marathi*, *Gujarati*, *Oriya*, *Sindhi*, *Assamese*, *Nepali*, *Urdu*, *Sanskrit*, *Tamil*, *Telugu*, *Kannada*, *Malayalam*, *Kashmiri*, *Manipuri*, *Konkani*, *Maithali*, *Santhali*, *Bodo*, and *Dogari*. *Hindi*, written in *Devnagari* script, is India's official language and has the most speakers, estimated to be more than 500 million. *Indic* scripts are a logical composition of individual script symbols and follow a common logical structure. This can be referred to as the "script composition grammar" which has no counterpart in any other set of scripts in the world. *Indic* scripts are written syllabically and are usually visually composed in three tiers where constituent symbols in each tier play specific roles in the interpretation of that syllable [1].

Automatic script identification in a multilingual environment is a challenging research problem over the last two decades [2]. Researchers have investigated OCR for a number of *Indic* scripts. However, most of these researches have been confined to the identification of isolated characters rather than the script. Unlike simple concurrence in *Roman* script, the *Indic* scripts are a composition of the constituent symbols in two dimensions. This implies that researchers first segment an *Indic* script word into its composite characters and then each composite character is decomposed into the constituent symbols or strokes that are finally recognized. Fig. 2 shows the general block diagram of the script identification system.

The process of identification of printed/handwritten script includes preprocessing and segmentation, feature extraction and recognition or classification processes. Preprocessing is necessary when the data or image is captured for further processing. Preprocessing is a method of enhancing the quality of an image for better understanding of the image. The choice of preprocessing method to be adopted on a document image depends on the type of application for which the image is used. The noise gets introduced in the document image during its acquisition and/or transmission over wired/wireless channels, as well as because of changing of some parameters of acquisition system in the OCR. Skew is an unavoidable distortion that is often introduced during scanning or