Information Sciences 314 (2015) 277–292

Contents lists available at ScienceDirect

Information Sciences

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

Multi-lingual date field extraction for automatic document retrieval by machine

Ranju Mandal [a,*], Partha Pratim Roy [b], Umapada Pal [c], Michael Blumenstein [a]

[a] School of Information and Communication Technology, Griffith University, Queensland, Australia
[b] Dept. of Computer Science & Engineering, Indian Institute of Technology, Roorkee, India
[c] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form 5 August 2014
Accepted 18 August 2014
Available online 2 September 2014

Keywords:
Robot reading
Robot retrieval of document
Date-based indexing
Handwritten date extraction
Date spotting
Multi-lingual documents

ABSTRACT

Robotic intelligence has recently received significant attention in the research community. Application of such artificial intelligence can be used to perform automatic document retrieval and interpretation by a robot through query. So, it is necessary to extract the key information from the document based on the query to produce the desired feedback. For this purpose, in this paper we propose a system for automatic date field extraction from multi-lingual (English, Devnagari and Bangla scripts) handwritten documents. The date is a key piece of information, which can be used in various robotic applications such as date-wise document indexing/retrieval. In order to design the system, first the script of the document is identified, and based on the identified script, word components of each text line are classified into month and non-month classes using word-level feature extraction and classification. Next, non-month words are segmented into individual components and labelled into one of text, digit, punctuation or contraction categories. Subsequently, the date patterns are searched using the labelled components. Both numeric and semi-numeric regular expressions have been used for date part extraction. Dynamic Time Warping (DTW) and profile feature-based approaches are used for classification of month/non-month words. Other date components such as numerals and punctuation marks are recognised using a gradient-based feature and Support Vector Machine (SVM) classifier. The experiments are performed on English, Devnagari and Bangla document datasets and the encouraging results obtained from the system indicate the effectiveness of the proposed system.

1. Introduction

Robotic intelligence has been applied in many fields to perform tasks which were previously performed by humans. At present various robot systems have been designed in many areas such as Engineering, Medical, and Industrial applications. Although there is a big effort in the robotics community in these areas [35,1], still to date, robots have been made little use of human readable text. To our knowledge, only a few pieces of work have undertaken towards realising a librarian robot to search and retrieve a book requested by a user [16,29,31]. The operation of such a robot starts when the user requests a book by its name or code, either through the Internet or by voice. The robot then locates the book in the library, extracts it and

* Corresponding author.

http://dx.doi.org/10.1016/j.ins.2014.08.037
0020-0255/

then takes it to the user. Some works have also been proposed towards a robotic system that is capable of detecting and reading wild text, a rich source of semantic information indigenous to man-made environments [15].

Nowadays huge collections of documents are available and automatic indexing or retrieval of relevant documents from these huge collections needs lot of human effort. For example, date-wise sorting of administrative documents of an organisation may require a lot of manual effort and also it is time consuming. So it is necessary to think of robotic applications to perform such tasks. Keeping this in mind, in this work we propose a date-based document indexing system for a robotic application. To our knowledge, there is no quality work on the application of robotics to document analysis such as document/form reading and document indexing by a robot.

In this paper we propose a system for automatic date field extraction from multi-lingual (English, Devnagari and Bangla scripts) handwritten documents. Date is a useful piece of information and it could be used as a key in various applications, for e.g., date-based document searching and indexing of document repositories such as administrative documents, historical archives and postal mail. Also, automatic extraction of date information is a challenging task due to different date patterns (Numeric and Semi-numeric dates consisting of different lengths), writing styles of different individuals, touching characters and confusion of classification during identification of numerals, punctuation and texts, etc. In multi-lingual and multi-script countries such as India, retrieval of multi-script documents using the date pattern can be very effective. An Indian state generally uses three official languages. For example, the West Bengal State of India uses Bangla, Devnagari and English as official languages. Hence, a single document may contain one or more of these three scripts. Fig. 1 shows an Indian handwritten Bangla postal document containing an English (Roman) date. English script is widely used in India with popular handwritten Indic scripts such as Bangla and Devnagari in a single document. Often, date information is written using English numerals in these scripts. Moreover, Bangla documents having Devnagari and English (Roman) script is also common. Because of multi-lingual behaviour, our date field extraction method consists of three major tasks namely: script identification, month word detectors, and numerical field and date pattern extraction. The proposed tri-lingual date extraction method can handle five cases such as Devnagari documents with only Devnagari or English date fields, Bangla documents with only Bangla or English date fields and English documents with English date fields. Two types of date field patterns (Numeric and Alpha-numeric) are considered for all the above cases. Hence, the proposed date extraction process from such documents will be very useful in searching and interpreting documents.

Because of the different writing formats of dates, script analysis is a necessary and important task in date field extraction from multi-script documents. We have shown some examples of handwritten documents containing date information in Fig. 2. It is important to note that, the date patterns appear in different formats in a document. Some of these formats of a single English date are 12/03/2012 or 12th March, 2012 or March 12, 2012 or 12-03-2012 or 12.03.2012 or 12.03.12, etc. Examples of Bangla dates are ১৫/০৭/১৯৯০ in dd/mm/yyyy format, ১লা বৈশাখ ১৪১১, ২রা আশ্বিন ১৪১২, ৪ঠা বৈশাখ ১৪১৪, ৫ই আষাঢ় ১৪১০, ২১ শে বৈশাখ ১৪১৪, ১লা জানুয়ারী ১৯৯০, ২রা ফেব্রুয়ারী ১৯৯০ in dd month yyyy format and ১৬-০৭-১০ in dd-mm-yy format, etc. Some sample formats of Devnagari dates are १२/०८/२००९ in dd/mm/yyyy format, २३ शे जनवरी, २०१२ in dd month yyyy format, etc.

1.1. Related works

Date field extraction from multi-lingual documents consists of three major tasks: script identification, month word identification, numerals and punctuation (e.g. slash '/', hyphen '-' and period '.') identification. To the best of our knowledge, there is no work on date field extraction from multi-script documents. However, to get the idea about the state-of-the-art work in script identification, word spotting, numerical field extraction and date field recognition some classical methods as well as a few recently published approaches are presented here.

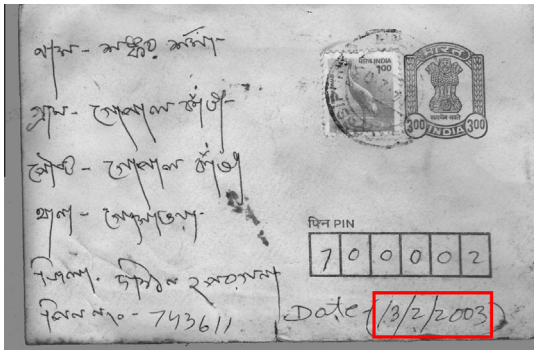

Fig. 1. Indian postcard containing a handwritten date field marked by a red rectangle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)