International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.5., 2015
(Sept.-Oct.)

RESEARCH ARTICLE

ISSN: 2321-7758

A STUDY ON IDENTIFICATION OF LANGUAGES AND ENCODINGS IN ONLINE DOCUMENT

ARCHITA[1], NIKITA SAGAR[2]
[1]M.Tech Student, [2]Head of the Department (CSE)
CSE Department, Indus Institute of Engineering and Sciences

International Journal of Engineering Research-online (IJOER)

ISSN:2321-7758
www.ijoer.in

ABSTRACT

Now-a-days the increase in popularity of portable computing devices such as PDAs and handheld computers, non keyboard based methods for data entry are receiving more attention in the research communities and commercial sector. The Pen-based and voice-based inputs are main options. Online Document can be written in any script. So, the first purpose is to identify the script, which is used to write the document Therefore, an online document analyzer must first identify the script before employing a particular algorithm for text recognition.

The multilingual, online document contains two or three scripts. Note that the writing style in English is mixed, with both cursive and handprint words. For lexicon-based recognizers, it may be helpful to identify the specific language of the text if the same script is used by multiple languages. A specific script like Roman may be used by multiple languages such as English, German and French.

Keywords: Scripts, Multilingual language, Analyser.

1. INTRODUCTION

The web contains text in many languages and scripts or we can say online documents may be written in different languages and scripts. Most of the text recognition algorithms are designed to work with a particular script and treat any input text has being written only in the script under consideration. Therefore, an online document analyzer must first identify the script before employing a particular algorithm for text recognition. Fig.1 shows an example of a document page containing six different scripts. . A specific script like Roman may be used by multiple languages such as English, German and French. The six scripts considered in this work, named Arabic, Cyrillic, Devnagari, Han, Hebrew, and Roman, cover the languages used by a majority of the world population (see Fig. 1). Handwriting recognition refers to the ability of a computer to receive intelligible written input. This paper represents

1. Allows the users to write a new script and then save it.
2. Users can write a script and then system can recognise it.
3. Users can view all the scripts saved in the database.
4. Users can delete any/all the saved scripts.

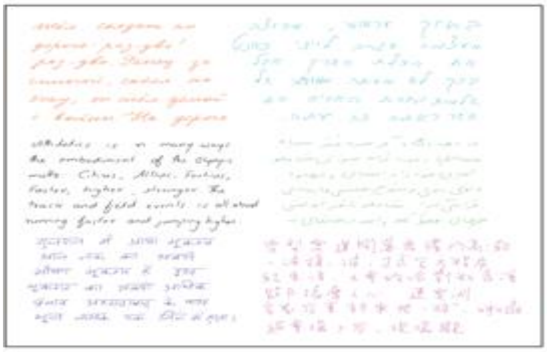


Fig1: multilingual document containing Cyrillic, Hebrew, Roman, Arabic, Devnagari, and Han scripts.

2. Literature Survey

Anoop M. Namboodiri, and Anil K. Jain, [1] present automatic identification of handwritten script facilitates many important applications such as automatic transcription of multilingual documents and search for documents on the Web containing a particular script. This paper proposes a method to classify words and lines in an online handwritten document into one of the six major scripts: Arabic, Cyrillic, Devnagari, Han, Hebrew, or Roman. The classification is based on 11 different spatial and temporal features extracted from the strokes of the words.

A.L. Spitz,[2] present language may refer either to the specifically human capacity for acquiring and using complex systems of communication, or to a specific instance of such a system of complex communication. The scientific study of language in any of its senses is called linguistics. The approximately 3000 languages that are spoken by humans today are the most salient examples, but natural languages can also be based on visual rather than auditory stimuli, for example in sign languages and written language

Anil Kumar Singh and Jagadeesh Gorla[3] Language identification becomes an important problem in the electronic world of many languages (Gordon 2005), even more so when multiple languages are mixed up in one document. Monolingual identification has been attempted by many researchers and it is now considered by many to be an almost solved problem. But multilingual identification has been rarely attempted. This is partly due to the fact that for a long time most of the documents on the Internet were monolingual.

`Johnson's method (Stephen 1993)[4] was based on characteristic 'common words' of each language. This method assumes unique words for each language. In practice, the test string might not contain any unique words

3. Identification of Script

Most of the published work on automatic script recognition deals with offline documents, i.e., documents which are either handwritten or printed on a paper and then scanned to obtain a two-dimensional digital representation. The proposed method uses the features of connected components to classify six different scripts (Arabic, Cyrillic, Devnagari, Han, Hebrew, Roman) and reported a classification accuracy of 88 percent on document pages.

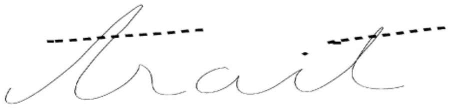
The properties of six different scripts used in this project are listed below.

3.1 Arabic: Arabic is written from right to left within a line and the lines are written from top to bottom. A typical Arabic character contains a relatively long main stroke which is drawn from right to left, along with one to three dots. The character set contains three long vowels. Short markings (diacritics) may be added to the main character to indicate short vowels.

3.2 Cyrillic: Cyrillic script looks very similar to the cursive Roman script. The most distinctive features of Cyrillic script, compared to Roman script is:

I. Individual characters, connected together in a word, form one long stroke.

II. The absence of delayed strokes.



The word "trait" contains three delayed strokes, shown as bold dotted lines here.

Delayed strokes cause movement of the pen in the direction opposite to the regular writing direction.

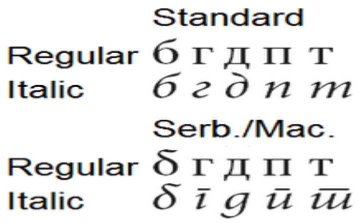
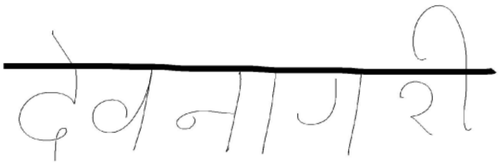


Fig 2: Standard Cyrillic letters.

3. 3 Devnagari: The most important characteristic of Devnagari script is the horizontal line present at the top of each word, called "Shirorekha". These lines are usually drawn after the word is written and hence are similar to delayed strokes in Roman script. The words are written from left to right in a line.



The word "devnagari" written in Devnagari script. The Shirorekha is shown in bold.

Devanāgarī is part of the Brahmic family of scripts of India, Nepal, Tibet, and South-East Asia. It is a descendant of the Gupta script, along with Siddham and Sharada

3.4 Han: Characters of Han script are composed of multiple short strokes. The strokes are usually drawn from top to bottom and left to right within a character. The direction of writing of words in a line is either left to right or top to bottom.

In many world languages, literacy has been promoted as a justification for spelling reforms. The People's Republic of China issued its first round of official character simplifications in two documents, the first in 1956 and the second in 1964. In the 1950s and 1960s, while confusion about simplified characters was still rampant, transitional characters that mixed simplified parts with yet-to-be simplified parts of characters together appeared briefly, then disappeared.

3.5 Hebrew: Words in a line of Hebrew script are written from right to left and, hence, the script is temporally similar to Arabic. The most distinguishing factor of Hebrew from Arabic is that the strokes are more uniform in length in the former.

3.6 Roman: Roman script has the same writing direction as Cyrillic, Devnagari, and Han scripts. We have already noted the distinguishing features of these scripts compared to the Roman script. In addition, the length of the strokes tends to fall between that of Devnagari and Cyrillic scripts. The features are extracted either from the individual strokes or from a collection of strokes. Here, we describe the features and their method of computation for each of the forecoming features an example follows shortly.

Fig 3: Roman Script Printed on Stone

The most important characteristics of online documents are that they capture the direction of strokes while writing the document. This allows us to analyze the individual strokes and use the additional direction for both script identification as well as text recognition. To identify the script used in the document, stroke properties as well as the spatial and temporal information can be used. The advantages of the proposed system are as follows:

1. Online documents are the main consideration.
2. Both spatial and temporal characteristics of the document are used to identify the scripts.
3. Easy segmentation of foreground and background of the online document.
4. Fewer error rate.

4. Multilingual Document

Today, Language identification is the main issue, even more when multiple languages are mixed up in one document. This paper present the multilingual language identification, it involves three parts. The first part is monolingual identification (Fig 4) Many methods with very high precision are available for this part. The second part is language enumeration i.e., finding out what languages are present in the document. The third part is segment identification i.e., identifying the language of segments of text in the document. If the segments are assumed to be single words, we can further divide the problem into word type identification and word token identification. In this first work on formulating the problem of multilingual language identification and solving it in a systematic way, we propose a method to solve the language enumeration and segment identification.

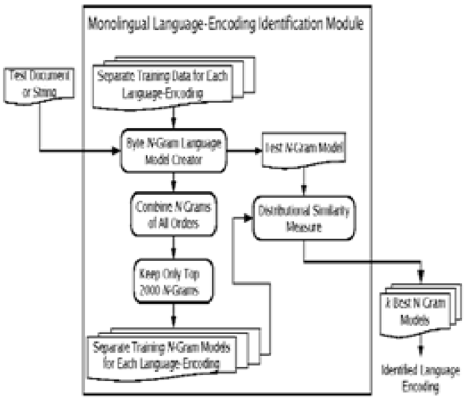

Monolingual Language-Encoding Identification Module
Test Document or String
Separate Training Data for Each Language-Encoding
Byte N-Gram Language Model Creator
Test N-Gram Model
Combine N-Grams of All Orders
Distributional Similarity Measure
Keep Only Top 2000 N-Grams
k Best N-Gram Models
Separate Training N-Gram Models for Each Language-Encoding
Identified Language Encoding

Fig 4: Monolingual Language Identification

Language Enumeration: Two algorithms are used for language enumeration:

- Mono K Best
- language Enumerator

The mono K Best algorithm returns K best n gram models or language encodings for a given word type or token. The language Enumerator algorithm returns the m best language encodings for a given document.

Word Type Identification: Once the best possible m language-encodings have been identified for the document, we can simply use the monolingual identifier to tag the language-encoding of each word type. The important point is that we only have to discriminate between m classes and m will be usually only 2 or 3.

Word Token Identification: In the current work, we assign the language-encoding class of a word token to be the same as that of the word type of which it is an instance. In other words, we are not taking the context of the token into account. We plan to explore how context can be used to improve token identification.

5. Future Scope

This paper can be implemented to identify the language and scripts using Java. In future, the user can develop script identification algorithm to recognize six scripts in any online document. The script classification algorithm can also be extended to do page segmentation, when different regions of the handwritten text are in different scripts.

Conclusion

The aim is to facilitate text recognition and to allow script-based retrieval of online handwritten documents. The classification is done at the word level, which allows us to detect individual words of a particular script present within the text of another script. This paper has been presented review of multilingual language identification.

REFERENCES

[1]. Anoop M. Namboodiri, and Anil K. Jain, "Online Handwritten Script Recognition" Ieee Transaction On Pattern Analysis And Machine Intelligence, Vol. 26, no. 1, January 2004.

[2]. A.L. Spitz, "Determination of the Script and Language Content of Document Images," IEEE Trans. Pattern Analysis and Machine Intelligence,

[3]. Anil Kumar Singh and Jagadeesh Gorla, "Identification of Languages and Encodings in a Multilingual Document", in Language Technologies Research Centre

[4]. A.K. Jain and Y. Zhong, "Page Segmentation Using Texture Analysis," *Pattern Recognition,* vol. 29, pp. 743-770, May 1996.

[5]. U. Pal and B.B. Chaudhuri, "Script Line Separation from Indian Multi-Script Documents," *Proc. Fifth Int'l Conf. Document Analysis and Recognition,* Sept. 1999.

[6]. C.Y. Suen, S. Bergler, N. Nobile, B. Waked, C.P. Nadal, and A. Bloch, "Categorizing Document Images Into Script and Language Classes," *Proc. Int'l Conf. Advances in Pattern Recognition,* pp. 297-306, Nov. 1998.

[7]. J.J. Lee and J.H. Kim, "A Unified Network-Based Approach for Online Recognition of Multi-Lingual Cursive Handwritings," Proc. Fifth Int'l Workshop Frontiers in Handwriting Recognition, pp. 393-397, Sept. 1996.

[8]. C.L. Tan, P.Y. Leong, and S. He, "Language Identification in Multilingual Documents," *Proc. Int'l Symp. Intelligent Multimedia and Distance Education,* Aug. 1999.

[9]. G.S. Peake and T.N. Tan, "Script and Language Identification from Document Images," *Proc. Third Asian Conf. Computer Vision,* pp. 96-104, Jan. 1998.