International Journal of Engineering Research and Technology.
ISSN 0974-3154 Volume 6, Number 1 (2013), pp. 171-179
© International Research Publication House
http://www.irphouse.com

Script Identification for Document Image Retrieval: A Survey

Madhushree B.[1], Nagashree T.S.[2] and Thanuja C.[3]

[1]Dept. of ISE, Sambhram Institute of Technology, Bangalore–560097.
E-mail: madhushreeb91@gmail.com
[2]Dept. of ISE, Sambhram Institute of Technology, Bangalore–560097.
E-mail: nagashreet@yahoo.com
[3]Dept. of ISE, Sambhram Institute of Technology, Bangalore–560097.
E-mail: thanu.reddy1985@gmail.com

Abstract

In recent years there are many multimedia documents captured and stored with the advances in computer technology and hence the demand for recognizing and retrieval of such documents has increased tremendously .In such environment the large volume of data and variety of scripts make manual identification unworkable. In such cases the ability to automatically determine the script ,and further the language of a document would reduce the time and cost of document handling. So the development of script identification from multilingual document image systems and then retrieving document image by matching with a query image (input image) has become an important task.
In this paper, we present a survey of methods developed to identify the script in document images automatically without manual intervention and then to access the document images by matching with the input query image,

General Terms: Script identification, Document image retrieval, Language identification Techniques, Phase based image matching, Feature extraction.

Introduction

Script identification is a key step that arises in document image analysis especially when the environment is multi script and language identification is required to identify the different language that exists in the same script.
Script identification has received relatively little attention in the document analysis field because one can normally deduce a document's script from its country of origin, or by examining the document. We are concerned with environments in which the

volume and variety of scripts makes such manual identification unworkable. For example, an office might have a large volume of incoming mail and reports in a variety of scripts that need to be converted to a character representation and routed to the appropriate reader. In such cases the ability to automatically determine the script, and further the language of a document, would reduce the time and cost of document handling. Script identification that runs accurately on a small number of characters could also be used to segment multiscript documents prior to language identification. Script identification facilitates many applications such as searching for a specific document in large digital libraries, searching online archives of document image containing a particular script, etc.

Document image retrieval is a very attractive field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. Good documents essentially play an important role in our day to day life. Complex documents present a great challenge to the field of document recognition and retrieval. With storage becoming cheaper and imaging devices becoming increasingly popular, efforts are on the way to digitize and archieve large quantity of multimedia data(text, audio, image and video).In response, extensive research is being carried out to make the digital content accessible to users through indexing and retrieval of relevant documents from such collection of images ,text ,audio ,video and audio. Most digital libraries aim at archiving books that are not available online. Success of text image retrieval systems mainly depends on the performance of identification and retrieval techniques used.

The combined presence of script identification and Document image retrieval would give an efficient way for document handling without manual efforts of human.

Proposed System

An efficient mechanism for retrieval of a Kannada word image from a bilingual document image collection is proposed. This involves four phases namely 1)Pre-processing 2)Query image formulation 3)Segmentation and Script identification and 4) Matching and retrieval.

Step 1:Pre-processing

Pre-processing involves preparing the source image for recognition of Kannada language. The source image is converted to binary image. This helps in comparison of input image with the query image.

Step 2: Query image formulation

Query image has to be formulated from the query word given as input to the system. English text entered, is translated to Kannada and converted as image by rendering. Convert this query image to binary image. This helps in comparison of input image with the query image.

Step 3: Segmentation and Script identification

Dilation helps in differentiating two words delimited by a space which is prerequisite for word level segmentation. Then identify the Kannada language by extracting vertical and horizontal line features from the bilingual image documents with the help of generated knowledge base.

Step 4: Matching and Retrieval

This is the stage where the documents matching with the search criteria are retrieved. Phase based image matching technique is used for matching query word with source image. If the matching score of query and source image is more than the threshold then words are matching, and the word will be highlighted in the document.

A survey of the techniques available for language identification and for image retrieval are discussed in this paper and organized as 3. A brief introduction to script identification is explained, which is followed by techniques under script identification in section 3.1 and previous related work in section 3.2. Further in section 4 introduction for document image retrieval is explained, which is followed by techniques under document image retrieval in section 4.2 and previous related work in section 4.3.Finally the paper in concluded in section 5.

Script Identification

Script Identification approaches can be broadly classified into two categories namely, local and global approaches. The local approaches analyze a list of connected components (Line, word, char) in the document images, to identify the script (or class of script). Here, success of Script classification mainly depends on the character segmentation or connected component analysis (Maximal region of connected pixels). In contrast, global approaches employ analysis of regions (block of text) comprising at least two lines (or words) without finer segmentation. Moreover, local approaches are slower when compared with the global approach

Few methods developed for script identification:

Cluster based method

In this method they developed a set of representative symbols (templates) for each script by clustering textual symbols from a set of training documents and representing each cluster by its centroid. Textual symbols" include discrete characters in scripts such as Cyrillic, as well as adjoined characters, character fragments, and whole words in connected scripts such as Arabic. To identify a new document's script, the system compares a subset of symbols from the document to each script's templates, screening out rare or unreliable templates, and choosing the script whose templates provide the best match.[1]Judith Hochberg, *et. al.* ,applied this technique for automatic script identification from images.

Method based on analysis of connected component profiles

The approaches based on connected component analysis generally use the intrinsic Morphological characteristics of the character sets or strokes of each script. It discovers frequent character or word shapes in each script by means of cluster analysis, then looks for instances of these in new documents. The system develops a set of representative Textual symbols (templates), defined as connected components, for each script by clustering textual symbols from a set of training documents and representing each cluster by its centroid. [2]Lijun Zhou1, *et al.* applied this method for Bangala/English script identification

Statistical identification technique

In this technique they identify scripts based on the density and distribution of vertical character runs. For each script studied, a script template is first constructed through a learning process. The script of the query image is then determined according to the distances between the query document vector and multiple learned script templates. The method first classifies the script into two broad classes: Han-based and Latin (or Roman)-based. This classification is based on the spatial relationships of features corresponding to the upward concavities. [9]A. Lawrence Spitz,*et al.*applied this technique to handle 13 Latin based languages, [7]Shijian Lu and Chew Lim Tan used this technique for script identification in degraded and disorted documents.

Prototype classification method

In this technique they identify scripts based on the density and distribution of vertical character runs. For each script studied, a script template is first constructed through a learning process. The script of the query image is then determined according to the distances between the query document vector and multiple learned script templates. The method first classifies the script into two broad classes: Han-based and Latin (or Roman)-based. This classification is based on the spatial relationships of features corresponding to the upward concavities. [9]A. Lawrence Spitz,*et al.*applied this technique to handle 13 Latin based languages, [7]Shijian Lu and Chew Lim Tan used this technique for script identification in degraded and disorted documents.

Support vector machines

SVM is a powerful tool for binary classification .In machine learning, support vector machines (SVMs, also support vector networks]) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other.[3]Ying-Ho Liu *et al.* proposed this technique in his paper for language identification of character images.

Visual discriminating Features

Language identification is one of the vision application problems. Generally human system identifies the language in a document using some visible characteristic features such as texture, horizontal lines, vertical lines, which are visually perceivable and appeal to visual sensation. This human visual perception capability has been the motivator for the development of this technique by [5]M.C. Padma,*et. al.* ,for language identification of Kannada, Hindi and English text words. With this context, in this technique an attempt has been made to simulate the human visual system, to identify the type of the language based on visual clues, without reading the contents of the document.

Method on profile based feature

The distinct characteristic features for several scripts are thoroughly studied from the nature for both top and bottom profiles. The proposed model is trained to learn thoroughly the distinct features of each script. Experimentation conducted involved with number of text lines for learning and number of text lines for testing. And then the k-nearest neighbor classifier is used to classify the test sample. [8]M. C. Padma*. P. A. Vijaya proposed this technique for script identification from trilingual documents.

Hierarchical classification method

This method uses features consistent with human perception. Such Features are extracted from the responses of a multi-channel log-Gabor filter bank, designed at an optimal scale and multiple orientations. In the first stage, the classifier groups the scripts into five major classes using global features. At the next stage, a sub-classification is performed based on script-specific features. All features are extracted globally from a given text block which does not require any complex and reliable segmentation of the document image into lines and characters. Thus this method is efficient and can be used for many practical applications which require processing large volumes of data.[4] Gopal Datt Joshi, *et. al.* ,proposed this technique for efficient identification of script from 10 Indian scripts. This proposed system achieves an overall classification accuracy of 97.11% on a large testing data set.

Document Vectorization

This method generates vectors from the nine zones segmented over the characters based on their shape, density and transition features. Script is then determined by using Rule based classifiers and its sub classifiers containing set of classification rules which are raised from the vectors. System proposed by [6] Abirami.S1and Murugappan. S2 identifies scripts from document images even if it suffers from noise and other kinds of distortions.There Results from experiments, simulations, and human vision encounter that the proposed technique identifies scripts and numerals with minimal pre-processing and high accuracy. In future, this can also be extended for other scripts.

Previous Work on Script Identification

No.	Authors	Method Proposed	Applied on Languages
1	Lijun Zhou1, *et al.*[2]	Connected Component Profiles	Bangla/English
2	Ying-Ho Liu[+], *et al.*[3]	Machine learning techniques	English, Chinese, Japanese
3	Judith Hochberg, *et al.*[1]	Cluster-based	Arabic,American, Chinese, Burmese, Cyrllic, Devangiri, Greek, Japanese, Ethiopic, Herbew, Korean, Thai, Roman
4	A. Lawrence Spitz, *et al.*[9]	Content of Document Images{statistical}	Latin based languages
5	Gopal Datt Joshi, *et al.*[4]	Hierarchical classification	English,Kannada,Korean,Malayalam, Tamil,Devanagiri,Gujurati,Oriya,Urdu, Gurumukhi
6	M.C. Padma, , *et al.*[5]	Visual discriminating features	Kannnada,Hindi,English
7	Shijian Lu and Chew Lim Tan[7]	Statiscal identification technique	Latin based languages
8	M. C. Padma*. P. A. Vijaya[8]	Profile based features	Kannnada,Hindi,English
9	Abirami.S1and Murugappan. S2[6]	Document vectorization.	Tamil,English,Hindi

Document Image Retrieval

Document image retrieval is a very attractive field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. The primary task of processing these complex documents is to isolate the different contents present in the documents. Once the contents are separated out, then they can now be called as indexed documents which are ready to use for a image retrieval system. Few are the below mentioned techniques developed by researchers for document image retrieval.

Techniques

Inexact string matching technique

In this method each word image is represented by a primitive string .Then the inexact string matching technique is applied to measure the similarity between the two primitive strings generated from the word images, from the measure of similarity the matching of two images are determined and then matched document images are retrieved.

Content based document image retrieval

In this technique the search will analyze the actual contents of the image rather than. metadata(keywords, tags), the word- content refers to colors, shapes, textures or any other information within a document image, analyzing these contents within a document and comparing it with query document image, a matched document image is retrieved[13].

Word spotting

In this technique they propose a shape code based word-image matching (word-spotting) technique for retrieval of multilingual documents written in Indian languages. Here, each query word image to be searched is represented by a primitive shape code using (i) zonal information of extreme points (ii) vertical shape based feature (iii) crossing count (with respect to vertical bar position) (iv) loop shape and position (v) background information etc. Each candidate word (a word having similar aspect ratio and topological feature to the query word) of the document is also coded accordingly. Then, an inexact string matching technique is used to measure the similarity between the primitive codes generated from the query word image and each candidate word of the document with which the query image is to be searched. Based on the similarity score, we retrieve the document where the query image is found [11].

Local feature analysis

This method presents a fast, accurate and OCR-free image retrieval algorithm using local feature sequences which can describe the intrinsic, unique and page-layout-free characteristics of document images. With a simple preprocessing step, the local feature sequences can be extracted without print-core detection and image registration. Then an efficient coarse-to-fine common substring matching strategy is applied to do local feature sequence match. Beyond a single matching score, this approach can locate the matched parts word by word [12].

Dynamic Time Wrapping

Dynamic Time Wrapping (DTW) computes a sequence alignment score for finding the similarity of words .The use of the total cost of DTW as a distance measure is helpful to cluster together Word images that are related to their root word. DTW is a dynamic programming based procedure to align two sequences of signals and compute a similarity measure and hence with the results of similarity measure two document images are matched and are obtained [14].

Phase based matching using Fourier Transforms

This method is viewed as a mathematical prism that separates function into various components also based on frequency content. Fourier transforms lets us characterize a function by its frequency content. The phase information obtained from 2D DFT (Discrete Fourier Transform) of images contains important information of image representation. The phase-based image matching is successfully applied to sub-pixel image registration tasks for computer vision applications and image recognition tasks for biometric authentication applications. Hierarchical block matching using phase information, i.e, phase-based correspondence matching, can find the corresponding points on the input image from the reference points on the registered image with sub-pixel accuracy[15].

Previous work On Document Image Retrieval

Authors	Method Proposed	Applied
Yue Lu and Chew Lim Tan[11]	word spotting.	English
Jilin Li,Zhi Gang Tan,Yadong hue and Ning Le[12]	local feature analysis	Asian based documents and Latin based documents
Million Meshesha,C.V. Jawahar[13]	content based retrieval.	English, Amharic and Hindi documents
A. Balasubramanian[14]	Dyanamic time wrapping	English, Devanagiri,Telugu.
Koichi Ito Takafumi Aoki[15]	Phase based matching using fourier transforms	For face recognition

Conclusion

In this paper, we have attempted to provide a survey on techniques used previously for script identification and document image retrieval. How each of the above mentioned technique works was explained in this paper. Certain techniques have their own advantages as well as disadvantages, so each technique is unique in its own way, hence understanding all the above techniques, we have chosen visual discriminating feature technique for script identification and phase based Fourier transforms technique for document image retrieval, to implement our proposed system.

References

[1] Judith Hochberg et al,'Automatic Script Identification from Images Using Cluster-based Templates'

[2] Lijun Zhou *et. al.*, Bangla/English Script Identification Based on Analysis of Connected Component Profile

[3] Language Identification of Character Images Using Machine Learning Techniques Ying-Ho Liu┼, Chin-Chin Lin┼┼, and Fu Chang.

[4] Script Identification from Indian Documents Gopal Datt Joshi, Saurabh Garg, and Jayanthi Sivaswamy

[5] Language Identification of Kannada, Hindi and English text words through visual discriminating features M.C. Padma, DR.P.A Vijaya.

[6] Multilingual document images Abirami.S1and Murugappan.

[7] Script and Language Identification in Degraded and Distorted Document Images Shijian Lu and Chew Lim Tan

[8] Script identification from trilingual documents using profile based features m. C. Padma* P. A. Vijaya

[9] Determination of the Script and Language Content of Document Images Lawrence Spitz, Member, IEEE

[10] S. Abirami, 2Dr. D.Manjula A Survey of Script Identification techniques for Multi-Script Document Images

[11] Information Retrieval in Document Image Databases Yue Lu and Chew Lim Tan, Senior Member, IEEE

[12] Document Image Retrieval with Local Feature Sequences Jilin Li, Zhi-Gang Fan, Yadong Wu and Ning Le

[13] Matching word images for content-based retrieval from printed document images Million Meshesha · C. V. Jawaha

[14] Retrieval from Document Image Collections Balasubramanian, Million Meshesha, and C.V. Jawahar.

[15] Face Recognition Using Phase-Based Correspondence Matching Koichi Ito Takafumi Aoki, Tomoki Hosoi Koji Kobayashi.