IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. V (Mar – Apr. 2015), PP 105-109
www.iosrjournals.org

A survey on Script and Language identification for Handwritten document images

Prasanthkumar P V[1], Midhun T P[2], Archana Kurian[3]
[1](Computer Science & Engineering Department, Vimal Jyothi Engineering College,Kannur,Kerala India)
[2](Computer Science & Engineering Department, Vimal Jyothi Engineering College,Kannur,Kerala India)
[3](Computer Science & Engineering Department, Vimal Jyothi Engineering College,Kannur,Kerala India)

***Abstract**: Offline Script and language identification serves as a forerunner for a multi lingual Optical Character Recognizer (OCR). OCR is a software used for digitizing the handwritten or printed document mages. Most of the OCRs are designed for dealing with a single script. So it can't convert a document which contain more than one scripts. A country having multi lingual culture like India, multi lingual OCR is an essential software requirement for automation of handwritten documents processing. Script Identification consists of three steps Pre-processing, Features extraction and Classification. Based on the features extracted from the document images the classifier discriminates the script of the documents. This survey is an overview of different script identification methods for handwritten document images. They are classified into different categories based on the features and classification algorithm used.*

***Keywords:** Handwritten documents, Optical Character Recognition, Script Identification*

I. Introduction

Script and Language identification is an interesting area of research in the domain of document processing. Optical Character recognition (OCR) is a process in which that optically scanned the paper document and then converted into computer processable electronics format. Many OCR algorithms have been developed over years are script specific in the sense that they can read characters written in one particular script only. It is essential to identify the Script and Language of the document in the multi-script environment before processing the document using OCR. Both printed and handwritten documents served as a medium of communication as well as a medium for recording facts. For example in Post Offices, Libraries, Railway Stations, Government Offices etc where we have to handle such documents which are written in multiple languages.

Script is defined as the graphic form of the writing system used to write statements expressible in language [1]. 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, Devanagari is used for writing a number of Indian languages like Sanskrit, Hindi, Konkani, Marathi, etc. Script identification relies on the fact that each script has

unique spatial distribution and visual attributes that make it possible to distinguish it from other scripts[1][2]. So, the basic task involved in script identification is to devise a technique to discover these features from a given document and then classify the documents script accordingly. In general, features necessary for recognition of different script characters depend on the structural properties, style and nature of writing which generally differs from one script to another.

This review on script and language identification for handwritten document images gives an overview of different methodologies developed so far. In section II challenges faced in the research of handwritten document images is presented. Section III discussed the review of the related works. Last section is for comparative analysis of the identified models as well as the inferences made from them.

II. Challenges

Script identification for handwritten document images is more challenging than printed document images [4]. These challenges made it impossible to directly apply algorithm for printed document to handwritten document images. Three main challenges are first, some scripts resemble each other more when handwritten than when printed. Second, handwriting styles are more diverse than printed fonts. Cultural differences, individual differences, and even differences in the way that people write at different times, enlarge the inventory of possible character and word shapes seen in handwritten documents. Third, problems typically addressed in pre-processing, such as ruling lines and character fragmentation due to low contrast, are common in handwritten documents due to the variety of papers and writing instruments used. Moreover Text lines in handwritten documents are curvilinear and the gaps between neighbouring words and lines are far from uniform. It is difficult to extend the methods to new languages, because they employ a combination of handpicked and trainable features and variety of decision rules.

III. Script Recognition Methodologies

D. Ghosh et al [1], divide the classification method into two broad categories based on the nature of the approaches and features used. They are structure-based and visual appearance based methods. Script identification methods that are based on extraction and analysis of connected components fall under the category of structure based methods. Visual appearance methods are often related to texture, a block of text corresponding to each script class forms a distinct texture pattern. Each of these categories further classified into page-wise, paragraph-wise, textline-wise and word-wise on the basis of the level at which they are applied. Script identification methods that use segment-wise analysis of character structure is also regarded as local approach. On the other hand, visual appearance based methods that are designed to identify script by analyzing the overall look of a text-block may be regarded as a global approach.

3.1 Linear Discriminant based

Hochberg, et al [4] proposed an algorithm for script and language identification from handwritten document images using statistical features based on connected component analysis. Documents are characterized by five connected component features which are relative vertical centroid, relative horizontal centroid, number of holes, sphericity and aspect ratio of the connected components in a document page. For each of the five connected component features, three document summary statistics mean, standard deviation, and skew are calculated. This created a fifteen-element vector for each document. A separate Fisher Linear Discriminant (FLD) is trained to separate each possible pair of script in the dataset. The classifier is tested through writer-sensitive cross validation and achieved a accuracy of 88% across six scripts Arabic, Chinese, Cyrillic, Devanagari, Japanese, and Roman. They have used the same method to discriminating Roman script English and German documents with 85% accuracy.

3.2 K-Nearest Neighbor Based Techniques

Dhandra and Hangarge [5][7] used nearest neighbor and K-nearest neighbor (KNN) algorithms to classify word images belonging to Kannada, Roman, and Devanagari scripts. By decomposing the word image in two directions at two levels using morphological transformation seven global features are obtained and other three dominant local features are computed based on connected components. These features are passed to KNN classifier for classification of the scripts. The Average and maximum recognition accuracy achieved is 96.05% and 99% respectively. This algorithm is insensitive to writing style, ink, size, noise and characters slant. Recently a Gaussian Mixture Model(GMM) was introduced by Mallikarjun Hangarge [6] to identify the script of handwritten words of Roman, Devanagari, Kannada and Telugu scripts. GMM is modeled using a set of six novel features derived from directional energy distributions of the underlying image. The standard deviation of directional energy distributions are computed by decomposing an image matrix into right and left diagonals. Furthermore, deviation of horizontal and vertical distributions of energies is also built-in to GMM. The model is tested on biscript, tri-script and multi-script level and achieved script identification accuracies in percentage as 98.7, 98.16 and 96.91 respectively.

To identify eight major scripts, namely Latin, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Tamil, and Telugu at block level, Rajput and Anita [8] proposed a scheme based upon features extracted using Discrete Cosine Transform (DCT) and Discrete Wavelets Transform (DWT). A KNN classifier is then employed for the identification purpose. They achieved an average accuracy rate of 96.4% for tri-script documents images.

Hiremath et al [9] proposed an approach for script identification using texture features. The scripts considered for the work was Bangla, Latin, Devanagari, Kannada, Malayalam, Tamil, Telugu, and Urdu. The texture features are extracted using the co-occurrence histograms of wavelet decomposed images. The correlation between the sub-bands at the same resolution exhibits a strong relationship and is significant in characterizing a texture. A KNN classifier is used for the identification of scripts. Average classification accuracy achieved is 97.5% for a single writer document with full text coverage, which decreases slightly with the increase in angle of orientation and decrease significantly with the increase in the number of writers.

3.3 Steerable Pyramid based

A method for Arabic and Latin text-block differentiation in both printed and handwritten scripts was proposed in Benjelil et al[10]. Literature explained an accurate and suitable designed system for script identification at word level which is based on steerable pyramid transform. The Steerable Pyramid(SP)[11], is a linear multi-scale, multi-orientation image decomposition, that provides a useful front-end for image processing and computer vision applications. The SP can capture the variation of a texture in both intensity and orientation. The overall Handwritten Arabic and Latin identification rate obtained is about 97.5%

3.4 Gabor function-based

Gabor function-based script recognition schemes have shown good performance, their application is limited to machine-printed documents only. Variations in writing style, character size, and inter-line and inter-word spacing make the recognition process difficult and unreliable when these techniques are applied directly on handwritten documents. Therefore, it is necessary to pre-process the document images prior to the application of Gabor filter so as to compensate for the different variations present. This has been addressed in the texture-based script identification scheme proposed in [12]. In the pre-processing stage, the algorithm employs denoising, thinning, pruning, mconnectivity, and text size normalization in sequence. Texture features are then extracted using a multichannel Gabor filter. Finally, different scripts are classified using fuzzy classification. In this proposed system, an overall accuracy of 91.6% is achieved in classifying handwritten documents written in four different scripts, namely Latin, Devanagari, Bengali and Telugu.

3.5 Neural Network Based Techniques

Fractal-based features, busy-zone based features, and Topological features along with an ANN classifier is used for word-wise Bangla, English, and Devanagari scripts identification by Roy and Majumder [13]. Multi-Layer Perceptron (MLP) based classifier for script separation, trained with 8 different word level features. Two equal sized data-sets, one with Bangla and Roman scripts and the other with Devanagri and Roman scripts, were prepared for the system evaluation and achieved accuracies of 99.29% and 98.43%.

Roy and Das[14] proposed a scheme for identification of scripts written by any of the 6 official languages via Bangla, Devnagari, Malayalam, Urdu, Oriya and Roman script of India using different features, namely, fractal dimension based features, component based features, topological features, and a Neural Network (NN) classifier is used for script identification. The scheme is independent of text size and there is no need for any normalization. The overall accuracy of the developed system was 89.48 percentages on the test set without rejection. S M Obaidullah et al [15] have proposed a work which is on six official languages of India. They have used very simple and efficient features at document level categorized under Abstract/Mathematical features, Structure based features and Script dependent features. Series of classifiers is used for classification. Overall accuracy of the proposed system is at present 92.8n% on the test set without rejection.

3.6 Support Vector Based Techniques

To identify the script of handwritten postal codes, Basu et al[16] grouped similar shaped digit patterns of Bangla, Urdu, Latin, and Devanagari in 25 clusters. A script independent unified support vector machine (SVM) based pattern classifier is then designed to classify the numeric postal codes into one of these 25 clusters. Based on these classification decisions a rule-based script assumption engine is designed to assume the script of the numeric postal code.

IV. Comparative Analysis

Most of the available works on recognition of Indian scripts are based on small databases collected in laboratory environments. Since the experiments were conducted independently using different data-sets so they do not reflect the comparative performance of these methods. The Table 1 shown below summarizes some of the benchmark work in script recognition handwritten document images. Various script features and classifiers used by different researchers are also listed in the table.

4.1 Better Pre-Processing for Higher Accuracy

The pre-processing steps generally comprises of binarization, gray level normalization, foreground and background noise removal, size normalization, removal of irrelevant information, skew and slant corrections, etc. Handwritten character samples are usually collected from individuals belonging to different age groups having different writing styles, professions, state of mind, writing medium, etc. So, pre-processing of the character image is an important step before the feature extraction and classification step. Better preprocessing leads to high performance

4.2 Structure-based Features

Structure-based features like character height distribution, character bounding box profiles, horizontal projections and several other statistical features do not depend on the document quality and resolution but on the overall size of the connected components [1]. However, these features are not invariant to character size and font and offer high performance only in separating distinctly different oriental scripts from others. Several different structural features like character geometry, occurrence of certain stroke structures and structural primitives, stroke orientations, measure of cavity regions, side profiles, etc that directly relate to the character shape have also been used for script characterization. One disadvantage with structure-based methods is that they require complex pre-processing involving connected component extraction. Also, extraction of structural

features is highly susceptible to noise and poor quality document images. Presence of noise or significant image degradation adversely affects the location and segmentation of these features, making them difficult or sometimes impossible to extract. Scripts having similar character shapes may be distinguished by their visual appearances.

4.3 Gabor filter based

Gabor filter offers a powerful tool to extract out visual attributes from a document. This has motivated many researchers to employ Gabor filter for script determination. Since texture feature gives the general appearance of a script, it can be derived from any script class of any nature. Accordingly, this feature may be considered a universal one. The discriminating power of a multichannel Gabor filter can be varied by having more channels with different radial frequencies and closely spaced orientation angles. Thus, this system is flexible compared to all other methods and can be effectively used in discriminating scripts that are quite close in appearance. The main criticism with this approach is that it cannot be applied with confidence to small text regions as in word-wise script recognition. Also, Gabor filters are not capable of handling variations in script size and font, inter-line spacing, etc [3].

4.4 Data collection

One major concern with most of the reported works in script recognition is the lack of any comparative analysis of the results. Experimental results given for every proposed method have not been compared with other benchmark works in the field. Moreover, the datasets used in experiments are all different. This is mainly due to the lack of availability of a standard database for script recognition research. Consequently, it is hard to assess the results reported in the literature. Hence, a standard evaluation test-bed containing documents written in only one script type as well as multi-script documents with mix of different scripts within a document is necessary. One important consideration in selecting the data-set for a script class is that it should reflect the global probability of occurrence of the characters in texts written in that particular script. Another problem of concern is for languages that constantly undergo spelling modifications over the years

V. Conclusion

This paper presents a comprehensive survey on script and language identification for offline handwritten document images. Different approaches proposed are classified into two categories, structure-based and visual appearance based. Most of the works are in Devanagari and Bangla scripts. To the best of my knowledge no works are reported for identifying south Indian scripts and languages using support vector machines up to word level.

Table: I Script Identification Methods

Research/ Authors	Features selected	Classifier Used	Script Classified	Result obtained
Hochberg et al	Relative Y centroid, Relative X centroid , number of white holes, sphericity, and aspect ratio	Linear Discriminant Analysis	Arabic, Chinese, Cyrillic, Devanagari, Japanese, Latin	Accuracy of 88% is achieved
Hangare and Dhandra	Vertical stroke density, Horizontal Stroke Density, Right diagonal stroke density, Left diagonal stroke density	KNN Classifier	English, Devanagari, Urdu	Accuracy of 97.83% (English), 93.00% (Devanagari) and 95.78% (Urdu)
Rajput and Anitha	Discrete cosine Transform and Discrete wavelet transform	KNN Classifier	Latin, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Tamil, and Telugu	Accuracy of 98% (KEH), 99.2%(MEH), 93%(PEH), 99.2%(TEH), 90% (GEH) and 99% (TeEH)
Hiremath et al	Texture features extracted by using DWT	KNN Classifier	Bangla, Latin, Devanagari, Kannada, Malayalam, Tamil, Telugu, and Urdu	Accuracy of 97.5% is achieved
Roy and Majumder	Fractal-based features, busy-zone based features, and Topological features	ANN classifier	Bangla, English, and Devanagari	Accuracy of 99.29% (BR) and 98.43%(DR)
Roy and Das	Fractal dimension based features, component based features,	Neural Network	Bangla, Devanagari, Urdu, Malayalam, Oriya and Roman	Accuracy of 89.48%

	topological features	Classifier		
Mohamed Benjelil et al	Mean, Standard Deviation, Kurtosis, Homogeneity, Energy, Correlation	Steerable Pyramid based	Arabic and Latin	Accuracy of 97.5%
Vivek Singhal et al	Gabor filter-based texture feature	Fuzzy Classifier	Devanagari, Bengali, Telugu, Latin	Accuracy of 91.6%
Sk.Md Obaidullah et,al	Structure Based, Mathematical based and Script dependent features	MLP Classifier	Bangla, Devanagari, Urdu, Oriya, Malayalam and Roman	Accuracy of 92.8%

References

Proceedings Papers:

[1]. Debashis Ghosh, Tulika Dube, and Adamane P Shivaprasad Script recognitiona review, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (2010).

[2]. Pal, Umapada and Jayadevan, Ramachandran and Sharma, Nabin Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques, ACM, March 2012

[3]. D GHOSH and AP SHIVAPRASAD, Handwritten script identification using probabilistic approach for cluster analysis, Journal of the Indian Institute of Science 80 (2013)

[4]. Judith Hochberg, Kevin Bowers, Michael Cannon, and Patrick Kelly, Script and language identification for handwritten document images, International Journal on Document Analysis and Recognition 2 (1999)

[5]. BV Dhandra and Mallikarjun Hangarge, Global and local features based handwritten text words and numerals script identification, Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on, vol. 2, IEEE, 2007

[6]. Mallikarjun Hangarge, Gaussian mixture model for handwritten script identification, arXiv preprint arXiv:1303.2751 (2013).

[7]. Mallikarjun Hangarge and BV Dhandra, Offline handwritten script identification in document images, International Journal of Computer Applications 4 (2010)

[8]. GG Rajput and HB Anita, Handwritten script recognition using dct and wavelet features at block level, IJCA Special Issue on: Recent Trends in Image Processing and Pattern Recognition, RTIPPR (2010)

[9]. PS Hiremath, S Shivashankar, Jagdeesh D Pujari, and V Mouneswara, Script identification in a handwritten document image using texture features, Advance Computing Conference (IACC), 2010 IEEE 2nd International, IEEE, 2010

[10]. Mohamed Benjelil, Slim Kanoun, R my Mullot, and Adel M Alimi, Arabic and latin script identification in printed and handwritten types based on steerable pyramid features, Document Analysis and Recognition, 2009. ICDAR09. 10th International Conference on, IEEE, 2009

[11]. Eero P Simoncelli and William T Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, ImageProcessing, 1995. Proceedings., International Conference on, vol. 3, IEEE, 1995

[12]. Vivek Singhal, Nishant Navin, and D Ghosh, Script-based classification of hand-written text documents in a multilingual environment, Research Issues in Data Engineering: Multi-lingual Information Management, 2003. RIDE-MLIM 2003. Proceedings. 13th International Workshop on, IEEE, 2003

[13]. Kaushik Roy and Kinshuk Majumder, Trilingual script separation of handwritten postal document, Computer Vision, Graphics & Image Processing, 2008.ICVGIP08. Sixth Indian Conference on, IEEE, 2008,

[14]. K Roy, S Kundu Das, and Sk Md Obaidullah, Script identification from handwritten document, Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on, IEEE, 2011

[15]. Kaushik Roy Sk Md Obaidullah, Supratik Kundu Das, A system for handwritten script identification from indian document, Journal of Pattern Recognition Research 8 (2013)

[16]. Subhadip Basu, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, and Dipak Kumar Basu, A novel framework for automatic sorting of postal documents with multi-script address blocks, Pattern Recognition 43 (2010)

[17]. Gao, Yangdong and Ding, Xiaoqing and Liu, Changsong, A Multiscale Text Line Segmentation Method in Freestyle Handwritten Documents Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011, IEEE

[18]. Nicolaou, Anguelos and Gatos, Basilios, Handwritten text line segmentation by shredding text into its lines, Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, IEEE, 2009.