A Review on Offline Handwritten Script Identification

D S Guru
Department of Studies in Computer Science, University of Mysore Mysore, Karnataka, India

M Ravikumar
Department of Computer Science, Kuvempu University

B S Harish
Department of Information Science and Engineering, S J College of Engineering Mysore, Karnataka, India

ABSTRACT
This paper presents a review of script identification on offline handwritten documents. Here we present the challenges involved in script identification, the applications of script identification, different methods used for extracting the features in different scripts and also the classifiers used for classification. Since, script identification plays an important role in analyzing the offline handwritten documents. we presented a brief overview and analysis of the existing methods.

Keywords
Handwritten Documents, Script Identification, OCR, Classifiers

1. INTRODUCTION
Script is defined as the graphic form of writing system which is used to express the written languages. Languages throughout the world are typeset in many different scripts. A script may be used by only one language or shared by many languages, 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., English, French, German and other European languages use different variants of the Latin alphabets, and so on. Languages use different scripts at different points of time and space. Malay language, which was earlier using jawi alphabets, is now using latin alphabets. Another example is Sanskrit that is mainly written in Devanagari in India but is also written in Sinhala script in Sri Lanka. Therefore there is an increased demand for handwriting recognition as lot of data, (such as addresses written on envelopes; amounts written on checks; names, addresses, identity numbers, and dollar values written on invoices and forms) were written by hand and they had to be entered into the computer for processing. Hence in this multilingual and multi-script world, Optical Character Recognition (OCR) systems need to be capable of recognizing characters irrespective of the script in which they are written. In general, recognition of different script characters in a single OCR module is difficult. This is because of features which are necessary for character recognition depend on the structural properties, style and nature of writing which generally differs from one script to another [8]. For example, features used for recognition of English alphabets are not good for recognizing Chinese alphabets. Another option for handling documents in a multi- script environment is to use a bank of OCRs (different OCR for different script) corresponding to different scripts. The characters in an input document can then be recognized reliably by selecting the appropriate OCR system from the OCR bank. However, it requires a priori knowledge of the script in which the document is written. Unfortunately, this information may not be readily available. At the same time, manual identification of the documents' scripts may be tedious and time consuming. Script identification is also useful in reading multi-script documents in which different paragraphs, text –blocks, text lines or words in a page are written in different scripts. Script recognition also helps in text area identification, video indexing and retrieval, and document sorting in digital libraries when dealing with multi-script environment [8].Script and language identification are important parts of the automatic processing of document images. A document's script (e.g., Cyrillic or Roman) must be known in order to choose an appropriate Optical Character Recognition (OCR) algorithm. Handwritten documents present three challenges for script identification. First, handwritten scripts are more difficult when compared to printed scripts. Second, handwriting styles are more diverse than printed fonts. Lastly, problems typically addressed in pre-processing, such as ruling lines and character fragmentation due to low contrast, are very challenging in handwritten documents [4]. India is a multi-lingual country where roman script is often used alongside different Indic scripts in a text document. To develop a script specific handwritten optical character recognition system, it is necessary to identify the scripts of handwritten text correctly. To develop a successful multi-lingual OCR system, separation or identification of different scripts is a very important step. In a multi-lingual country like India, it is an utmost essential for designing an OCR system. India has more than 22 official languages and 12 different scripts are used for these languages. Moreover English is taught and used largely almost all over India. Most of the published methodologies on automatic scripts identification are discussed on printed text documents [1]. Automatic handwritten script identification from document images facilitates many important applications such as sorting, transcription of multilingual documents and indexing of large collection of such images. To make a successful multilingual OCR, script identification is very essential before running an individual OCR system. Most of the published work has identified a number of approaches for determining the script/language of handwritten documents. They are classified into four categories (a) methods based on the analysis of connected components, (b) methods based on the analysis of characters, words and text lines, (c) methods based on the text blocks, (d) methods based on the analysis of hybrid information of connected components, text lines etc [5].Automatic script identification is a challenging research problem in a multilingual environment over the last few years. All existing works on automatic language identification are classified into either local approach or global approach. In local approach, the features are extracted from a list of connected components such as line, word and character, which are obtained only after segmenting the underlying document image. So, the success rate of classification depends on the effectiveness of the pre-processing steps. But, it is difficult to find a

common segmentation method that best suits for all the script classes. Due to this limitation, local approaches cannot meet the criterion as a generalized scheme. In contrast global approaches employ analysis of regions comprising of at least two text lines and hence fine segmentation of the underlying document into line, word and character, is not necessary. Consequently, the script classification task is simplified and performed faster with the global approach than the local approach. So, global schemes can best suited for a generalized approach to the script identification problem [10].In short, automatic script identification is crucial to meet the growing demand for electronic processing of volumes of documents written in different scripts. This is important for business transactions across Europe and orient, has great significance in a country like India which has many official state languages and scripts. Due to this there has been a growing interest in multi-script OCR technology during recent years [8]. Also, the existing methods work well for printed documents. Unfortunately, there are several limitations for handwritten script identification. However, only few attempts were made towards the handwritten script identification in the literature. Further, majority of the work attempted on handwritten script identification is based only one language. However, there is a great increase in the demand towards handwritten script identification using bilingual and trilingual documents. Since in India most of the official documents are trilingual (documents with English, Hindi and regional language like kannada), handling handwritten trilingual documents is a very challenging task. Thus the above challenges made us to work on trilingual handwritten documents, and in this work we have made an initial attempt to review some of the works done in handwritten script identification.

2. RELATED WORK

Hangare and Dhandra [5] proposed texture as a tool to determine the script of handwritten document image. It is based on the observation that text has a distinct visual texture. Experiments are carried out with KNN classifier by varying the number of neighbors (K= 3, 5, 7, 9, 11, 13, 15) and the performance of the algorithm is found optimal when K = 5 for text blocks and K= 3 for text lines respectively. Hochberg et al [4] have used linear discriminant analysis for classification of scripts in a document and tested using writer sensitive cross- validation. Extracted five connected component features viz., Relative Y centroid, Relative X centroid, Number of white holes, Sphericity and Aspect ratio. Documents were characterized by the mean, standard deviation and skew of five connected component features. Namboodiri and Jain [2] have used a SVM classifier to classify words and lines in an online handwritten document of different scripts; classification is based on spatial and temporal features of strokes. The features were extracted either from the individual strokes or from a collection of strokes. The extracted features are namely., Average stroke length, Shirorekha strength, Shirorekha confidence, Stroke density, Aspect ratio, Reverse distance, Average horizontal stroke direction and Average vertical stroke direction. Sarkar et al., in [1] have first extracted the text lines and words from document pages using a script independent neighboring component analysis technique. Then designed multi layer perceptron (MLP) based classifier for script separation, trained with 8 different word level holistic features and back propagation algorithm. Behrad et al., [9] used a method for script identification based on curvature scale space features. The proposed features are rotation and scale invariant and can be used to identify scripts with different fonts. Above method use cluster based weighted support vector machine for classification and recognition. The algorithm extracts the required features using principal component analysis (PCA) and linear discrimination analysis (LDA) algorithms. The extracted features are then classified using a new classification algorithm called Cluster Based Weighted Support Vector Machine (CBWSVM). Rajput and Anitha [6] have proposed a novel method towards multi-script identification at block level. The recognition is based on features extracted using discrete cosine transforms (DCT) and wavelets of Daubechies (is a wavelet used to convolve image data) family i.e., features are extracted by transforming the image in time domain to the image in frequency domain. KNN classifier is adopted for recognition purpose and the classifier computes the Euclidean distances between the test feature vectors with that of the stored features and identifies the k-nearest neighbor. Finally, the classifier assigns the test image to a class that has the minimum distance with voting majority. The corresponding script is declared as recognized script. Ghosh et al.,[8] presented several methods for automatic script identification developed so far. They mainly belong to two broad categories- structure based and visual appearance based techniques and gives an overview of the different script identification methodologies. Patil and Ramakrishnan [3] have evaluated the effectiveness of Gabor and discrete cosine transform (DCT) features independently using nearest neighbor, linear discriminant and support vector machines (SVM) classifiers. The combination of Gabor filter bank with either SVM or NN classifier handles the important issue of script identification at word level quite well. For most cases, Neural network classifier performs at par with SVM and they both outperform LDC. However, the actual performance is script dependent. The fact that LDC does not perform well as the other classifiers clearly indicates that the classes are not linearly separable. Now, SVM handles the non-linearity parametrically.While the Neural network classifier handles it non- parametrically. Moussa et al [10] have proposed multilingual automatic identification of handwritten and printed scripts. The proposed scheme is based on morphological transform of text line images and secondly on fractal analysis features of both (i) original textures of 2-D images, (ii) vertical and horizontal projection profile. The method is based on global texture analysis, by extracting fractal multi dimensions features. 12 features are obtained using the two techniques which are based on the classifiers; they are fractal multi dimension and K– Nearest Neighbor and radial basic function. Roy and pal [13] have proposed an automatic scheme for word-wise identification of hand-written scripts for Indian postal automaton. In the proposed scheme, at first, document skew is corrected. Next, using a piecewise projection method the document is segmented into lines and then lines into words. Finally, using different features like, water reservoir concept based features, fractal dimension based features, topological features, scripts characteristics based features etc., a neural network classifier is used for word-wise script identification. Biramani and Manjula [12] give a survey on identification of multi-scripts in document images. They have listed different classifiers and different existing methods for feature extraction. Finally they have concluded that these approaches can address complex tradeoffs between accuracy, time complexity and.the quality of the image. Table 1 presents the comparative analysis on various approaches used in handwritten script identification

3. CONCLUSION

This paper presents detailed survey on script identification of offline handwritten documents. Methods used for script identification, the datasets taken from different authors for script identification, the challenges involved in script identification, quantitative analysis, features and classifiers were used for comparative analysis. Further, our future work is aimed at quantitative and qualitative analysis of both offline and online handwritten script identification

4. ACKNOWLEDGEMENTS

The authors would like to thank Mr. S Manjunath, Assistant Professor, JSS College of Arts, Commerce and Science College, Ooty Road, Mysore, for his valuable comments and technical support rendered during the preparation of this paper.

Table 1. Script Identification Methods

Researchers	Features	Classifier	Dataset	Result
Sarkar et al [1]	Horizontalness feature, segmentation-based feature and foreground-background transition feature	Multilayer perceptron. Trained with back propagation algorithm	Bangla and Roman (1600 words) Devanagari and roman(1600 words)	Accuracies are 99.29 % and 98.43 % respectively.
Namboodiri and Jain [2]	Spatial and temporal features of strokes	Support vector machine (SVM)	13379 words (1,423Arabic, 1002 Cyrillic , 3173 Devanagari, 1981 Han , 2261 Hebrew, and 3575 Roman scripts)	Accuracy of 95.1% is obtained
Patil and Ramakrishnan [3]	Gabor and Discrete cosine transform (DCT)	Nearest neighbor and SVM	20000 words (Kannada , Devanagari and Roman scripts) printed	Accuracy of 89% for tri script document
Hochberg et al [4]	Relative Y centroid, Relative X centroid , number of white holes, sphericity, and aspect ratio	Linear Discriminant Analysis	498 documents (Arabic, Chinese, Cyrillic, Devanagari, Japanese and Roman scripts)	Accuracy of 88 % is achieved
Hangare and Dhandra [5]	Vertical stroke density, Horizontal Stroke Density, Right diagonal stroke density, Left diagonal stroke density	KNN Classifier K= 3	300 tri lingual documents (Devanagari, English and Urdu)	Accuracy of 97.83% (English), 93.00% (Devanagari) and 95.78 % (Urdu)
Rajput and Anitha [6]	Discrete cosine Transform and Discrete wavelet transform	KNN Classifier K= 1	800 multi script documents (English, Hindi, Kannada, Telgu, Malayalam , Tamil, Gujarathi and Punjabi)	Accuracy of 98 % (KEH), 99.2%(MEH), 93%(PEH), 99.2%(TEH), 90% (GEH) and 99% (TeEH)
Moussa et al [10]	Fractal analysis features	KNN Classifier and Radial Basic Function (RBF)	Arabic and Latin	96.64 % (KNN) and 98.72 % (RBF)

5. REFERENCES

[1] Ram sarkar,Nibaran Das, Subhadip Basu, MahantapasKundu., 2010.Word level script identification from bangla and devanagari handwritten texts mixed with Roman script. Journal of computing, Vol 2 , pp 103 – 108.

[2] Anoop M Namboodiri and Anil K Jain., 2002, On- line script recognition. 16th International Conference on Pattern Recognition, pp 736-739

[3] Peeta basu patil and A G Ramakrishnan., 2008. Word level multi-script identification. Pattern recognition letters, Vol 29, pp 1218 – 1229.

[4] Judith Hochberg , Kevin Bowers, Michael Cannon and Patrick Kally., 1999. Script and language identification for handwritten document images. International journal of Document analysis and recognition, vol 2, pp 45 - 52.

[5] Mallikarjun hangare and B V Dhandra., 2010. Offline handwritten script identification in document images. Intenational journal of computer applications.Vol 4, pp 6 – 10.

[6] G G rajput and Anith H B., 2010. Handwritten script recognition using DCT and wavelet features at block level. Recent trends in image processing and pattern recognition, pp 158 – 163.

[7] Joachim Schenk, Johannes Lenz, Gerhard Rigoll., 2009. Novel script line identification method for script normalization and feature extraction in on-line handwritten whiteboard note recognition.Pattern Recognition, Vol 42 , pp 3383 – 3393.

[8] D Ghosh, T Dube and A P Shivaprasad., 2009. Script Recognition – A review. IEEE Transactions on pattern analysis and machine intelligence, Vol 32, pp 2142 - 2161

[9] Alireza Behrad, Khoddami and Mehdi Salehpour., 2010.A novel framework for farsi and latin script identification and farsi handwritten digit recognition . Journal of automatic control, Vol 20, pp 17 – 25.

[10] S Ben Moussa, A Zahour, A Benabdelhafid, A M Alimi. 2008. Fractal – Based system for Arabic/Latin, printed/Handwritten script identification.19th International conference on pattern recognition, pp 1 – 4

[11] Guo Xian tan, Christian viard – Gaudin, Alex C kot., 2009. Information retrieval model for online handwritten script identification. 10th Intenational conference on document analysis and recognition, pp 336-340.

[12] S.Abirami and D. Manjula ., 2009. A Survey of Script Identification techniques for Multi-Script Document Images. International Journal of recent trends in engineering, Vol 1, pp 246 – 249.

[13] K Roy and U Pal., 2006. Word-wise Hand-written Script Separation for Indian Postal automation. Proceedings of International workshop frontiers in Handwriting recognition, pp 521 – 526

.