A Survey of Neural Network based Script Recognition using Wavelet Features

Sadanand R. Leshappanavar[1], Anoop Sharma[2]
[1]*Lecturer, Department of Computer Science and Engineering, Singhania University, Jhunjunu, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Singhania University, Jhunjunu, India*

***Abstract*: Script Recognition is a key step that arises in document image analysis especially when the environment is multi script and required to identify the different scripts that exists in the same script. South Indian Language such as Kannada, Telugu, Tamil and Malayalam requires identification of scripts in many applications. This project is to identify script of the given word among 100 script samples using suitable features, which can differentiate script of South Indian Languages namely Kannada, Telugu, Tamil and Malayalam by using more powerful feature vectors obtained by applying Discrete Wavelet Transform (DWT) to each of the script images. Learning Vector Quantization Neural Network (LVQ-NN) trained with the feature vectors is used as a recognizer. The combination of wavelet features and LVQ-NN gave expected results. 92% to 95% accuracy is obtained for 70 to 100 script images.**

***Keywords*: Script recognition, South Indian languages, DWT, LVQ-NN.**

1. Introduction

As the world moves closer to the concept of the “paperless office,” more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches, and modifications, as well as to prolong the life of such records [1]. A great portion of business documents and communication, however, still takes place in physical form and the fax machine remains a vital tool of communication worldwide. Because of this, there is a great demand for software, which automatically extracts, analyzes, and stores information from physical documents for later retrieval. All of these tasks fall under the general heading of document analysis, which has been a fast growing area of research in recent years [10].

A very important area in the field of document analysis is that of optical character recognition (OCR), which is broadly defined as the process of recognizing either printed or handwritten text from document images and converting it into electronic form[3]. Today, many algorithms have been presented in the literature to perform this task for a specific language, and these OCRs will not work for a document containing more than one language/script [7]. Therefore, a multilingual document page may contain text words in more than one regional language. So, multilingual OCR is needed to read these documents.

2. Script recognition

Script Recognition is a key step that arises in document image analysis especially when the environment is multi script and required to identify the different scripts that exists in the same script. Script Identification [10] facilitates many important applications such as sorting the images, selecting appropriate script specific text understanding system and searching online archives of document image containing a particular script.

Script Identification approaches can be broadly classified into two categories namely, local and global approaches [10]. 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.

3. Problem statement

South Indian Language such as Kannada, Telugu, Tamil and Malayalam requires identification of scripts in many applications. Problem of script identification of South Indian Language is a major challenge as the scripts of these languages resembles each other in one or other way. So this project is to identify script of the given word among 100 script samples using suitable features, which can differentiate script of South Indian Languages namely Kannada, Telugu, Tamil and Malayalam.

4. Review literature

This project on South Indian script recognition is basically in the field of research. To gain better knowledge, techniques and solutions regarding the procedures that we want to follow, we studied the various research papers on existing systems. These papers are much more relevant with our project. All these study helped us with clarifying our target goals. Some of the recognition techniques and knowledge that we became familiar with are discussed briefly in this Literature Review section.

1. A Generalized Frame Work for Script Identification

Gopal Datt Joshi, Saurabh Garg, Jayaswami Sivaswamy.

This paper models the script identification as a texture classification problem and examines a global approach inspired by human perception. Paper proposes a general framework to address the problem of script identification using two key features of frame work they are

a. A hierarchical strategy to help group given script into homogeneous script classes before attempting to classify them into single script classes. That is document image is processed to make it suitable for analysis.

b. The use of global analysis. That is by using Log Gabor filtering divide the document into fixed size text block.

2.Neural Network based System for Script Identification in Indian Document.

S Basavaraj Patil, N V Subbareddy Sadhana

This paper describes a neural network based script identification system. The system includes a feature extractor and a modular neural network. Feature extractor contain two stage where in first stage document image is dilated using 3x3 masks in horizontal, vertical, right diagonals and left diagonal directions. In the second stage average pixel distribution is found in these resulting images. The modular network is a combination of separately trained feed forward neural network classifiers for each script. It recognizes 64x64 pixel document images.

3. Word Level Script Identification for Scanned Document Images.

Huanfeng Ma, David Doermann

This paper compares the performance of three classifiers used to identify the script of words in scanned document images. In both training and testing, a Gabor filter is applied and 1 channels feature is extracted. Three classifiers Support Vector Machine (SVM), Gaussian Mixture Model and k-Nearest Neighbor (k-NN) are used to identify different script at the word level. These three classifiers are applied to a variety of bilingual dictionaries and their performance is compared.

4. Language Identification of Kannada, Hindi and English words through Visual Discriminating Features

M C Padma, P A Vijaya

The objective of this paper is to propose visual clues based procedures to identify Kannada, Hindi and English text portions of the Indian multilingual document. Visual discriminating features of four languages serves useful visual clues for language identification. The system identifies the features based on presence or absence of Horizontal lines, Vertical line, Variable sized blocks and blocks with more than one component.

5. Global and Local Features Based Handwritten Text Words and Numerical Script Identification

B V Dhandra, Mallikarjuna Hangarge

The script identification scheme proposed in his paper has to phases. First phase reports the script identification of text words using global and local features, extracted by morphological filters and regional descriptors of three major Indian languages/scripts (Kannada, Roman and Devnagari). For classification of text words and numerals, a K nearest neighbor algorithm is used.

6. Word-wise Sinhala Tamil and English Script Identification Using Gaussian Kernel SVM

Sukalpa Chanda, Srikant Pal, Umapada Pal

In this paper, a Support Vector Machine based technique is proposed for word wise identification of languages scripts from single document using three features. They are Structural feature, Topological feature and Water reservoir principle. Two stage classification schemes are used. In first stage SVM classifier is used on feature set to identify Sinhala script from other two scripts. In the second stage SVM is performed on feature set for classification for identification between English and Tamil scripts with highest accuracy.

7. Recognition of an Indian Script using Multilayer Perceptrons and Fuzzy Features

Shamik Sural, P. K. Das

This paper presents a multi-stage character recognition system for an Indian script, namely, Bengali using fuzzy features and multilayer perceptrons (MLP). The fuzzy features are extracted from Hough transform of character pattern pixels. It first defines a number of fuzzy sets on the Hough transform accumulator cells. The fuzzy sets are then combined by t-norms to generate feature vectors from each character. A set of fuzzy linguistic vectors is next generated from these feature vectors. The MLPs used for classification have the fuzzy features as inputs. The MLP outputs also represent the belongingness of an input pattern to different fuzzy character pattern classes.

8. A Novel Approach to Script Separation

Ranjith Kumar, Vamsi Chaitanya, C. V. Jawahar

This paper describes a new approach for script separation. A character level script separation scheme is combined with a Viterbi algorithm to get an optimal sequence of scripts which could generate such a text. Viterbi Algorithm provides a dynamic programming based solution to the identification of optimal state sequences from the given set of observations. A novel method to script separation without the help of structural and textural features is proposed in this paper.

9.Two-stage Approach for Word-wise Script Identification

Sukalpa Chanda, Srikanta Pal, Katrin Franke, Umapada Pal

A two-stage approach for word-wise identification of English (Roman), Devnagari and Bengali (Bangla) scripts is proposed. The 1st stage allows identifying scripts with high speed, yet less accuracy when dealing with noisy data. The advanced 2nd stage processes only those samples that yield low recognition confidence in the first stage. For both stages rough character segmentation is performed and features are computed on segmented character components. Features used in the 1st stage are a 64-dimensional chain-code-histogram feature, while

400-dimensional gradient features are used in the 2nd stage. Final classification of a word to a particular script is done via majority voting of each recognized character component of the word.

10. Script Identification from Indian Documents

Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy

In this paper, we present a scheme to identify deferent Indian scripts from a document image. This scheme employs hierarchical classification which 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.

11. Word-wise Hand-written Script Separation for Indian Postal automation

K. Roy, U. Pal

In this paper, an automatic scheme for word-wise identification of hand-written Roman and Oriya scripts is proposed for Indian postal automation. In the proposed scheme, using a piece- wise 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 (NN) classifier is used for word-wise script identification.

12. A Survey of Script Identification techniques for Multi-Script Document Images.

S. Abirami, D. Manjula

This paper attempted to provide some background information about the past researches on both global based approach as well as local based approach for script identification in document images. Both the systems can perform Script/Language identification in document images at document, line and word level. To be specific, Global based approaches gives an excellent result only if good quality document image prevails whereas local approaches can be adopted in low quality document images also. These approaches can address complex tradeoffs between algorithm speed, image quality and identification accuracy.

13. Information Retrieval Model for Online Handwritten Script Identification

Guo Xian Tan, Christian Viard-Gaudin, Alex C. Kot

This paper proposes a novel approach for online handwritten script identification based on the Information Retrieval (IR) model. It attempts to identify among three script families; Arabic, Roman and Tamil scripts. This paper discussed an approach that involved the use of IR techniques to build prototypes to model the script families as stochastic distributions of frequency vectors, thus providing a simple and effective method of indexing and retrieving the identity of the script. The advantage of the IR method is that it allows the frequency of occurrence of features pertaining to that script to be taken into account through the usage of the tf-idf (term frequency &inverse document frequency) combination, hence similar scripts such as Hangul and Chinese which share common features can be better differentiated using their relative contribution to the frequency vectors.

14. Script Identification of Camera-based Images

Linlin Li, Chew Lim Tan

This paper reports a statistical script identification technique that determines the script of document images, especially camera-based images which suffer from perspective distortion. The identification technique represents a document image by a frequency vector of affine invariant signatures of characters, and identifies the script by comparing the vector with pre prepared script templates.

15. Morphological Reconstruction for Word Level Script Identification

B. V. Dhandra, Mallikarjun Hangarge

This paper identified a tool of morphological opening by reconstruction of an image in different directions and regional descriptors for script identification at word level, based on the observation that every text has a distinct visual appearance. The proposed system is developed for three Indian major bilingual documents, Kannada, Telugu and Devnagari containing English numerals. The nearest neighbor and k-nearest neighbor algorithms are applied to classify new word images.

16. Bangla/English Script Identification Based on Analysis of Connected Component Profiles

Lijun Zhou, Yue Lu, Chew Lim Tan

This paper presents a novel and efficient technique for Bangla/English script identification with applications to the destination address block of Bangladesh envelope images. The proposed approach is based upon the analysis of connected component profiles extracted from the destination address block images, it does not place any emphasis on the information provided by individual characters themselves and does not require any character/line segmentation.

17. AI Approach to Hand Written Devnagiri Script Recognition

Dileep Kumar

This paper proposes a system that uses an AI approach to integrate information from diverse sources. It has three levels of abstraction low, medium and high. At each level of abstraction, knowledge appropriate to that level is used to identify the components of the higher level concept. At low level of abstraction, a given document is segmented into components, at the medium level features are extracted from the segment, and at the high level segments are recognized with the help of available contextual information. To deal with imprecise information a new approach based on the fuzzy logic concept has been suggested.

18. Lexicon-based Word Recognition Using Support Vector Machine and Hidden Markov Model

Ahmad A R, Viard Gaudin C, Khalid M

This paper focus on online word recognition using the support vector machine (SVM) for character recognition. SVM's use of structural risk minimization (SRM) principle has allowed simultaneous optimization of representational and discriminative capability of the character recognizer. It uses SVM in isolated character recognition environment using IRONOFF and UNIPEN character database then demonstrate the practical issues in using SVM within a hybrid setting with HMM for word recognition by testing the hybrid system on the IRONOFF word database and obtained commendable results.

19. Texture for Script Identification

Andrew Busch, Wageeh W Boles, Sridha Sridharan

This paper investigates the use of texture as a tool for determining the script of a document image, based on the observation that text has a distinct visual texture. An experimental evaluation of a number of commonly used texture features is conducted on a newly created script database, providing a qualitative measure of which features are most appropriate for this task. Strategies for improving classification results in situations with limited training data and multiple font types are also proposed

20. Techniques for Language Identification for Hybrid Arabic-English Document Images

Ahmed M. Elgammal, Mohamed A. Ismail

This paper proposes three efficient techniques that can be used to discriminate between text written in Arabic script and text written in English script are presented and evaluated. These techniques address the language identification problem on the word level and on text line level. The characteristics of horizontal projection profiles as well as run length histograms for text written in both languages are the basic features underlying these techniques.

5. Conclusion

Script Recognition is a process that has no fixed solution [18]. The challenges, researchers face lies in finding the attributes that can clearly distinguish the scripts that are trying to recognize. However, no mathematical approaches or scientific techniques can be used to find these attributes. Research is mostly done on a trial and error basis. In addition, since different scripts have different attributes and style, the document attribute that is successful in identifying one script may not be successful in another script. In addition, the attributes must also be independent of the quality and variations of the document to handle multiple fonts, sizes and poor quality. This paper attempted to provide some background information about the past researches on both Hidden Markov Model approach as well as Wavelet based approach for script identification in document images. These approaches can address complex tradeoffs of quality and identification accuracy.

References

[1] Gopal Datt Joshi, Saurabh Garg, Jayaswami Sivaswamy , “A Generalized Frame Work for Script Identification”. International Journal of Document Analysis and Recognition (IJDAR).

[2] S Basavaraj Patil, N V Subbareddy, “Neural Network based System for Script Identification in Indian Document”. Sadhana Vol. 27, Part1, February 2002, pp. 83-97.

[3] Huanfeng Ma, aDavid Doermann, “Word Level Script Identification for Scanned Document Images”. www.lampsrv02.umiacs.umd.edu.

[4] M C Padma, P A Vijaya, “Language Identification of Kannada, Hindi and English words through Visual Discriminating Features”. International Journal of Computational Intelligence System, Vol.1, No.2 (May, 2008), 116-126.

[5] B V Dhandra, Mallikarjuna Hangarge, “Global and Local Features Based Handwritten Text Words and Numerical Script Identification”. International Conference on Computational Intelligence and Multimedia Applications 2007.

[6] Sukalpa Chanda, Srikant Pal, Umapada Pal, “Word-wise Sinhala Tamil and English Script Identification Using Gaussian Kernel SVM”. IEEE society 978-1-1-4244-2175-6/08

[7] Shamik Sural, P.K.Das, “Recognition of an Indian Script using Multilayer Perceptrons and Fuzzy Features”. Sixth International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, 2001, pp. 1120-1124.

[8] Ranjith Kumar, Vamsi Chaitanya, C. V. Jawahar, “A Novel Approach to Script Separation”. www.cvit.iiit.ac.in-papers-ranjith03novel

[9] Sukalpa Chanda, Srikanta Pal, Katrin Franke, Umapada Pal, “Two-stage Approach for Word-wise Script Identification”. 2009 10th International Conference on Document Analysis and Recognition.

[10] 1Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy, “Script Identification from Indian Documents”. Springer-Verlag Berlin Heidelberg 2006, LNCS 3872, pp. 255–267.

[11] K. Roy, U. Pal, “Word-wise Hand-written Script Separation for Indian Postal automation”. Published Website: www.hal.archives-ouvertes.fr

[12] S.Abirami, Dr. D.Manjula, “A Survey of Script Identification techniques for Multi-Script Document Images”. International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[13] Guo Xian Tan, Christian Viard-Gaudin, Alex C. Kot, “Information Retrieval Model for Online Handwritten Script Identification”. 2009 10th International Conference on Document Analysis and Recognition.

[14] Linlin Li, Chew Lim Tan, “Script Identification of Camera-based Images”. www.comp.nus.edu.

[15] B.V.Dhandra, Mallikarjun Hangarge, “Morphological Reconstruction for Word Level Script Identification”. www.cscjournals.org.

[16] Lijun Zhou, Yue Lu, Chew Lim Tan, “Bangla/English Script Identification Based on Analysis of Connected Component Profiles”. www.comp.nus.edu.

[17] Dileep Kumar, “AI Approach to Hand Written Devnagiri Script Recognition,” TENCON '91.1991 IEEE Region 10 International.

[18] Ahmad A R, Viard Gaudin C, Khalid M, “Lexicon-based Word Recognition Using Support Vector Machine and Hidden Markov Model”. 2009 10th International Conference on Document Analysis and Recognition.

[19] Andrew Busch, Wageeh W Boles, Sridha Sridharan, “Texture for Script Identification”. IEEE Transactions On Pattern Analysis and Machine Intelligence, vol. 27, no. 11, November 2005.

[20] Ahmed M. Elgammal, Mohamed A. Ismail, “Techniques for Language Identification for Hybrid Arabic-English Document Images,” 2009 10th International Conference on Document Analysis and Recognition.