International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Classification of Handwritten Multilingual Documents

Raushan Kumar Singh[1], Akhilesh Pandey[2]

[1]M. Tech Scholar, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
[2]Assistant Professor, Suresh Gyan Viahr University, Jaipur, Rajasthan, India

Abstract: ***In the Branch of computer Science handwritten script recognition is demanding part. Projected work is highlighting on the "block level technique. In this wok we use the feature extraction technique for the recognition and we combined approach of Discrete Cosine Transform (DCT) and discrete wavelets Transform (DWT) for feature extraction and neural network (feed forward back propagation) classifier for classification and recognition purpose. The projected work has been experiment on three handwritten scripts Hindi, English and Urdu. For this work we create a database. That contains 9862 handwritten samples, written in three scripts.***

Keywords: Multi-script documents, handwritten script, Discrete Cosine Transform, Wavelets

1. Introduction

In the today scenario number of researchers done a very challenging phase for the recognizing the multi script data set and they have been proposed these multi scripts (Hindi, English, Urdu) can be easily identified. However these all algorithm not be give the satisfaction results .Handwritten multi script data for the reorganization can be divided into two classed i.e. online Handwritten script and offline Handwritten script .Online handwritten script artificial arrangement with program conversation with multi script. Which are written on special digitalization and tablets PCs, where sensor pic up pint point and woman of pint switched between script to script. Multi Script depends upon scanned handwritten manuscript which deals with data set. The chief detached of handwritten Multi Script acknowledgment (HMSR) is to identify the Multi Scripts in required presentation from image format so that these could be simply corrected. Multi-line handwritten script confirm that many in the field of design gratitude research surveys and offline cursive script word acknowledgment presented obtainable presented inch target is similar to the work of CLA and advancement innovation advancement information. OCR heir is described in India which finished script survey analysis survey comprehend benchmark database. Many technology must go hand-written multiple scripts recognition acknowledgment recognition but still so little satisfied divided into three parts, the first one introduced in India about the automatic recognition of handwritten and regional official script. This script contains nine regions and then divided into four subgroups based on t-efficiency and identification accuracy. Like the concepts of artificial intelligence neural networks are secondhand to achieve the effort, the attention can fix. Explore how the idea of the general human recognize diagnose recognize text and the process used to develop simulation machine. Distinguish multiple scripts to develop these intelligent machines are not a relaxed commission; this is since more than one script could be printed in diverse conducts. There are very limitations in adequacies perfections and handwritten deviation as position place position, noise and angle, so that multi-script handwriting recognition challenges to the implementation of home appliances. . Identify existing script relies on different features missing like DCT and DWT introduction .the OCR technique is applied on the Devnagari script on paper. In paper metadata describing the text in paragraph, page and line level. Extraction of paragraph from & segmentation of paragraphs into lines are also been established& implemented. Different methods for Amharic term gratitude in unconstrained handwritten manuscript using HMMs describe in In this first approach is to build a connection from the root word character model portfolio structure and composition of the second approach appeals HMM models together to form the word model. In the Persian name for a subset of the paper offline Arabic Persian handwriting recognition algorithms available. Here are using RBF neural genetic and K-means clustering algorithm and permutation networks. This article is about the Indian language street name recognition works. We know that some street names that contain two or more words, then it is concatenated to create a word. Many study has been completed to in the related parks, such as cognitive science artificial processing, image recognition, pattern intelligence, further research is under way to improve the accuracy and efficiency to solve the multi-script handwriting recognition problems. Offline multi-script handwriting recognition is the area of many researchers working the field of decoration gratitude. Some method has been functional to the multi-script handwriting grateful, nevertheless immobile it is known under circumstances fewer competence and accurateness.

Existing script identification depends on the different feature extraction like DCT and DWT presented in [3].the OCR technique is applied on the devanagri script on [4] paper. In [5] paper metadata describing the text in paragraph, page and line level. Tools to extract paragraphs from pages, segment paragraphs into lines have also been developed. two approaches for Amharic word recognition in unconstrained handwritten text using HMMs describe in [6].in which first approach builds word models from combined features of constituent characters and in the second method HMMs of constituent characters are concatenated to form word model. In [7] paper offline arbiclFarsi handwritten recognition algorithm on a subset of Farsi name is proposed. There have use RBF neural network and combination of GA and K-Means clustering algorithm. The [8] paper is works on street name

recognition on Indian language. we know that some street name contain two or more than words so it is concatenate that's word and create in a single word. Hence, in this paper, we present a multiple feature based approach that combines Discrete Cosine Transform (DCT) and Wavelet based frequency contents for three Indian scripts including English, Hindi and Urdu. The classification is done using feed forward back propagation neural network classifier. The experiments are carried out on the database at block level.

2. Background Information

On behalf of multi-script handwriting recognition, HMSR machine depends on the learning process, the feed forward back-propagation algorithm requires input from the user. In this learning process, training and testing multi-script has been completed. A library of information that is stored in the text segment which for future comparison. This library helps script and refused to accept more. For example: the script was written more than 50 times, 40 of which font are used for neural network training and the rest of the library used to test the network. The importance of growth affectedly neural network in the past 15 years. A large number of universities and companies are using neural networks and works based on neural networks available on the market.

Neural networks can be used as the human brain; the neural network structure is the same machine structure of the human brain. There in custom IC hundreds or even thousands of neurons. In the aggregate, an interest in learning, nonlinear dynamics and parallel computing increasingly stimulated renewed attention in artificial neural networks. There are various actual applications, such as configuration identification, system identification and noise.

The neural network is widely used for the removal and the like. When the positive result came out the numerous appreciation used. The neural networks accepted by many non-Indian and Indian scripts accurately and effectively. There are too many applications, it can easily take advantage of neural networks is tough outdated methods to solve solved. Neural network is composed of three layers, hidden layer, output layer and input layer. Each layer consists of small interconnected handling elements. These elements are organized with every other by weighted links. Each unit has a separate function, but the combination of these units display complex behavior. Neural network is a massively parallel distributed processor, has a natural tendency to store research knowledge to make it available for use. Neural networks like the human brain. Neural network to acquire knowledge, knowledge of the human brain obtained from the learning process. Neural networks have many advantages over conventional systems. For example: - This is a stupid noise and easier to handle, because it contains fewer people working than other traditional statistical analysis. Neural network algorithm can be solved without a solution or a problem, and its solution is too complex algorithms and found the problem. It has fewer errors because it can respond to anything, and small changes do not usually enter the result in a change in the output. This behavior of neural networks shows its importance.

On the origin of that data acquirement process, Script recognition can be categorized into following two parts:

1. Online Script Recognition
2. Offline Script Recognition

Offline Handwriting Recognition is the recognition that has been perused from a superficial and digital storage format in word gray scale process. After wards life deposited, it is predictable to complete additional dispensation to permit excellent recognition. If the on-line handwritten script recognition, handwriting is seized and deposited in numerical form through different means. Frequently, a singular pen is used in combination with an electron donating superficial. As the pen transfers on the surface, a continuous two-dimensional organizes of the point of time is expressed as the tools and is maintained in order. It is widely believed online handwritten text recognition methods have reached better results than their corresponding offline. This may be due to the additional information in the case of on-line, such as the order of the direction, speed and handwritten strokes are captured fact.

The main difference between online and offline recognition script is that the script online with real-time identification of appropriate information, but there is no offline data.

A. Handwritten Multi Script Recognition

Multifunctional handwriting recognition script (HMSR) is an area of pattern recognition has been the subject of considerable research, because the last few decades. There is too much utilization (i.e. India offices, such as banks, sales tax, railways, embassies, etc.) in English, Hindi and Urdu languages. Many forms and procedures are filled in these languages; sometimes these forms must be directly scanned. If you do not HMSR structure, and then the image is captured directly and have the option to edit those items. Script handwriting recognition (HSR) is a fully automated computer processes the text thanksgiving optical scanning and digitization of web scripts. The main purpose of a system is to distinguish HMSR multi-script, which is in the form of digital images without any manual intervention. This is extracted by matching the search from the given image and the image of the script function between the model libraries is complete. The library helps Functional differences between the images of the script; this confusion contempt correct script identification. First HMSR image search system using matching data from the user input, and causes the preprocessing stage, the feature extracted from the extraction and Image model library, and then the script classification.

B. Pre-processing

In HMSR, typical preprocessing operations include:

1. Binarization
2. Noise reduction
3. Skew detection

The main objectives of Pre-processing methods are:

- In preprocessing technique we perform 2 operation
- Binarization
- Thinning

After pre-processing phase, a cleaned image is available that goes to the segmentation phase. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of Script analysis.

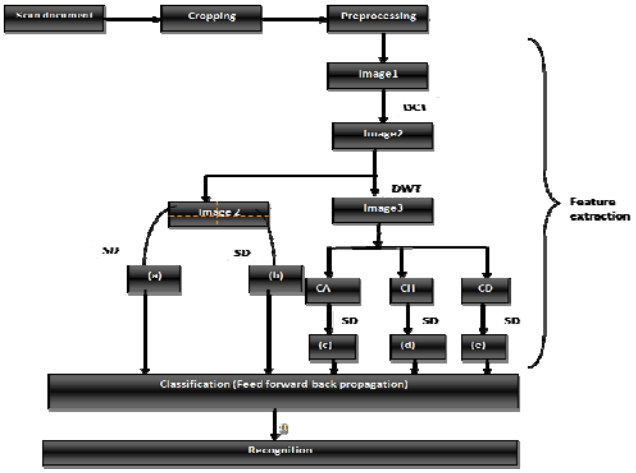


Figure 1: Block Diagram of Script Identification

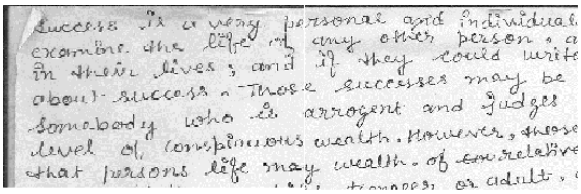


Figure 2: Script Sample of English Language

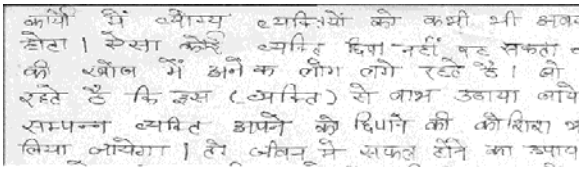


Figure 3: Script Sample of Hindi Language

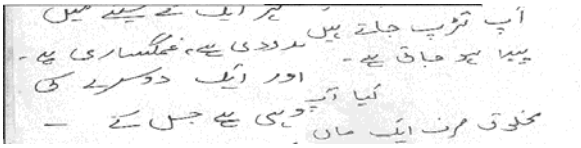


Figure 4: Script Sample of Urdu Language

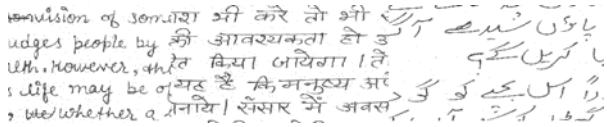


Figure 5: Combined Sample of multi script

C. Segmentation

It is intended for the decomposition of an image is divided into a configuration script for each symbol sub-image operations. Segmentations are key condition script controls the effectiveness of the straight system. Some approaches can be used followed by a square-like segmentation strategy according to the text and to identify the type of cut based segmentation and classification methods for classification. To achieve wide use, it is a segmentation method has the following properties are important:

1. Capture group or sensibility significant area, which often reflects the image of a global problem. Two central issues are what are important to provide accurate Securitizations perception, and you can specify the same for a given partition. Performance should be precisely defined division caused better facilitate a better understanding of the method and different methods. The actual method to be used in edge detection quickness' anatomical or other similar low-level visual processing technology, which means close to the linear operation with a low constant factor. For example, the video processing operations may be used in several applications per frame division.

D. Feature Extraction

Each script has some features, which play a significant role in pattern recognition. English, Hindi and Urdu script has many special features. Description, such a classification model feature extraction task becomes very easy to contain information about a pattern in the shape of a proper driver. These handwritings splinter scrutiny system in HMSR feature extraction stage, and selected a customary of landscapes that can be castoff to categorize abnormal script section. Mainly, this is the heart HMSR stage system, since the results be contingent on these topographies. Article abstraction are assumed to family, is included in the program for measuring information related to the shape of a pattern, the sort pattern so that the task is facilitated through the formal name of the program. Tangled in building recognition system in which different design issues, is perhaps one of the most extensive set of features to choose from. Feature extraction for exploratory data projector so that the concept of high-dimensional data to better understand and clustering data structure. The computational requirements are reduce to quotation of the characteristics of great discriminate dimensionality reduction, in the feature extraction of the image. However, feature extraction rule, projection timings exploratory objective is to minimize the error function of data, such as mean square error or difference from the inter mode, the purpose of feature extraction and classification is divided into classes the better enhancement. Therefore, the best feature extraction (for specific Standard) for exploratory data prediction is not automatically the best in the class can be divided into functions, and vice versa. In particular, two or more classes can have primary function is similar. In addition, article abstraction for examining data prognosis for two or believable data visualization, and classification typically require more than two or three characteristics. Therefore, not mostly for classification, and vice versa feature extraction paradigm exploratory data projection.

3. Representation of Script Features

Currently, in India, India handwritten script standard database is unavailable. Therefore, the training and test data classification scheme is to collect from different sources. Are in English, Hindi and Urdu script handwritten documents belonging to different people in different industries to collect. The document scanned at 300 dpi and gray scale image storage. Size 512 × 512 pixel image block, and then manually extract files from different regions of the image. It should be noted that the handwritten text block may contain two or more lines, rows, different font sizes (large and small) and variable bit between words and characters. We do not perform any processing, homogenization parameters. It ensures that the area of at least 50% block of text containing the text. These blocks are equivalent to a part of the handwritten document, and then the two values, so that the text and the background on behalf of a representative value of 0. In the proposed system, using morphological opening around the boundary noise is removed. This operation also removes non-contiguous pixel level.

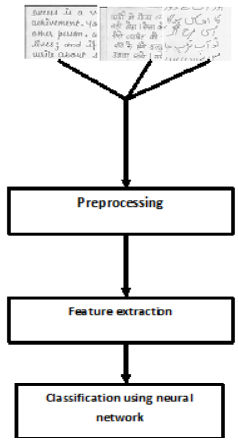
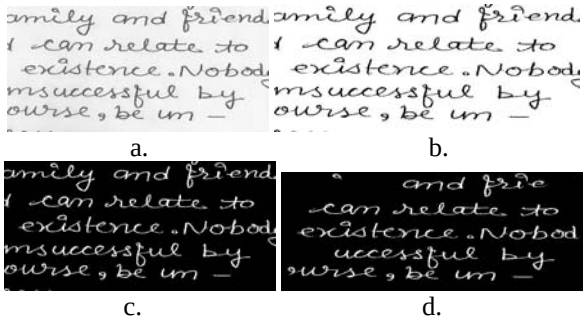
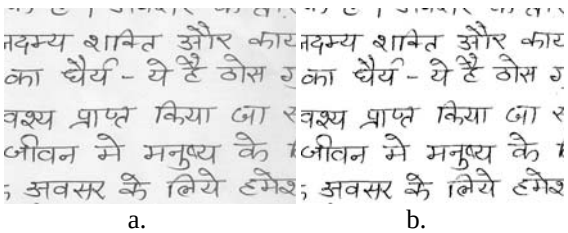


Figure 6: Block diagram of Methodology

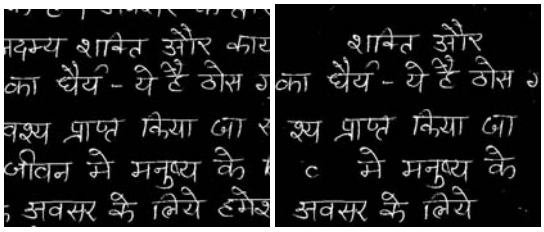


a. b. c. d.

Figure 7a: Original Cropped Image of English Script b. Black & White Image c. Invert color d. Clear component clear border

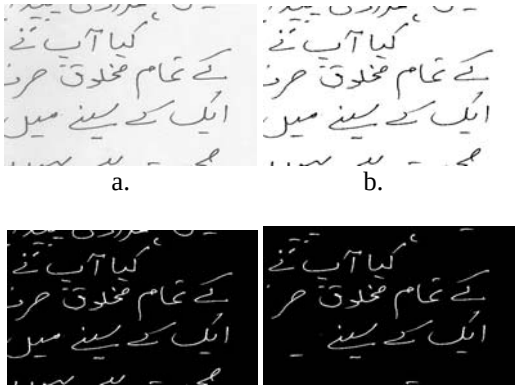


a. b.



c. d.

Figure 7a: Original Cropped Image of Hindi Script b. Black & White Image c. Invert color d. Clear component clear border



a. b. c. d.

Figure 8a: Original Cropped Image of Urdu Script b. Black & White Image c. Invert color d. Clear component clear border

4. Results

For the reorganization of handwritten script that is prepared by different people in different location .The total dataset is 9862 there the Hindi sample is 3373, English sample 3269 and Urdu is 3220. In the total dataset is divided in to two parts one part is training purpose, other part is testing purpose. For recognition of each script. Features are calculated and safe for the training purpose. The neural network is having three type of the layer, one is the input layer, second is hidden layer and third is the output layer. If the increases the number of neurons in the hidden layer the result will be increases and decreases on given script.

In the back propagation algorithm one layer is behave like an input layer. Second one is the hidden layer and last one is the output layer. If increase the number of neuron in hidden layer then required memory allocation problem can be happened and also the required result are not acquired if the value of tolerance is increased can take more number of cycle for learning purpose to obtaining the results . But learning is not up to mark up the result not to acquire desire.

Scripts	No. of samples	Train/test	Recognition result
Hindi	3373		
English	3269		
Urdu	3220	4932/4930	92.70%

Table 1: Result of Multiple Scripts

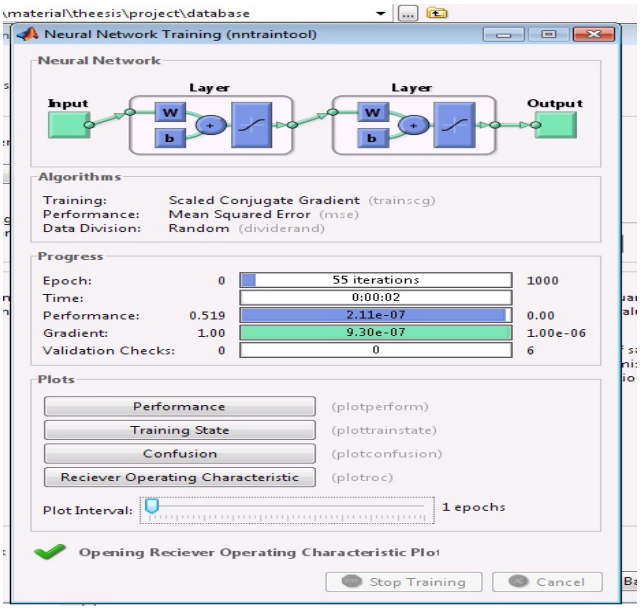


Figure 9: Diagram of NN Training

Reference

[1] Nabin Sharma. With Co-Authored with U. Pal, and R. Jayadevan, "Handwriting recognition in Indian regional scripts: A survey of offline techniques"

[2] Ram sarkar, nibaran das, subhadip basu, mahantapas kundu, mita nasipuri and dipak kumar basu," cmaterdb1: a database of unconstrained handwritten bangla and bangla-english mixed script document image", international journal on document analysis and recognition Volume 15, number 1 (2012), 71-83, doi: 10.1007/s10032-011-0148-6, 2012.

[3] G. G. Rajput and Anita H. B.," Handwritten Script Recognition using DCT and Wavelet Features at Block Level",2010.

[4] Jayadevan, R. Pune Inst. of Comput. Technol., Pune, India Kolhe, S.R. ; Patil, P.M. ; Pal, U.," Offline Recognition of Devanagari Script: A Survey", Volume: 41 , Issue: 6,Product Type: Journals & Magazines,2011.

[5] AlKhateeb, J.H.," A new approach for off-line handwritten Arabic word recognition using KNN classifier", 18-19 Nov. 2009.

[6] Assabie, Y.,' HMM-Based Handwritten Amharic Word Recognition with Feature Concatenation", Document Analysis and Recognition, ICDAR '09. 10th International Conference, 2009.

[7] Bahmani, Z., Alamdar, F., Azmi, R., Haratizadeh, S.," 8) Off-line Arabic/Farsi handwritten word recognition using RBF neural network and genetic algorithm", Intelligent Computing and Intelligent Systems (ICIS), IEEE International Conference on 2010.

[8] Pal, U., Roy, R.K., Kimura, F.," Handwritten street name recognition for Indian postal automation", Document Analysis and Recognition (ICDAR), International Conference on 2011.

[9] Liangrui Peng, Changsong Liu, Xiaoqing Ding, Hua Wang, "Multilingual document recognition research and its application in China," dial, pp.126-132, Second International Conference on Document Image Analysis for Libraries (DIAL'06), 2006.

[10] U. Pal and B. Chaudhuri. Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line. In International Conference on Document Analysis and Recognition, pages 790-794, 2001.

[11] U.bhattacharya,T.K Das,A.Datta,S.K.Parui,B.B Chaudhuri,"A hybrid scheme for hand printed numeral recognition based on a self- organizing network and MPL Classifiers,Int.J.Pattern Recognitoin Artificial Intelligence".16(2002) 845-864.

[12] K. H. Aparna, V. Subramaniam, M. Kasirajan, G. V. Prakash, V. S. Chakravarthy and S. Madhvanath, "Online handwrting recognition for Tamil", in the Proceedings of 9th International Workshop on Frontiers in Handwriting Recognition(IWFHR), pp. 438-443, 2004.

[13] C. V. Lakshmi and C. Patvardhan, "A high accuracy OCR system for printed Telugu text", in the Proceedings of Conference on Convergent Technologies for Asia-Pacific Region (TENCON 2003), Vol. 2, pp. 725-729, 2003.

[14] Lei Han, Jue Zhong, Arkady Voloshin, Image analysis and data processing of time series fringe pattern of PCBs by using moiré interferometry,in: Proceedings of HDP'04, 2004, pp. 141–145.

[15] Ping Zhong, Chenjie Song, Nian Luo, Method of extracting high-resolution digital moiré fringe in warpage measurement, Physical and Failure Analysis of Integrated Circuits, IPFA, 2009, pp. 527–530.

[16] V. Ablavsky and M.R. Stevens, "Automatic Feature Selection with Applications to Script Identification of Degraded Documents," Proc. Int'l Conf. Document Analysis & Recognition, Edinburgh, pp.750-754, Aug. 2003.

[17] D.Dhanya, A.G Ramakrishnan and Peeta Basa pati, "Script identification in printed bilingual documents," Sadhana, vol. 27, part-1, pp. 73-82, 2002.

Authors Profile

Raushan Kumar Singh is a M.Tech Scholar in computer science department with information communication at Suersh Gyan Vihar university. He does work on pattern recognition in multiple language script and work with the help of MAT LAB. .His interest area is pattern recognition and Neural Network

Akhilesh Pandey is an Asst. Professor in department of computer science and engineering Suresh Gyan Vihar University, Jaipur. He did his MCA from IGNOU in 2002 and after that he worked as a faculty member in different engineering college. After that he acquired his M. Tech. (CSE) at Sharda University, Gr. Noida, India, His area of Interest is Pattern Recognition and neural network.