ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

Survey on Existing Techniques for Writer Verification

Ms. Rashmi Welekar[1], Ms. V. Swetha D. Rao[2]
[1]Assistant professor, Dept. of CSE, RCEOM College, Nagpur, Maharashtra, India
[2]PG Student [CSE], Dept. of CSE, RCEOM College, Nagpur, Maharashtra, India

Abstract: This paper presents a survey of the literature on handwriting analysis and writer verification schemes and techniques up till date. The paper outlines an overview of the writer identification schemes mainly in English, Arabic, Bangla, Malayalam and Gujrati languages. Taxonomy of different features adopted for online and offline writer identification schemes is also drawn at. The feature extraction methods adopted for the schemes are discussed in length outlining the merits and demerits of the same. In automated writer verification, text independent and text dependent methods are available which is also discussed in this paper. An evaluation of writer verification schemes under multiple languages is also analyzed by comparing the recognition rate. New method proposed for identifying writer using slant, orientation, eccentricity enabling to identify writer's mental state by features associated.

Keywords: character recognition, character segmentation, writer verification, segmentation, feature extraction

I. INTRODUCTION

Character recognition has been one of the most active topics in pattern recognition for several decades. It is the process of converting an image representation of a document into digital form. The document image can be printed or handwritten. Handwritten data is converted to digital form either by scanning the writings on paper or by writing with a special pen on an electronic surface such as a digitizer combined with a liquid crystal display. The two approaches are termed as off-line and on-line handwriting, respectively.

A handwritten character recognition system consists of four phases: pre-processing, feature extraction, classification and post-processing. Among them, feature extraction is one of the most important factors in achieving high recognition performance. The remaining part of this paper shows briefing about the feature extraction concepts and in later sections it shows some classifiers used for English language.

II. CONCEPT OVERVIEW

Off-line Handwriting deals with the recognition of handwritten words after it was written. Moreover, there is little or no control in most offline scenarios of the type of medium and instrument used. The artifacts of the complex interactions between medium, instrument, and subsequent operations such as scanning and binarizations present additional challenges to algorithms for offline Handwriting. It is, therefore, generally regarded as much more difficult compared to its online counterpart.

There can be mainly two approaches for the word recognition purpose for any handwritten documents.

1) Analytical approach: It treats words as a collection of simpler subunits such as characters and proceeds by segmenting the word into those subunits, and then identifies the units. As an example to identify any word using this approach the letters are to be identified first then they are used for the total word recognition.

2) Holistic approach: It treats the word as a single, indivisible entity and attempts to recognize it using features of the word as a whole. So, in this approach the whole word image features are used for the recognition purposes. Researchers have utilized many different approaches for both the segmentation and recognition tasks of word recognition. Some researchers have used conventional, heuristic techniques for both character segmentation and recognition some have used a convex hull based and recursive contour following algorithms, while others have used heuristic techniques for segmentation followed by *ANN* based methods for the character/word recognition process. Hidden Markov Model based techniques are also used widely for both offline and online hand-written document recognition.

III. CHARACTER RECOGNITION SYSTEM

The character recognition system can be divided as segmentation of text document into character and recognition of character. The whole process is shown as

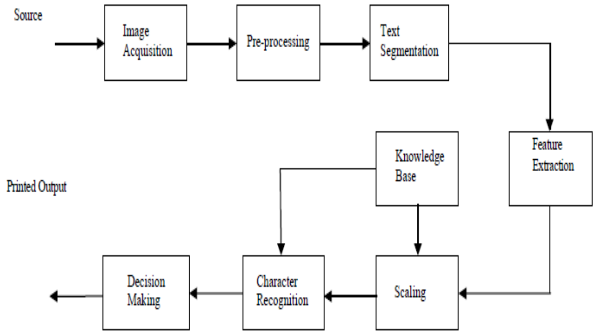

Figure1: Block diagram of character recognition system

a) Image Acquisition
The input images are acquired from documents containing text by using scanner as an input device or using Adobe Photoshop or Paint. Acquired images are then stored in Hard Disk in JPG picture format. This image is then passed for preprocessing.

b) Pre-Processing
The scanned image is converted into binary image. At first, the RGB image is converted into grayscale image and then binary image i.e. an image with pixel 0 (white) and 1 (black). After converting the image, the unnecessary pixels (0s) from the original image is removed.

c) RGB to Grayscale and Gray to RGB Conversion
In practical cases most of the images are generally color (RGB), but it is complex to work with a three-dimensional array. So it needs to convert the RGB image into the grayscale image. The RGB to grayscale conversion is performed by MATLAB command.

I = rgb2gray(f)

For ease of analysis, the grayscale image is converted into binary image by using the following MATLAB command.

BW = im2bw(I)

IV. SEGMENTATION

Text segmentation is a process where the text is partitioned into its elementary entities i.e. characters. The total performance of the character recognition process depends on the accuracy of the segmentation process of the text into the characters. In the segmentation phase, first the document is segmented into text lines, the text lines are segmented into text words and then the words are segmented into characters.

a) Line Segmentation
Text line segmentation is performed by scanning the input image horizontally. Frequency of black pixels in each row is counted to separate the line. The position between two consecutive lines, where the number of black pixels in a row is zero denotes a boundary between the lines.

b) Word Segmentation
In English text there is a minimum gap between two consecutive characters and two consecutive words. The minimum gap between two consecutive words is greater than two consecutive characters.

c) Character Segmentation
For character segmentation from the word, the vertical scan is performed. The starting boundary of a character is the first column where the first black is found.

V. FEATURE EXTRACTION

The writer identification task lies in the definition of a feature space common to all the handwritten documents. In this study we have extended this principle to the whole document database. Following the segmentation of the handwritten document, a set of binary features is defined thanks to a clustering procedure. In this manner the feature set is adapted to the handwritings of the database under study. We briefly recall the main characteristics of the clustering procedure. Several sequential clustering phases are iterated with random selection of the elements. Each of them provides a variable number of clusters. The invariant clusters are defined as the groups of patterns that are always clustered together during each sequential clustering phase.

VI. METHODS USED

i) **Multilayer Perception**
Srihari proposed a large number of features for the writing which can be classified into two categories. a) Macrofeatures – They operate at document/paragraph/word level. The parameters used are gray-level entropy and threshold, number of ink pixels, number of interior/exterior contours, number of four-direction slope components, average height/slant, paragraph aspect ratio and indentation, word length, and upper/lower zone ratio. b) Microfeatures – They operate at word/character level. The parameters comprise of gradient, structural, and concavity (GSC) attributes. Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. We represent the error in output node j in the n th data point by $e_j (n)=d_j(n)-y_j(n)$, where d is the target value and y is the value produced by the perceptron.

ii) Bayesian Classifiers
Zois and Anastassopoulos implemented writer identification in 2000 and verified using single words. After image thresholding and curve thinning, the horizontal projection profiles were resampled, divided into 10

segments, and processed using morphological operators at two scales to obtain 20-dimensional feature vectors. Classification was performed using either a Bayesian classifier or a multilayer perceptron.

iii) k-nearest neighbor classification

Writer identification scheme suggested by Marti and Hertel and Bunke, text lines was the basic input unit from which text-independent features are computed using the height of the three main writing zones, slant and character width, the distances between connected components, the blobs enclosed inside ink loops, the upper/lower contours, and the thinned trace processed using dilation operations.

KNN is an *non parametric lazy learning* algorithm. These samples are the *kn nearest-neighbors* of **x**. It the density is high near **x**, the cell will be relatively small, which leads to good resolution. If the density is low, it is true that the cell will grow large, but it will stop soon after it enters regions of higher density.

iv) Hidden Markov Model

Schlapbach et al. implemented an HMM based writer identification and verification method . An individual HMM was designed and trained for each writer‘s handwriting. To determine which writer has written an unknown text, the text is given to all the HMMs. The one with biggest result is assumed to be the writer. The identification method was tested by using documents gathered from 650 writers. The identification accuracy was 97%. Also, this method was tested as a writer verification method. This was achieved by a collections writings from 100 people and twenty unskilled and twenty skilled imposters, who forged the originals. Experimentations results obtained showed about 96% overall accuracy in verification. Thus it is obvious, that this method can be extended to other languages by applying some changes on feature extraction phase.

HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be considered as the simplest dynamic Bayesian network. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.

v) Manhattan Distance

Vladimir Pervouchine implemented a writer identification scheme based on high frequent characters. In this method, the high frequent characters (‗f‘,‘d‘,‘y‘,‘th‘) are first identified, and then according to the similarity of those characters, the writer is selected. The similarity is calculated with respect to the features (such as height, width, slant, etc.) associated with the characters. The number of features associated with each character is different (e.g. ‗f‘ had 7 features while ‗th‘ had 10 ones). A simple Manhattan distance was used in the classification phase.

VII. PROPOSED METHOD

Our method is based on converting individual character written by writers on digitally from color image to gray scale image .with every writer we are storing six key features like slant, orientation, eccentricity, of character , major axis, minor axis , total area of single word and writer’s name. As soon as new writer’s handwriting samples are taken the classifier which check through the system and can predict whether it’s known (old) writer or unknown (new) writer, along with it would also specify new writer’s handwriting has probalistic match with known writer.

With the survey we are able to conclude as how writer’s handwriting is being matched mainly for checking over forgery, handwritten thesis. Still areas like mental state of writer, as how was the flow of characters, were characters in cursive form or simply written separately does give writer’s present state of mind.

VIII. SUMMARY

We have seen through this paper as different methods used for identifying handwriting ; written by different writers . We have also seen as how our proposed method is easy to use as it only identifies writer but when further enhanced with flow, indentation of characters from baseline can help in understanding the personality of writer along with his mental state.

IX. REFERENCES

[1] Jomy John, Kannan Balakrishnan, Pramod K V, “A System for Offline Recognition of Handwritten Characters in Malayalam Script”, I J Image, Graphics and Signal Processing, MECS , vol 4, April 2013 .

[2] Ankush Acharyya, Sandip Rakshit,Ram Sarkar,Subhadip Basu,Mita Nasipuri, “Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach” ,IJSCI, vol 10, issue 2, no 2, March 2013.

[3] Md Mojahidul Islam, Md. Imran Hossain & Md Kislu Noman, “Bangla Character Recognition System is Developed by using Automatic Feature Extraction and XOR operation “ , Global journal of computer science and technology graphics & vision, volume 13, issue 2, version 1.0, 2013.

[4] Shashank Mathur, Vaibhav Aggarwal, Himanshu Joshi, Anil Ahlawat, “Offline Handwriting Recognition using Genetic Algorithm” , International Book Series “ Information Science and Computing”

[5] Seeraj M, Suman Mary Idicula, “ A survey on Writer Identification Schemes” , International Journal of Computer Application(0975-8887), volume 26, no 2, july 2011.

[6] Avani R. Vasant, Sandeep R. Vasant, Dr G. R Kulkarni, “Gujarati Character Recognition : The State of Art

Comprehensive Survey", ISSN : 0975-6760, Volume 02, issue 01 , nov 2011.

[7] D Impedovo, G. Pirlo, R Modugno. " New Advancements in Zoning –Based Recognition of Handwritten Characters", International Conference on Frontiers in handwriting Recognition, 2012.

[8] Atallah Mahmoud AL Shatnawi, Farah Hanna AL Zawaideh, Safwan AL Salaimeh, Khairuddin Omar, "Offline Arabic Text Recognition- An overview" ,WCSIT, ISSN 2221-0741,vol 1 , no-5,2011.

[9] J. Pradeep , E. Shrinivasan, S.Himavathi," Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System using Neural Network", IJCSIT, vol 3, no 1, feb 2011.

[10] Hanan A. Aljuaid, Dzulkifi Muhamad ,"Offline Arabic Character Recognition using Genetic Approach : A Survey", 2008.

[11] Qui-Feng Wang, Fei Yin, Cheng-Lin Liu, " Handwritten Chinese Text Recognition By Integrating Muliple Contexts", IEEE, vol 34, no 8, August 2012.

[12] Jawad H. Alkhateeb, Olivier Pauplin, Jinchang Ren, Jianmin Jiang, " Performance of Hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition", knowledge based systems, vol 24, February 2011.

[13] F. Shahabi, M. Rahmati , "Comparision of Gabor Based Features for Writer identification of Farsi/ Arabic Handwriting".

[14] Sargur N. Srihari, Sung-Hyuk Cha, Sangjik Lee, " Establishling Handwriting Individuality Using Pattern Recognition Techniques", IEEE, 2001

[15] G. Louloudis, N. Stamatopoulos, B. Gatos, " ICDAR 2011 Writer Identification Contest", ICDAR, 2011.

[16] Marius Popescu, Radu T. Ionescu, "The Story of the characters, the DNA and Native Language", Proceedings of English Workshop on Innovative Use of NLP for building Educational Applications, June 2013.

[17] Meenu Bhatia, "Offline Handwritten Signature Verification using Neural Network", IJAIEM, May 2013.