ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.5

HMM Segmentation Approach for Offline Cursive Handwritten Words

Sourabh Sagar[1], Sunanda Dixit[2]
M.Tech Student[1], Associate Professor[2]
Department of Information Science and Engineering
Dayanand Sagar College of Engineering, Banglore, India

Abstract:
Hidden Morkov Model (HMM) based offline cursive handwritten word segmentation method is proposed in this method. Data set consists Handwritten words which are in the cursive format images and is taken as input and these images consists of noise and these noises are removed by pre-processing method. Pre-processing method includes word image acquisition which is RGB image for further steps RGB image is converted to Gray image. Later Thresholding is applied on gray image. Thinning and Skeletonization is applied on thresholded image .Finally noise is removed from the handwritten word image and pre-processed binary matrix is shown in the form of matrix. Over segmented words are divided by potentially segmented column (PSC) and HMM method.

Keywords: HMM, RGB Image, Gray Image, Thinning and Skeletonization

I. INTRODUCTION

Handwritten words are scanned first than that image format is taken as input for the segmentation process. Binary format of the input image is formed using proper algorithm to make the further process. In real application, handwriting segmentation is used to Segment bank check and segment address of city's name in English. Handwriting segmentation plays important part in handwritten text document segmentation that is in digital format. Nowadays, Segmentation on unconstrained handwriting is a difficult challenge because there are many variant of handwriting style. Even after 30 years of research and achievement in handwriting segmentation, developing algorithms for segmentation unconstrained handwritten words text is still an open problem. Cursive Handwritten segmentation is done on many types of cursive words. There is algorithm development going on in Arabic handwriting, Chinese handwriting, English handwriting, and many languages. The significance of the piece of paper cannot be ignored towards enhancing the people's memory and in facilitating communication between people. People will write the important thing in paper later they will use it for further process. Writing important things on the Paper is still a good way of storing the data in the form of handwritten text. Most of the Historical data are present in handwritten words. In today's life also storing information on paper plays important role, in banking system, Postal Department and Insurance companies etc. Many researchers are attempting to simulate intelligent behavior and mimic the human brain's ability to read and recognize the handwritten or printed characters from the paper surface so that the computer can understand this script and process the data. Algorithm based cursive handwritten word segmentation becomes easy for further process of analysis. Though Optical Character Recognition (OCR) technology is developing rapidly nowadays, offline cursive English handwritten word segmentation is always an open problem in segmentation, mainly due to the cursive nature of English letters, the variations in shape and size, and the existence of dots and diacritical marks. As ligatures and overlaps often occur within a word, English words are usually split into one or more sub-words. In the English alphabet, there are 26 basic characters, whose shapes depend on their position in the words or sub-words. Specifically English character consists of The 21 consonant letters and 5 vowels.

II. LITERATURE SURVEY

Busam A et.al [01] proposed Arabic character segmentation for high accuracy of character recognition. They used Arabic heuristic segmented (AHS) algorithm for segmentation process. The AHS performs three operations first one it removes dots second one is detection of ligature detection and third one is additional methods. Dots removing make the less error in bad segmentation point. Ligature detection calculates the distance between foreground and background pixels of word image histogram. An additional method includes close/open holes detection. Results show improved segmentation. Rajiv Jain et.al [02] proposed a new method for writer identification, which emulates the approach taken by forensic document examiners. It mixes the shape and curvature features. For input word generate a pseudo-alphabet. Calculate similarity in term of distance between the pseudo and input words. There approach achieves a Top-1 identification rate of 96.5% on the benchmark IAM dataset, reducing the error rate of previous approaches by 50%.
Hesham M. Eraqi et.al [03] proposed a segmentation of Arabic handwritten words. They used new algorithm for Offline Arabic handwritten segmentation which is Douglas-Peucker algorithm creates linear curves based on standard characters direction. Pre-processing is performed to remove the noise. The segmentation is tested with 1400 Arabic handwritten words. IFN/ENIT database is used. A result shows the improved segmentation. Ashok Kumar Pant et.al [04] proposed off-line Nepali handwritten character recognition using neural networks.

Features are extracted from Nepali handwritten character images. Accuracy and efficiency are analyzed by classifier. They used the data set Nepali handwritten numerals, vowels and consonants. Recognition accuracy for numeral dataset 94.44%, Recognition accuracy for vowel dataset 86.04% and Recognition accuracy for consonant dataset 80.25% is obtained. Saeeda Naz et.al [05] proposed identification of Urdu language written in Nasta'liq writing style. Being a cursive nature, Urdu has no standard dataset available publicly. The basic motive of preparing UCOM offline database is to compile Urdu text and make it available to research community free of cost. Evaluate the UCOM dataset they took 50 text line images as train dataset and 20 text-lines as test dataset. The 0.04~0.06% error was reported on subset of UCOM offline dataset. It is planned to extend their database up to 300 writers. Currently, UCOM database almost covers all characters with different variations in addition to Urdu numeric data. At first phase, data was gathered from 100 individuals. As the dataset is tagged with user identity, thus it is used for writer identification. Other future task includes writer identification, apply different feature extraction approaches, and apply different classifiers to recognize the text and word recognition with the use of dictionary and language modeling. Ning Li et.al [06] proposed HMM based character identification. Training is done by Baum-Welch Algorithm segmentation is done by Viterbi Algorithm. IFN/ENIT database is used for 161 models. Results show that there proposed features can make good relationship between adjacent characters and are sufficiently robust, especially when characters are shifted up or down and when the handwriting width varies. Patrick Doetsch et.al [07] proposed a English and French word identification technique to exhibit a adapted topology for long short-term memory. A neural network is used for the classification purpose. They further propose a well-organized training framework based on mini-batch training on series plane united with a series chunking approach. They used English and French handwriting data set. Training is done neural network models which outperform state of the art recognition results. Made Edwin Wira Putra et.al [08] proposed a novel system Research in offline handwriting recognition for unconstrained text remains a difficult challenge. Some problems such as noise in image skew of text, cursive letters, and various handwriting styles are at rest an unwrap problem. Previously Slant is corrected by Kimura algorithm used for slant correction with the help of slant prediction. Therefore, there paper proposes to create handwritten character into graph with string representation based on structural approach. The purpose is to improve identification. levenshtein distance is used for likeness calculation in the distance among graphs calculation. An ETL-1 AIST database is used. Levenshtein distance accuracy is 84.69% on digits and 67.01% on alphabet with 5% size of data for training and value 10 for string representation length. Results show the improved slant correction. Pankaj Kumawa [09] proposed SVM-HMM based Offline Handwriting recognition. Persons handwriting varies from time to time Hence HMM is method is good usage. It creates large number of patterns for the same character. However, the performance of the system depends entirely on the feature vectors. SVM gives better efficiency and HMM gives good performance. Combination of SVM-HMM gives better accuracy. Results shows that high performance. Shanjana C et.al [10] proposed Malayalam handwritten character segmentation. Input images are taken from dataset for segmentation. They segment the touching characters in Malayalam language And also it segment contains breaks in the characters and different handwriting styles, fonts etc. Segmentation of character is done by applying vertical projection and connected components are also removed. Features are extracted by using algorithm and extracted features are fed to the classifier for identification of the character. Then the unique features for identifying each character are extracted and given to a classifier.

III. METHODOLOGY

In this project proposing offline cursive handwritten word segmentation is used by applying HMM.OCR process is used for HMM based word segmentation. Cursive word segmentation procedure is little bit complex suitable for the different cursive word format. Tallness and size of the word image is calculated after completion of pre-processing method. The word picture appears in matrix format. The values of the matrix are stored for further process. All these identified columns are termed as PSC (Potential Segmentation Columns).In this paper proposed a HMM based cursive handwritten word. Figure 1 shown is the structural design of proposed system.

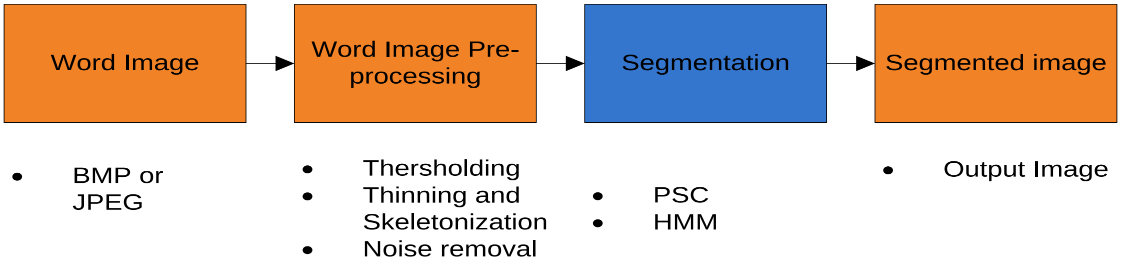


Figure.1. Architecture of Proposed System

3. 1 Word Image Pre-Processing

To make proper identification pre-processing is done on input image. Output of pre-processed image is noise free. Cursive Handwritten words are taken as Input images from local database. RGB image is taken as input for pre-processing method. Further Gray scale conversation done to a RGB image. Further Thresholding is done on gray image. Further thresholded image is applied for Thinning and Skeletonization finally noise is removed from the word image.

Handwritten Word Image Acquisition

In image acquisition from the dataset cursive handwritten word images are taken as input. The input word images are in JPEG or BMP formats. These images are taken from digital camera or by a scanner. Scanned text is stored as an image.

A. RGB to Greyscale Conversion

24 bit original RGB image and is converted to gray scale image by considering the weighted sum of all the 3 colors of RGB image.

If $f(x\ ,y)$ is RGB image than Gray conversation RGB image is given by Eq.1

$$g(x\ ,y) = 0.2989 * f_R + \ 0.5870 * \ f_G + 0.1140 * \ f_B \qquad (1)$$

Wheref_R, f_G, f_B are Red, Green and Blue colors of the RGB image $f(x\ ,y)$ respectively.

B. Thresholding

Binary matrix format of the gray scale image is called as Thresholding. This binary matrix consists of 0 and 1 where 0 is background image and 1 is foreground image. Thersholding is done in two approaches

- Global Thresholding
- Local Thresholding

Entire pixel of the image is taken in global thresholding and that entire pixel value is taken and divided each pixel by that entire number. It is performed to check it is foreground pixel or background pixel. In local thresholding based on neighbor pixel calculating many threshold values for each pixel. In proposed method gray thresh inbuilt MATLAB function is used for local thresholding to alter the input handwritten character image into binary.

C. Normalization

Method of converting all the variable size input images to fixed size images is called as size normalization. Normalize the all input images to the predefined size.

Normalization of the word image is finished by 2 methods

I. **Thinning:** Method of changing a pattern from one form to another with less thickness is called as thinning. Threshold image output is taken as input for the thinning.

II. **Skeletonization:** Method of making reduced shape of order 1 pixel without changing the important formation of the picture is called as skeletonization. Transformation from one form to reduced form deletes boundary points of a area of the object.

D. Noise Removal

Noise may add in input image during the Scanning process. This method is essential to remove the noise. This method makes next process smooth. MATLAB's built-in function 'bwareaopen' is used to take away the small objects from the output of skeletonized image.

3. 2 Segmentation Technique

Many segmentation techniques are developed because of different languages have different segmentation methods are present. One language is unlike from another language.

A. Potentially Segmented Column (PSC)

Foreground values from the pre-processed images the form of matrix is represented. Threshold value is fixed to pre-processed image for segmentation process to defeat the over segmentation problem. After applying threshold value to the pre-processed image the column is called as PSC.

Over segmentation occurs in 3 cases

1. two characters in the word image are not touching each other
2. two characters in the word image are connected by a ligature
3. characters are Open Characters

For case 1 addition of foreground pixels of the columns in this area are 0.For case 2 sum of Foreground pixels in these columns crossing this ligature are 1.For case 3 sums of foreground pixels in these columns is 1. Experimenting several times the rate of threshold is situated to a value 7 to minimize the over segmentation problem. Over segmented images are separated by red color if it is less or equal to distance 7.later PSC are changed to a single SC (Segmentation Column). Below Figure 8 represents segmented image.

B. Hidden Markov Model (HMM)

HMM is used for segmentation process. Cursive handwritten word Segmentation method is completed by applying viterbi algorithm from HMM. Majority of liable sequence of states is determined by viterbi algorithm through HMM. In HMM true states are hidden. Observe the state indirectly predicting true state. Observe the character which is non similar exact same thing as the state. Figure 2 shows the viterbi algorithm steps.

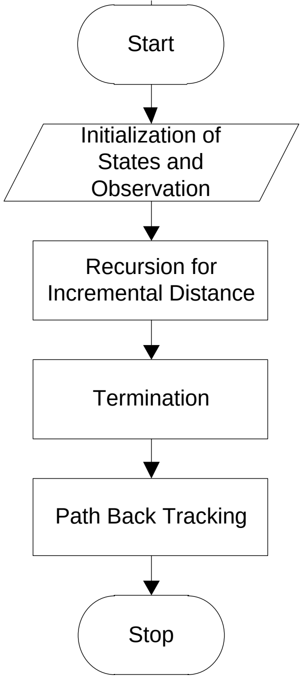

Start
Initialization of States and Observation
Recursion for Incremental Distance
Termination
Path Back Tracking
Stop

Figure.2. Viterbi Algorithm

The steps followed during the implementation of the projected segmentation method are mentioned below:

Proposed System

Input: Cursive Word Image

Output: Segmented letters

Start

Step 1. Pre-processing is done to remove the noise in the image by performing thresholding, binarization, thinning, and noise removal and cropping. This pre-processed word image is taken for the segmentation in Figure 3 (a).

Step 2. Making white pixels that is foreground pixels as the black pixel that is background pixels and the process is called as inverting. Count the only the foreground white pixels represented by 1 in the matrix of inverted image output in each column of the word image as shown in Figure 3 (b)

Step 3. To display PSC (Potential Segmentation Columns) convert the binary format to a RGB format in different color other than black and white and is exposed in Figure 3 (c).

Step 4. Over segmented words in the picture is separated by PSCs in red color as shown in Figure 3 (d). The addition of foreground white pixels is 0 by setting threshold point is a PSC in each column of word image. Potentially segmented column vertically cuts the word image.

Step 5. Vertically segmented word foreground pixel is set threshold 7 .if the distance is equal or less than 7 are separated by red color later that large red color is fused into a single column called as Segmentation Column (SC) as shown in Figure 3 (e).

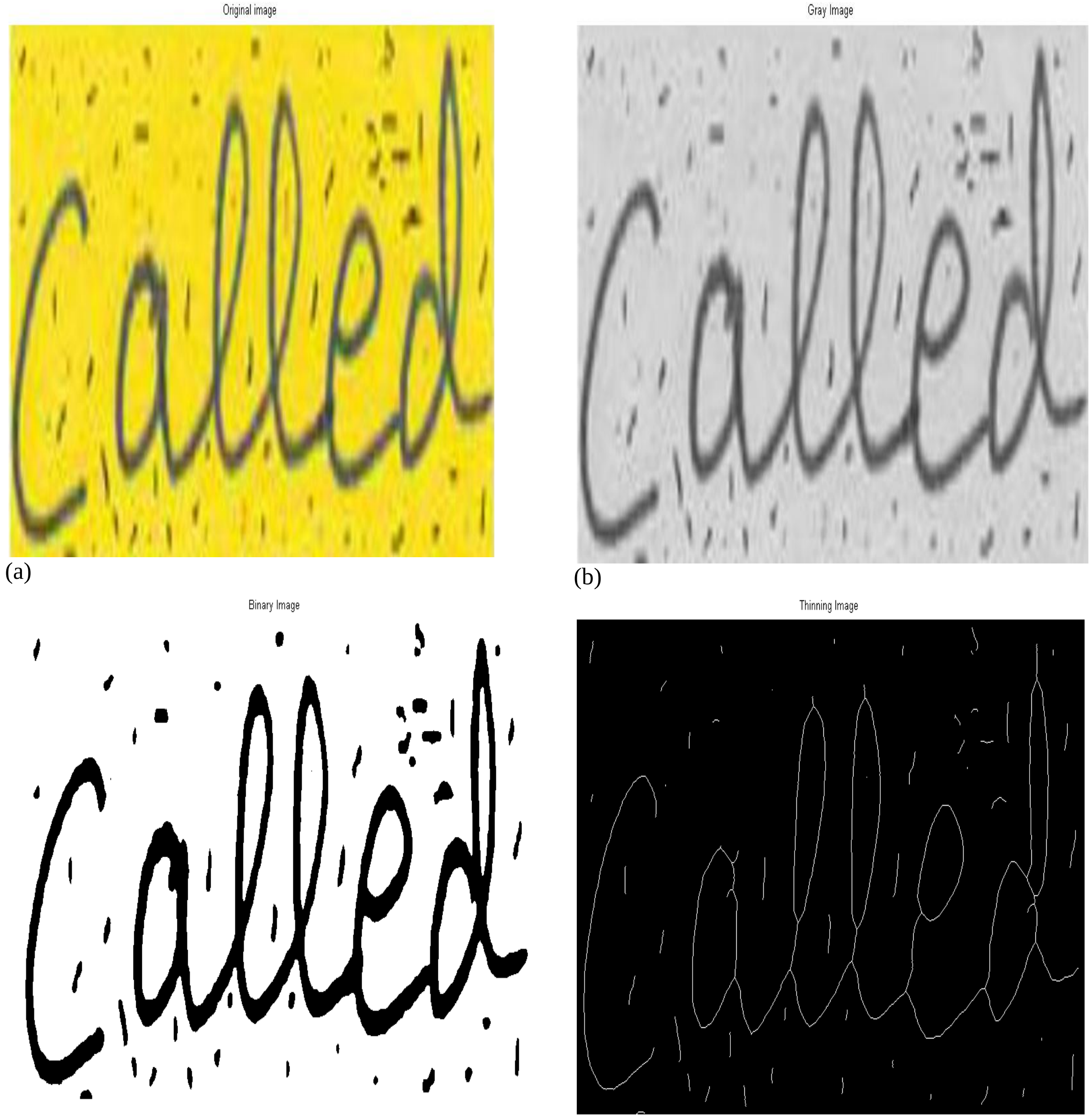
Step 6. The background has been changed to white background from black background in sort to get the concluding segmented picture as shown in Figure 3 (f).

Step 7. The cursive handwritten words segmentation done by applying HMM virtbrie Algorithm as shown in Figure 3 (g).

Stop

IV. RESULTS

Figure 3 represent the experimental result of proposed method. Original image for proposed method is shown in figure (a), next this input image converting to gray color image, using RGB color image processing method which is given away in figure (b), next binary image is shown in figure (c), Thinned image is given away in (d), pre-processed image is given away in (e), after applying PSC image is given away in figure (f), finally get a Retrieved segmented Result which is given away in figure (g) respectively.



(a) (b)

(c) (d)

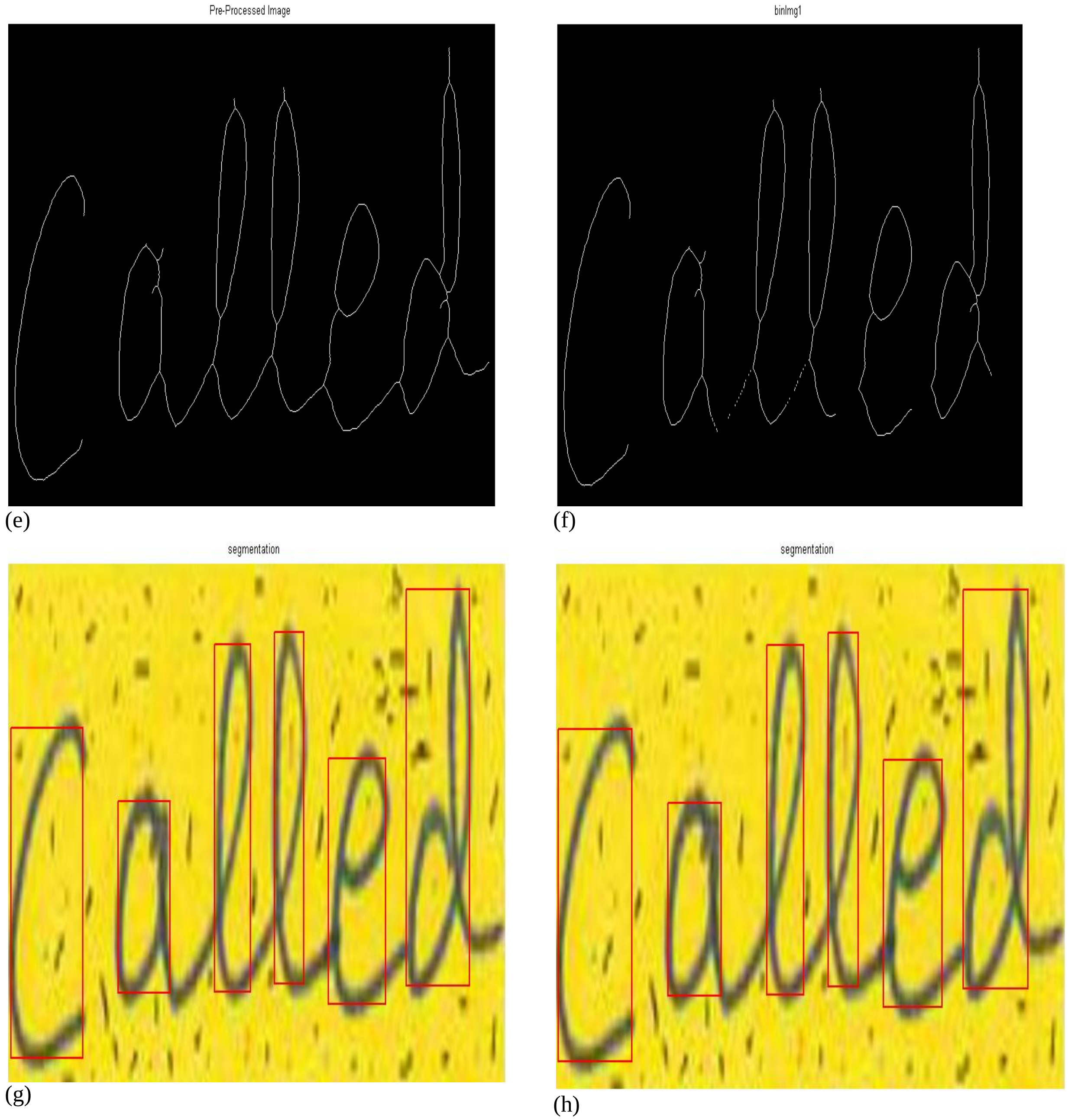


Figure.3. (a) Input Image; (b) Gray Image; (c) Binary Image; (d) Thinned Image (e)Pre-processed Image ; (f) PSC Image ; (g) Segmented Image;(h) HMM Segmented Image

V. CONCLUSION AND FUTURE SCOPE

In graphology and forensic science offline cursive handwritten character segmentation plays important role. Character segmentation is difficult aspect because user's handwriting tends to differ depends on type of Pen they use, the writing style and surface so on. Beside, Handwriting is not considered as part of unique biometric property. Segmentation cursive handwritten word is unlike other segmentation method. In this occupation presented a new method of HMM based segmentation method. This is used for improving over segmentation of cursive English handwritten characters. Image processing is improved, useful and emerging research fields in engineering sector, the image processing algorithms such as segmentation techniques, features extraction module are again changing every day. In future the for the purpose of enhanced machine learning and computer vision algorithm develops the system performance.

VI. REFERENCE

[1]. Al Hamad, Husam A, "Over-Segmentation of Handwriting Arabic Scripts Using An Efficient Heuristic Technique", Wavelet Analysis and Pattern Recognition (ICWAPR), IEEE, pp. 180-185, 2012.

[2]. Jain, Rajiv, and David Doermann, "Writer Identification Using An Alphabet Of Contour Gradient Descriptors", Document Analysis and Recognition (ICDAR), IEEE, pp. 550-554, 2013.

[3]. Eraqi Hesham M, and Sherif Abdelazeem, "A New Efficient Graphemes Segmentation Technique for Offline Arabic Handwriting", Frontiers in Handwriting Recognition (ICFHR), IEEE, pp. 95-100, 2012.

[4]. Pant, Ashok Kumar, Sanjeeb Prasad Panday, and Shashidhar Ram Joshi, “Off-Line Nepali Handwritten Character Recognition Using Multilayer Perception And Radial Basis Function Neural Networks”, Third Asian Himalayas International Conference . IEEE, pp. 1-5, 2012.

[5]. Ahmed Saad Bin, “UCOM Offline Dataset-An Urdu Handwritten Dataset Generation”, Int. Arab J. Inf. Techno, pp.239-245, 2017.

[6]. Li N, Xie X Liu W, and Lam K M, “Combination of Global and Local Baseline-Independent Features for Offline Arabic Handwriting Recognition”, In Pattern Recognition (ICPR), 2012 21st International Conference, IEEE, pp. 713-716, 2012.

[7]. Doetsch, Patrick, Michal Kozielski, and Hermann Ney, “Fast and Robust Training of Recurrent Neural Networks for Offline Handwriting Recognition”, Frontiers in Handwriting Recognition (ICFHR), 14th International Conference, IEEE, pp. 279-284, 2014.

[8]. Putra, Made Edwin Wira, and Iping Supriana, “Structural Offline Handwriting Character Recognition Using Levenshtein Distance”, In Electrical Engineering and Informatics (ICEEI), 2015 International Conference , IEEE, pp. 31-36, 2015.

[9]. Kumawat P, Khatri, A and Nagaria B, “Offline Handwriting Recognition Using Invariant Moments and Curve Let Transform with Combined SVM-HMM Classifier”, In Communication Systems and Network Technologies (CSNT), 2013 International Conference, IEEE, pp. 144-148, 2013.

[10]. Shanjana C, and Ajay James, “Character Segmentation in Malayalam Handwritten Documents”, Advances in Engineering and Technology Research (ICAETR), International Conference, IEEE, PP.1-4, 2014.

[11]. Sharma, Om Prakash, M. K. Ghose, and Krishna Bikram Shah, “An Improved Zone Based Hybrid Feature Extraction Model For Handwritten Alphabets Recognition Using Euler Number”, International Journal of Soft Computing and Engineering, PP.504-508, 2012.

[12]. Sarker, Goutam, Monica Besra, and Silpi Dhua, “A Programming Based Handwritten Text Identification”, Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances, IEEE, 2015.

[13]. Choudhary A, Rishi R, Ahlawat S, “A New Character Segmentation Approach for Off-Line Cursive Handwritten Words”, Procedia Computer Science, pp. 88-95, 2013.

[14]. Huang J, Liu Z, and Wang Y, Joint Scene Classification and Segmentation Based On Hidden Markov Model, IEEE Transactions on Multimedia, Vol.7,No.3, pp.538-550,2015.

[15]. Wshah, Safwan, Gaurav Kumar, and Venu Govindaraju, “Multilingual Word Spotting In Offline Handwritten Documents”, Pattern Recognition (ICPR) 21st International Conference, pp. 310-313, IEEE, 2012.

[16]. Xinyan, Cao, and Zou Yingyong, “Segmentation of Chinese Handwritten Text”, Computer Science and Network Technology (ICCSNT), 2nd International Conference, pp. 367-370, IEEE, 2012.