ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Offline Handwritten Word Recognition using Multiple Features with SVM Classifier for Holistic Approach

Shruthi A [1], M S Patel [2]

M.Tech Student, Department of Information Science and Engineering, DSCE, Bangalore, India

Research Scholar, Department of Information Science and Engineering, DSCE, Bangalore, India

ABSTRACT*:* Offline Handwritten Word Recognition (HWR) is an interesting research field in the domain of digital image processing. Offline Handwriting recognition has achieved a high attention for many years due to its key contribution in the digital libraries extension. Offline recognition of handwritten words by multiple and different writers is still an exposed problem, because of multiple challenges including strong variability in writing style and each person have their own control over writing. English Handwritten words recognition is most important now days because English script is widely used language in the world and most of the countries used as official language. To identify the handwritten words, each word take is an individual entity so holistic approach is used. In the proposed method reports multiple features namely: density features, long run features and structural features for extraction in the input handwritten document image. Next phase is classification and this phase is most important for word recognition, for classification Support Vector Machines (SVM) classifier is used. To estimate the performance, own dataset is created and sample words for testing collected from multiple people. There are so many applications in HWR like, postal address identification, historical document conversion, signature verification and etc. This presented work gives 88.13% of recognition rate.

KEYWORDS*:* Handwritten Word Recognition (HWR), dataset, density features, long run features, structural features, classifiers.

I. INTRODUCTION

Handwriting word recognition of scanned or photographed document image is still a most difficult and unresolved issue in computer science and in the field of research. Handwritten text identification is most essential because historical documents are arising strongly in current years. To save the historical documents for upcoming generation, identify the postal address, writer identification, bank check recognition and for other applications there are so many researches are going on in the field of image processing.

Image processing is a method of processing an image using several operations of mathematics with any signal processing, here input may be an image, video, photograph, or scanned document, after preprocessing output may be an image or group of related parameters of image. In image processing there are so many categories like, Medical Image Processing, Satellite Image Processing, Document Image Processing and etc. Document Image processing is used to process the handwritten documents for analyze, recognition and produce the secondary information for further use. Document image processing involves mainly two approaches namely: Offline Recognition and Online Recognition. Overall approaches of document image processing are shown in the fig. 1.

Online Recognition is used to recognize the handwritten documents where writing is done on an electronic notepad, tab and etc, by a special pen and at that time temporal information is stored like movement of the pen on notepad will give the shape of the letter or word, velocity of the pen and these information used in the classification phase [1].

In Offline Recognition is to identify the text documents, this approach are used for printed and handwritten document recognition. Recognition is done after a text is written on a paper. Printed document recognition is easy

DOI: 10.15680/ijircce.2015.0306138

compare to handwritten text because in printed document is in the standard format like font style, font size and other details may help in recognition process. Handwritten words identification is difficult because each writer has unique style of writing and they have control over writing.

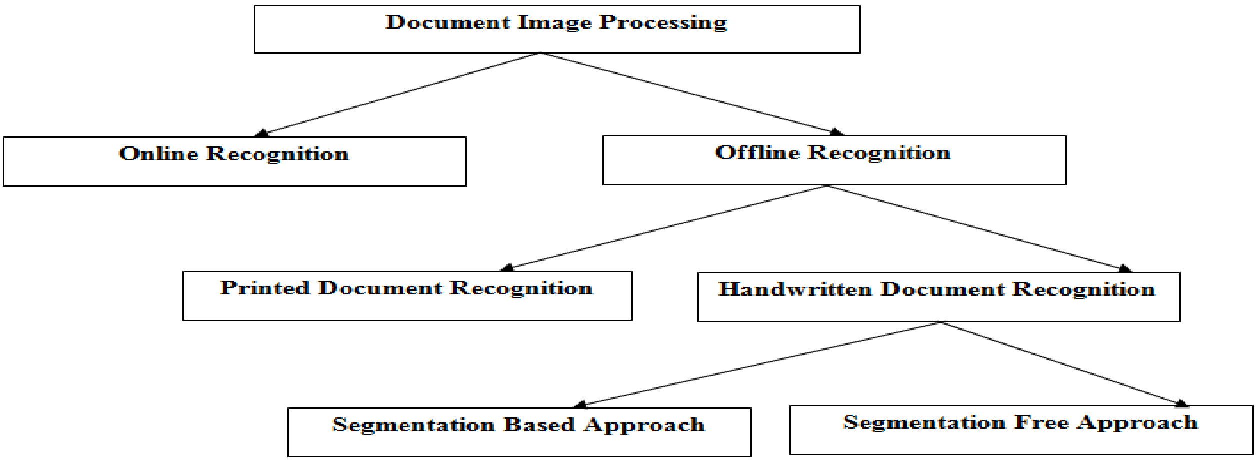
Document Image Processing
Online Recognition
Offline Recognition
Printed Document Recognition
Handwritten Document Recognition
Segmentation Based Approach
Segmentation Free Approach

Fig. 1 Approaches of document image processing

Handwritten Word Recognition (HWR) involves segmentation based approach and segmentation free approach. In Segmentation Based Approach is to divide word into characters and recognize the each letter in the word. Sometimes segmentation is more difficult because of cursive nature of the words. In Segmentation Free or Holistic Approach is treat the word as a whole entity and recognize the word based on the shape, structure and other features of the word. Handwritten Word Recognition (HWR) is used in many applications like signature verification, bank cheque verification, forensic department, historical documents recognition, postal address identification and many more [2,3].

In recognition system involves major five steps especially, Data collection, preprocessing, segmentation, features extraction and classification. For data acquisition many researchers used standard data sets available globally and some of researches uses own data sets for processing. In preprocessing phase many mathematical and morphological operations are applied on input document image for gray scale conversion, normalization, binarization, baseline detection, skew correction, slant detection, slant correction, noise removal and etc. If the input document image consists of a sentence than that will be segmented into words than each word is consider as distinct substance. Than word is taken for further process, features like density features, structure based features, hierarchical features and other features are extracted from the input image. Which are the features extracted those are comparing with the training image features if the matching class is presented than that image is recognized and for classification process classifiers are used like support Vector Machine (SVM), Neural Networks, K Nearest Neighbor (KNN), Hidden Markov Model (HMM) and etc [4,5].

II. RELETED WORK

There are so many researches on handwritten word recognition holistic approach are held but still it is a open problem for examine. Handwritten word recognition researches are done in many languages like English, Arabic, Hindi, Urdu, Devanagari, Kannada, Telugu and other languages.

Douglas J. Kennard et .al brings the method wordwarping to recognize the handwritten words. To compute 2-D geometric warps they used automatic image morphing and in this method stroke of the each image in training phase is extracted and those are stored, these strokes are used to align the stroke in testing image for this automatic morphing is used. This method is to compare the two words how those are similar to each other and for this process distance maps are used [6].

Anuja Naik, M S Patel proposed best feature extraction method to recognize the handwritten words. In preprocessing steps skew and slant correction are performed. Least pixel in each column of image finds out for skew correction and to find the angle input image is rotated as per rotation angle. Slant is an angle so that find the image angle from vertically clockwise, threshold image contour and connected components in chain representing edges of stroke and if those detected edges are close to vertical than it is considered as slant. Slant is too corrected by related transformation. To determine the baseline for image lower and upper black pixels are used. In skeletonization, Gaussian filters are applied on input image to remove the noise with that convolution is used for smoothing of image. For reduction of width of stroke, thinning and erosion algorithms are applied. In feature extraction phase they extracted structural features and for classification Euclidean distance method is applied [7].

Soulef Nemouchi et .al presented multiple classifiers for handwritten words recognition in Arabic script and for the experiments they used Algerian city names. In this method, prime attention on feature extraction and classification phases. There are three features are extracted namely Zernike moments, structural features and Freeman code is determined by contour of the image. In classification phase use the four classifiers namely, K Nearest Neighbor algorithm (KNN), Fuzzy C-Means algorithm (FCM), K-Means algorithm and Probabilistic Neural Network (PNN). In the experimental results this method carried out 80% of accuracy [8].

Ahlam Maqqor et .al presented offline cursive handwritten Arabic word recognition used by multi-stream HMM approach. Two methods are used to extract a set of simple statistical features. Thresholding or binarization, normalization, filtering, smoothing and skew detection operations are applied to text image to extract the word feature simplify in preprocessing phase. In feature extraction involves two phases, In sliding window method features are extracted from right to left of binary image and image is divided into N number of sliding windows, in Vertical Horizontal 2-dimentional (VH2D) method projections are take place with angle of 45 and 135. Multi-stream approach is used and that involved multi-classifiers, multi-model approach, multi-band approach and multi-stream formalism [9].

Youssouf Chherawala et .al proposed feature design for offline Arabic handwriting recognition. They evaluate the automatically learned features performance and that is compared with handcrafted features. The recognition model is based on the connectionist temporal classification (CTC) neural networks and long short-term memory (LSTM). HMM model is used as classifier for this method. Multidimensional LSTM network is able to automatically learn features from the input document image. The IFN/ENIT database is used as benchmark for Arabic word recognition [10].

Silky Bansal, Munish Kumar, and Mamta Garg proposed an approach for recognize handwritten city name written in Gurumukhi script for postal automation. In this approach, offline holistic approach used in which they considered the whole word as an individual entity. For recognizing words they used a tree-diagonal feature extraction technique in which a tree structure comprises of zoning and diagonal feature extraction technique. For classification of images, Support Vector Machine (SVM) and k-Nearest neighbor (KNN) classifiers [11].

Anne-Laure Bianne-Bernard et .al proposed HMM modeling with dynamic and contextual information for HWR. To design the dependent units, based on the decision tree clustering a state tying method is instigate here. This method is applied for recognition of handwritten documents and for experiments; three standard datasets are used like, Rimes, IAM and openhart [12].

Ankush Acharyya et .al proposed HWR holistic approach using MLP based classifier. In the holistic approach, for recognition each word is take it as whole based on the shape of the word image. To extract the features input image is divided hierarchically with depth of 5. Neural network based classifier used to classify word images belonged to different classes [13].

B Gatos et .al proposed efficient off-Line cursive handwriting word recognition. In this approach, for normalization of input image there are two modes combining and robust hybrid characteristics are extracted. In support of preprocessing, skew and slant are corrected, thickness of the stroke is normalized. Hybrid features extraction there are two different features are combined. Word image is divided into number of zones and for each zone a densities are calculated. Next, from the upper and lower projections of the word image area is calculated [14].

III. PROPOSED METHOD

In our proposed method, handwritten words are recognized for English script and it involves four major steps those are Data acquisition, Preprocessing, feature extraction and classification. Architecture of proposed system is shown in the below figure 2. In the proposed method involves two phases: training and testing. In training phase, first step is to take the scanned document image as an input and it is going for further process. In preprocessing, input image is converting into binary image and involves many operations for image enhancement. Multiple features are extracted in feature extraction method and similar features are made it as a class in classification. In testing phase, extracted features on test image are comparing with the classes of trained features; if features are matched then image is recognized otherwise image is not recognized correctly.

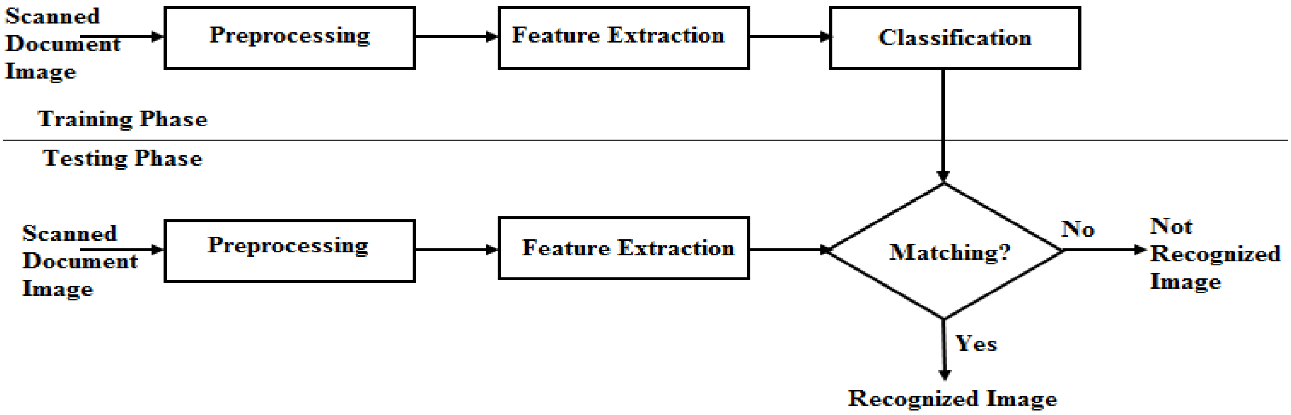

Fig 2: Architecture of proposed method

A. Data Acquisition:

In many researches uses standard data set for their experiments, in our proposed method own data set is created and used for experiments. Handwritten words are collected from more than 200 people with the all age groups and from different professions. Karnataka districts names are written by citizens for data collection and handwritten document is scanned through scanner than each word is cropped and resized.

B. Preprocessing:

In the preprocessing step is to occupy several operations on input document image for image enhancement. Input image is converting into grayscale image, skew is correction, grayscale image is converting into binary image and block corner are removing for the next process shown in figure 3. Skew correction: first for input image, skew is detected than only it is going for correction. Baseline of the word image is detected and based on the projection of the image, angle of the image is calculating by the following formula.

$$Slope=atan(p(1))*180/pi$$

After the angle calculation, image is rotated in counter clockwise direction around its center point than new baseline is given for skew corrected image.

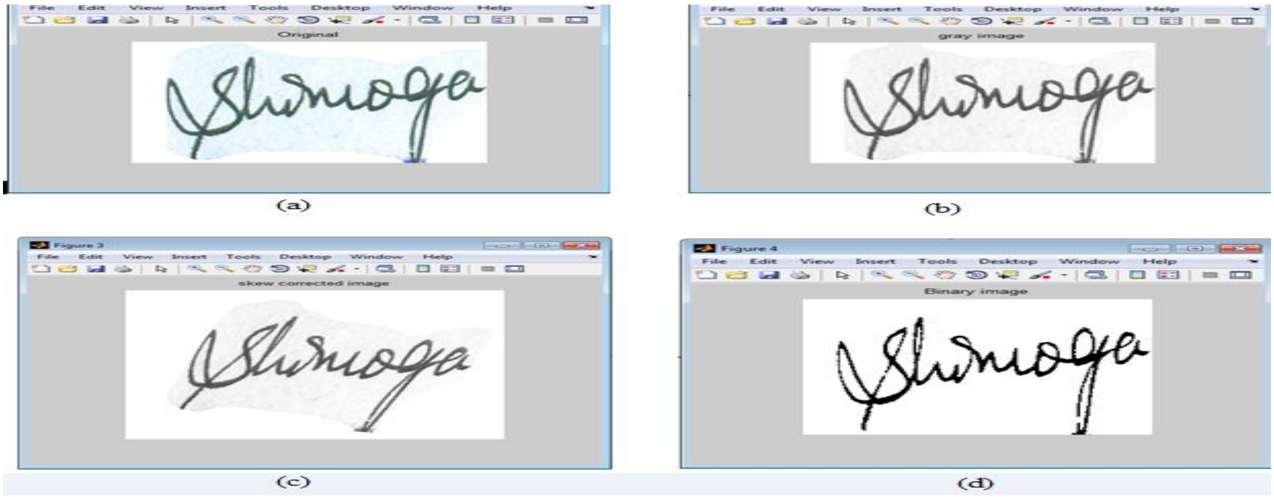

Fig.3 (a) Input image (b) gray scale (c) skew corrected (d) binary image

C. Feature Extraction:

Feature extraction stage is major step because handwritten words recognition is based on the extracted features of text image. In this stage multiple features are extracted and stored in dataset. Extracting features are:

- Density Based Features
- Structural Features
- Longest Run Features

In density based feature, design of feature set is most challenging task when handwritten words are recognizing. Difficulty is increases if the document images are handwritten words. For this problem, in the proposed system, density based, structural and long run features are used. For density based features lower pixels are extracted from bottom to top of the image and for every column first black pixel is extracted that black pixel is corresponding to the column of lowest one. Density features are distributed on the image so that foreground pixel densities are takeout from the vector image. Structural features are based on the shape of the image like, area, perimeter, major axis, minor axis, roundness, form factor and compactness of the image. In four directions longest run features are calculating, longest consecutive black pixels lengths are adding for each row and column. Center of gravity for each word is generating to discriminating the information of an image of text document.

D. Classification

Features extracted on each word image are stored, similar features are classified and labeled in training phase. In this step testing image features are extracted and comparing with training classes than assign the matching class for this process Support Vector Machine Classifier (SVM) is used.

IV. EXPERIMENTAL RESULTS

To test and estimate the proposed work the own dataset is used. For the experiment, we have taken 30 districts names of Karnataka from more than 50 people this can apply in postal services. In training, 35 samples are taken from each districts, the given below table 1 shows the results for varying data samples like 10, 20, 30 and 35.

Data samples	Districts names (DN)								Average accuracy (%)
	DN1		DN2		DN3		DN4		
	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative	True Positive	True Negative	
10	9	1	9	1	8	2	9	1	87.5
20	19	1	17	3	17	3	18	2	88.75
30	27	3	26	4	25	5	26	4	86.66
35	32	3	31	4	30	5	28	7	89.62
Total average accuracy (%)									88.13

Table.1 Proposed Recognition performances

Success Percentage of the proposed method is calculated by the following formula

$$\text{Success Percentage} = \frac{\text{Number of correctly recognized words}}{\text{Total number of words in testing}} * 100$$

V. APPLICATIONS

There has been significant growth in the application of off-line handwriting recognition during last decade.

- Signature Verification
- Forensic Science
- Bank Check Recognition
- Handwritten Address Interpretation
- Historical Manuscript conversion etc

VI. CONCLUSION

Handwritten word recognition is challenging task and it requires higher level of accuracy. Most of the techniques used for HWR are script dependent and holistic approach is avoiding the challenges of character segmentation. Handwritten words identification is difficult and open problem because writing style is different from one another. For this issue, Density based, structural and longest run features are extracted for classification. Support vector machine classifier is used for classification. For the experiments dataset is created and used here. Applications of HWR are extent and used in many fields. Proposed method achieves average of 88.13% of accuracy.

REFERENCES

1. Pooja Yadav & Ms. Neha Popli," Handwriting Recognition System – A Survey," IJETST Volume 01 ,Issue 03, Pages 405-410, ISSN 2348-9480,May 2014.

2. N.Azizi, N.Farah and M.Sellami, "Off-line Handwritten Word Recognition Using Ensemble of Classifiers Selection and Feature Fusion," Journal of Theoretical and Applied Information Technology,2005 – 2010.
3. Jino P. and Kannan Balakrishnan, "HWR for Indian Lnguages: A Comprehensive Survey," Econometric Institute research papers,"Feb 2014.
4. Ashwin S Ramteke, Milind E Rane, "A Survey on Offline Recognition of Handwritten Devanagari Script," International Journal of Scientific & Engineering Research Volume 3, Issue 5, ISSN 2229-5518, May-2012.
5. Yousri Kessentini, Thierry Paquet and AbdelMajid Benhamadou, "A Multi-Stream HMM-Based Approach for Off-line Multi-Script Handwritten Word Recognition," journal pattern recognition letters volume 31, issue 1, January 2010.
6. Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg, "Wordwarping for Offline Handwriting Recognition," ICDAR, Beijing, September 2011.
7. Anuja Naik and M S Patel, "Offline English Handwritten Word Recognizer Using Best Feature Extraction," IJACTE Volume 3, ISSN 2319-2526, Issue -2, 2014.
8. Soulef Nemouchi, Labiba Souici Meslati and Nadir Farah, "Classifiers Combination for Arabic Words Recognition Application to Handwritten Algerian City Names," ICISP, volume 7340, Pages 562-570, Agadir Morocco, June 2012.
9. Ahlam Maqqor, Akram Halli, and Khaled Satori, **"**A Multi-Stream HMM Approach to Offline Handwritten Arabic Word Recognition," International Journal on Natural Language Computing (IJNLC), Vol. 2, No.4, Aug 2013.
10. Youssouf Chherawala, Partha Pratim Roy and Mohamed Cheriet, "Feature Design for Offline Arabic Handwriting Recognition: Handcrafted vs Automated?," ICDAR, Washington, ISSN 1520-5363,2013.
11. Silky Bansal, Munish Kumar, and Mamta Garg, "A New Approach for Handwritten City Name Recognition," ICAET, ISBN: 978-1-63248-028-6, 2014.
12. Anne-Laure Bianne-Bernard et.al, "Dynamic and Contextual Information in HMM Modeling for Handwritten Word Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence VOL. 33, NO. 10, Oct 2011.
13. Ankush Acharyya,Sandip Rakshit,Ram Sarkar,subhadip Basu and Mita Nasipur, "Handwritten Word Recognition using MLP based Classifier: A holistic approach," IJCSI,vol.10,issue 2,no 2,march 2013.
14. B. Gatos, I. Pratikakis, A.L. Kesidis and S.J. Perantonis,**"**Efficient Off-Line Cursive Handwriting Word Recognition,". Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, Oct. 2006.
15. Andreas Fischer, Emanuel Indemuhle, Horst Bunke, Gabriel Viehhauser and Michael Stolz, "Ground Truth Creation for Handwriting Recognition in Historical Documents," 9th IAPR international workshop on document analysis systems, pages 3-10, ISBN 978-1-60558-773-8, USA, 2010.