Offline Handwritten Word Recognition: A Survey

M S Patel
Research Scholar
VTU Belgaum
Department of Information Science and Engineering
DSCE, Bangalore

Rohith Kumar
M. Tech Student
VTU Belgaum
Department of Information Science and Engineering
DSCE, Bangalore

Abstract- **Offline Handwritten Word Recognition (HWR) is one of the interesting fields in the image processing and pattern recognition application for the past few decades. Handwriting is the most natural mode of collecting, storing, and transmitting information which serves for communication of humans and machines. Many approaches are presented to recognize the handwritten documents or paper. These approaches focus on how we recognize hand written words and documents. Selecting the appropriate feature extraction methods and classifier is the most important thing in the handwritten word recognition. Hence achieving good recognition and better accuracy. This paper provides an overview of offline handwritten word recognition in English, Arabic, Hindi and Kannada languages.**

Keywords- **Offline Handwritten Word Recognition (HWR), pattern recognition, feature extraction, classifier.**

I. INTRODUCTION

Image processing is the modification or manipulation of a digitized image in order to enhance its quality. Different types of image processing are satellite image processing, medical image processing, document image processing, etc. Document image processing is the process of analyzing the documents and the preparation of secondary information that produces the most substantive elements of the documents contents. There are mainly 2 approaches in the document image processing. Online and offline approach.

In Online HWR the trajectories of pen tip movements are recorded and evaluated to identify intended information. Here writing is done using a stylus on an electronic notepad or a tablet where temporal information, such as the position and velocity of the pen along its trajectory, is available to the recognition algorithm. On the other hand, offline

HWR deals with the recognition of handwritten words after it has been written.

Further offline recognition can be split into holistic and segmentation based approach. Holistic approach treats the word as a whole, whereas segmentation based approach is the divide and conquer method where each character is separately recognized.

Applications of offline HWR are

- Postal address identification.
- Writer's handwriting identification.
- Bank check recognition.
- Signature Verification in banks.
- Historical documents.
- Identifying the words in inscriptions.
- Palm leaf manuscript

Although several works have been taken place under the HWR still this field is a open problem for the research people. This survey focuses mainly on offline handwritten word recognition of various languages like English, Arabic, Hindi and Kannada. The remaining of the paper is organized as follows. In section 2, we discuss about the handwritten word recognition system. Section 3 deals with the survey on different languages. In the Section 4, conclusions are drawn.

I. HANDWRITTEN WORD RECOGNITION SYSTEM

A. Data acquisition

Normally handwritten words are collected from the persons of various ages, sex, education, and occupations. The A4 size paper sheet having the data written by various writers is digitized using the scanner. The images were stored in jpeg, gif, tiff or any other standard format.

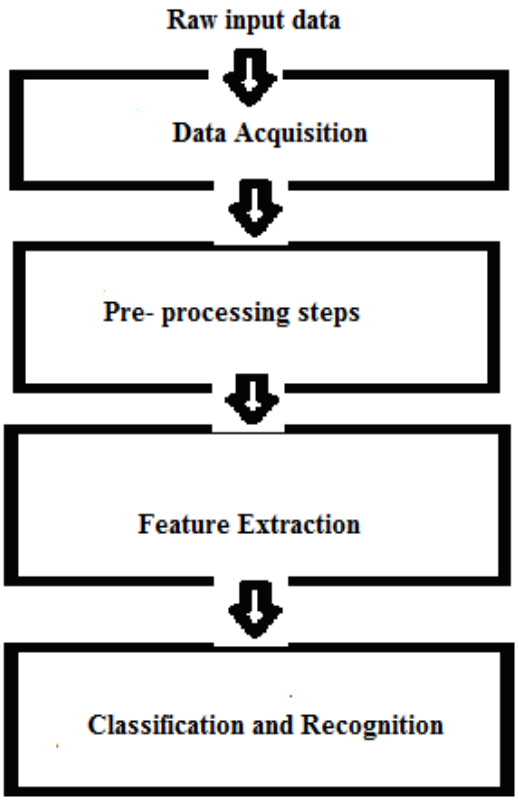

Fig. 1 Handwritten word recognition system

B. Preprocessing

In this step different operations are performed on the scanned images. Initially noise present in the image has to be removed. Commonly occurring noises in the document image are salt and pepper noise, Gaussian noise, ink blobs, stray marks, marginal noise, clutter noise [17]. Other tasks performed during this stage are thresholding, binarization, edge detection, skew detection and correction, size normalization, dilation and filling etc.

C. Feature extraction

This is the very important stage of handwritten word recognition system. The main objective of feature extraction is to extract all the essential features of the scanned image. The selection of the appropriate feature extraction method is the important factor in getting high recognition rate. The widely used feature extraction techniques are Zoning, Projection profiles, Hough transform, Chain code, Fourier descriptor etc. In this phase, we can extract features like structural, statistical or both. The feature extraction methods extract the features for classification and recognition of words. There are mainly three types of feature extraction techniques.

- Structural features
- Statistical features
- Hybrid features

D. Classification and recognition

Extracted features in the feature extraction steps are used to classify the images by assigning labels to these features. The data set collected is divided into training data set and testing data set. Bayesian classifier, Binary tree classifier, Nearest Neighbor classifier, Neural networks, Hidden Markov Model (HMM) and Support Vector Machines (SVM) are some of the classifiers that are used in this stage.

III. LITERATURE SURVEY

A. English

English is a West Germanic language originated in England. It is the official language of more than 60 countries. And it is the most commonly used language all over the world. English language has 26 letters. With 5 vowels and 21consonants.Upper case and lower case letters make total of 52 alphabet characters. Approximately 359 million people speak English as a first language.

B. Gatos et .al [1] proposed an off-line cursive on a novel combination of two different modes of word image normalization and robust hybrid feature extraction. Here two types of features are combined. The first feature which creates the set of zones by dividing the image and calculates the density of the character pixels in each zone. Second feature calculates the area that is formed from the projections of the upper and lower profile of the word. Two classifiers are used here. Namely Minimum Distance Classifier and the Support Vector Machines (SVM). 80.76% accuracy is achieved using IAM database.

Ankush Acharyya et .al [2] presented a holistic approach to recognize the offline handwritten words using Multi Layer Perceptron (MLP) classifier. Words are taken from the CMATERdb1.2.1 dataset. In this paper longest run features are used. These features are computed in four directions; row wise (east), column wise (north) and along the directions of two major diagonals (northeast and northwest). To get the more discriminating information of a particular word image, hierarchical partitioning is done till depth 5. Recognition rate achieved is 83%.

Rodolfo Luna-Pérez, Pilar Gómez-Gil [3] described an unconstrained handwritten word recognition using a combination of neural networks. Author presented a novel method for classification of isolated handwritten words based on three components: a Self Organizing Map (SOM) for non-supervised classification of segments of a word, a function measuring probabilities of each segment belonging to a specific cluster and a Simple Recurrent Network (SRN) for temporal classification of a sequence of feature vectors obtained from segments forming the word. A Feed-Forward (FF) network, FF-SOM network are the classifiers used. 86.5% accuracy is achieved. IAM benchmark database is used for testing.

Shaolei Feng et .al [4] reported Hidden Markov Model (HMM) for alphabet-soup word recognition. This approach first uses a joint boosting technique to detect potential

characters –called as alphabet soup. In the second stage dynamic programming algorithm to recover the correct sequence of characters is described. A Hidden Markov Model is used to recognize a sequence of characters of fixed length given the character detection results. Here 85% words are recognized correctly.

B. Arabic

Arabic is the native language of Arab countries. It has 290 million native speakers. And the official language of 27 countries. Arabic is the third most spoken language after English and Chinese. Arabic Abjad is the Arabic script used to write Arabic language. It is written from right to left. [5] The basic Arabic alphabet contains 28 letters.

Ahlam MAQQOR et .al [5] proposed a approach to cursive Arabic word recognition. The main objective of this system applies a multi-stream approach of two types of feature extraction methods. First one is based on local densities called as sliding window. and configurations of pixels and features a projection based on vertical, horizontal and diagonal 45 °, 135 ° - is the VH2D approach. By using multi-stream HMM 83.8% accuracy is obtained. Experiment is done on 200 Arabic words.

Ilya Zavorin [6] et.al described combining different classification approaches to Arabic handwritten word recognition. In this paper author spoke about the problem of offline Arabic handwriting recognition of pre-segmented words. Parts of Arabic Word (PAW) Segmenter, Ranking Lexicon Reducer, HMM Classifier are used here. Experiment is done on IFN/ENIT corpus of Tunisian village and town names.73% accuracy is achieved when combining the multiple classifiers.

Alex Graves and Jurgen Schmidhuber [7] presented a offline Arabic handwritten word recognition using multidimensional recurrent neural networks. Author combined two methods in neural networks. Multidimensional recurrent neural networks and connectionist temporal classification. Instead of using single recurrent connection multidimensional recurrent neural networks are used. Because of this 91.7% accuracy is obtained. IFN/ENIT database of handwritten Arabic words is used for the experiment.

Volker Märgner et .al [8] described a offline handwritten word recognition of Arabic words using HMM method. Grey valued pixels of the normalized word image are used as features in the feature extraction steps. Sliding window and Karhunen-Loève Transform (KLT) are applied. Sequence of transformed feature vectors are used as the input to the HMM classifier. IFN/ENIT - Database is used in the experiment and got 89.77% recognition rate is achieved.

C. Hindi

Hindi is the national language of India. It is written in Devanagari script also called Nagari. Hindi is the native language of 280 million people. There are 11 vowels, 2 modifiers and 36 consonants makes total of 49 letters in Hindi language.

Brijmohan Singh et .al [9] proposed a Curvelet Transform Based Approach to offline handwritten Devanagri word recognition. Principal Component Analysis (PCA) of the coefficients is used to reduce the size of feature vector to about 200 dimensions. For the recognition process Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) classifiers are used. K-NN produced better results than the SVM classifier and obtained 93.21% accuracy on Devanagari handwritten words.

Vaibhav Dedhe, Sandeep Patil [10] reported a handwritten Devanagri special characters and word recognition using neural networks. The neural classifier consists of two hidden layers besides an input layer and an output layer. To extract the information of the boundary of a handwritten character, the eight-neighbor adjacent method is used. Proposed method has provided accuracy up-to 90% for special characters of Devanagari script.

Naresh Kumar Garg et .al [11] described a offline handwritten Hindi text recognition using SVM method. The shape based features were extracted by applying many heuristics based on the shape of the character for each feature. Total 59 features are selected in the feature selection phase.89.6% recognition rate is achieved.

Shailendra Kumar Dewangan [12] reported real time recognition of handwritten Devanagari signatures using Artificial Neural Networks (ANN). Different features of signature such as height, slant, length etc are extracted and used for training of the Neural Network. Authors collected total 500 genuine signatures. The accuracy rate achieved by the proposed Devanagari handwritten signature recognition system was 96.12 %.

D. Kannada

Kannada is a language spoken in South Indian state of Karnataka, Kannada whose Native speakers are called Kannadigas and they are 50 million; it is one of the 30 most spoken languages in the world. Kannada is a southern Dravidian language. It has its own script derived from Bramhi script. Kannada language uses 49 phonemic letters; [16] it is divided into 3-groups, vowels (Swaragalu-13), consonants (Vyanjanagalu-34) and Anusvara and Visarga called as modifiers.

Thungamani.M et .al [13] proposed Kannada offline handwritten text recognition using Support Vector Machine (SVM) using Zernike moments. In the preprocessing step Skew estimation and correction, and Slope and slant correction has been done. Zernike Moments have been widely used as the invariant global features for word recognition. It is used as feature vector for recognizing images. The recognition rate achieved is 94 %.

B.V.Dhandra et .al [14] presented a Kannada writer's handwriting text recognition. A set of features based on Discrete Cosine Transform, Gabor filtering and gray level co-occurrence matrix, are used in the feature extraction stage. Experiment is done using the features of Discrete Cosine Transform (DCT), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM). Experimental results

showed that the Gabor energy features are more potential than the DCTs and GLCMs based features for writer identification. It has got a higher recognition rate of 88.5%. K-NN classifier is used in the recognition stage.

Keshava Prasanna et .al [15] reported a knowledge based information retrieval for syntactic analysis of Kannada script. Levenshtein edit distance technique is used as the word correction technique. The main data structure used in this work is the Ternary Search Tree (TST). MList is the other data structure used. An input word is taken from the user and it is searched for in a static data dictionary. The data dictionary is implemented using TST. Very good recognition rate is achieved in this work.

Krupashankari S Sandyal and M S Patel [16] proposed offline handwritten word recognition in Kannada language. For the future extraction Corner points, curves and loops have been used. Freeman's chain code (FCC) is the technique for the boundary extraction. Euclidean distance and DTW algorithm are the classifiers. Total of 1200 data samples are collected from the different age groups.92% accuracy rate is achieved during the experiments.

Table 1 Brief summary on literature survey

Authors	Language	Features used	Classifiers	Database	Accuracy
B. Gatos et .al	English	Zones and upper and lower profile of the word	Minimum Distance Classifier and the SVM	IAM	80.76%
Ankush Acharyya et .al	English	Longest run features	MLP classifier.	CMATERdb1.2.1	83%
Rodolfo Luna-Pérez, Pilar Gómez-Gil	English	Feature extractor based on non-supervised clustering, probability of belongs to k-most cluster	Feed-forward (FF) network, FF-SOM network	IAM	86.5%
Shaolei Feng et .al	English	Joint boosting technique	HMM	Collected by George Washington's secretaries	85%
Ahlam MAQQOR et .al	Arabic	Sliding window, VH2D approach	Multi-stream HMM	Own dataset	83.8%
Ilya Zavorin et.al	Arabic	Large-scale features	HMM	IFN/ENIT	73%
Alex Graves and Jurgen Schmidhuber	Arabic	The hierarchical structure	Multidimensional recurrent neural networks	IFN/ENIT	91.7%
Volker Märgner et .al	Arabic	Grey valued pixels of the normalized word image, sliding window	HMM	IFN/ENIT	89.77%
Brijmohan Singh et .al	Hindi	Curvelet based approach	SVM classifier	Own dataset	93.21%
Vaibhav Dedhe, Sandeep Patil	Hindi	Eight-neighbor adjacent method	Neural networks	Own dataset	90%
Naresh Kumar Garg et .al	Hindi	Bars, end points, loops, crossings, presence of particular horizontal and vertical lines, groves, curves and projection profiles	SVM classifier	Own dataset	89.6%
Shailendra Kumar Dewangan	Hindi	Hu's Moment Invariants	Artificial Neural Network	Own dataset	96.12 %.
Thungamani.M et .al	Kannada	Zernike Moments	SVM classifier	Own dataset	94%
B.V.Dhandra et .al	Kannada	Discrete Cosine Transform, Gabor filtering and gray level co-occurrence matrix	K-NN classifier	Own dataset	88.5%
Keshava Prasanna et .al	Kannada	Auto completion and spell checking.	Ternary search tree(TST)	Own dataset	-
Krupashankari S Sandyal and M S Patel	Kannada	Corner points, curves and loops	Euclidean distance and DTW algorithm	Own dataset	92%

IV. CONCLUSION

Offline handwritten word recognition is one of the interesting fields of research in the image processing. Though lot of work has been done still it has got many opportunities to do the work in this field. Reasons for this are, in the most of the languages lack of availability of standard datasets and accuracy rate obtained. By using good combinations of feature extraction methods and classifiers it is possible to achieve the good results. In our survey paper different methods used for preprocessing, feature extraction, classification are discussed. Survey has done on English, Arabic, Hindi, Kannada languages. We believe that this survey will helpful for researchers in this field.

V. REFERENCES

[1] B. Gatos, I. Pratikakis, A.L. Kesidis, S.J. Perantonis, "Efficient Off-Line Cursive Handwriting Word Recognition", Tenth International Workshop on Frontiers in Handwriting Recognition, Oct. 2006
[2] Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu, Mita Nasipuri, "Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013
[3] Rodolfo Luna-Perez, Pilar Gomez-Gil, "Unconstrained Handwritten Word Recognition Using a Combination of Neural Networks", ISBN: 978-988-17012-0-6, WCECS 2010
[4] Shaolei Feng Nicholas R. Howe R. Manmatha, "A Hidden Markov Model for Alphabet-Soup Word Recognition", Dept. of Computer Science, University of Massachusetts, Amherst, 2008

[5] 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, August 2013
[6] Ilya Zavorin, Eugene Borovikov, Ericson Davis, Anna Borovikov, Kristen Summers, "Combining Different Classification Approaches to Improve Off-line Arabic Handwritten Word Recognition", SPIE-IS&T/ Vol. 6815 681504-1, 2008
[7] Alex Graves, Jurgen Schmidhuber, "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks", In NIPS, PP. 545-552,2008
[8] Volker Margner, Haikal El Abed, Mario Pechwitz, "Offline Handwritten Arabic Word Recognition Using HMM -a Character Based Approach without Explicit Segmentation", SDN06, PP. 259-264, 2006
[9] Brijmohan Singh, Ankush Mittal, M.A. Ansari, " Handwritten Devanagari Word Recognition: A Curvelet Transform Based Approach", ISSN : 0975-3397 Vol. 3 No. 4 Apr 2011
[10] Vaibhav Dedhe, Sandeep Patil, "Handwritten Devnagari Special Characters and Words Recognition Using Neural Network", International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655, 2013
[11] Naresh Kumar Garg, Dr. Lakhwinder Kaur, Dr. Manish Jindal, "Recognition of Offline Handwritten Hindi Text Using SVM", International Journal of Image Processing (IJIP), Volume (7): Issue (4): 2013
[12] Shailendra Kumar Dewangan, "Real Time Recognition of Handwritten Devnagari Signatures without Segmentation Using Artificial Neural Network", I.J. Image, Graphics and Signal Processing, 2013
[13] Thungamani.M, Dr Ramakhanth Kumar P, Keshava Prasanna, Shravani Krishna Rau, "Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.7, July 2011
[14] B.V.Dhandra, Vijayalaxmi.M.B, Gururaj Mukarambi, Mallikarjun.Hangarge, "Writer Identification by Texture Analysis Based on Kannada Handwriting", International Journal of Communication Network Security ISSN: 2231 – 1882, Volume-1, Issue-4, 2012
[15] Keshava Prasanna, Dr Ramakhanth Kumar P, Thungamani.M, ShravaniKrishna Rau, " Knowledge Based Information Retrieval for Syntactic analysis of Kannada Script", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.7, July 2011
[16] Krupashankari.S.Sandyal, M.S.Patel, "Offline Handwritten Kannada Word Recognition", 07th IRF International Conference, ISBN: 978-93-84209-29-2, 2014
[17] M. Manomathi, S. Chitrakala, "Skew Angel Estimation and Correction of Noisy Document Images", ACC 2011, Part 3, CCIS 192, pp. 415-424, 2011