ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Offline Kannada Handwritten Word Recognition Using Locality Preserving Projection (LPP) for Feature Extraction

M.S. Patel[1], Rohith Kumar[2], S.C. Linga Reddy[3]

Dept. of ISE, Dayananda Sagar College of Engineering, Bengaluru, Affiliated VTU, India[1]

Dept. of ISE, Dayananda Sagar College of Engineering, Bengaluru, Affiliated VTU, India[2]

Dept. of CSE, Alpha College of Engineering, Bengaluru, Affiliated VTU, India[3]

ABSTRACT: Offline Handwritten Word Recognition (HWR) plays a major role in the field of image processing and pattern recognition. Compared to online recognition, handwritten words cannot be identified easily because of the variations in the handwriting styles, type of paper used, quality of the scanner etc. In our paper we have focused on the Kannada handwritten word recognition. Large number of characters present in the Kannada language makes it as a open problem for the researchers. Major steps in offline Kannada HWR are preprocessing, feature extraction, and classification. Locality Preserving Projections (LPP) method is used here for the feature extraction. For the classification Support Vector Machines (SVM) is used. Result is compared with the K-Means classifier. Experimental results show that SVM is better than K-Means classifier for our data set.

KEYWORDS: Offline Handwritten Word Recognition (HWR), preprocessing, feature extraction, classification, Locality Preserving Projections (LPP), Support Vector Machines (SVM), K-Means classifier.

I. INTRODUCTION

Offline Handwritten Word Recognition (HWR) is an important research filed in the area of image processing. When the generation is migrating towards the digitalized world, it's necessary to adopt the changes. It is not possible to save the historical documents, writer's books for many years in the original format. But once it is digitized, then it's very easy to use such documents for the generation to generations. HWR has got many real world applications, so it has become a potential leading research field in document image processing.

Recognizing the Kannada handwritten words is very complex in nature. This is because of the number of characters and shapes present in the Kannada language. The proposed method is having training and testing phase. In training phase images are preprocessed and features are extracted. In the testing phase, trained features of images are compared with the features of test image. Comparing the trained images with testing image is similar in features; then the word will be recognized.

We are mainly using the Locality Preserving Projections (LPP) for feature extraction. Support Vector Machines (SVM) is used for classification and recognition. Major steps in offline Handwritten Word Recognition (HWR) are data acquisition, preprocessing, feature extraction and classification. By using all these steps Kannada handwritten words are identified.

Applications of offline HWR are Postal address identification, writer's handwriting identification, bank cheque recognition, signature Verification in banks, historical documents, identifying the words in inscription, palm leaf manuscript.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Although several works have been taken place under the HWR still this field is a open problem for the research people. Here we are using Locality Preserving Projections (LPP) to extract the features and Support Vector Machines (SVM) for recognition. The remaining of the paper is organized as follows. In section 2, we discuss about Kannada language. Literature survey is discussed in section 3. Section 4 deals with the proposed method. Experimental results are shown in section 5. In the Section 6 conclusions are drawn.

II. KANNADA LANGUAGE

A. Kannada

Kannada is the official language of Karnataka state. More than 30 million people speak Kannada as the first language. Around 11 million people use Kannada as the second language. It is the 27th most spoken language in the world. Kannada has its own script derived from Bramhi script. Kannada script has a set of 49 characters. They are classified into three categories: Swara (vowels), vyanjana (consonants), and yogavahakas. There are 13 vowels, 34 consonants and 2 yogavahakas. Modifier glyphs (Half-letters) from the vowels and yogavahakas are used to alter the 34 base consonants. Additionally, a consonant emphasis glyph called vattakshara (subscript) exists for each of the 34 consonants. This gives a total of (544*34) + 15=18511 distinct characters in Kannada language [20]. Hence identifying the Kannada words is complex task. Following figures shows the vowels and consonants of Kannada language respectively.

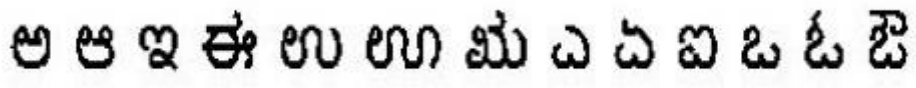

Fig 1: Vowels of Kannada script

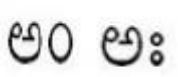

Fig 2: Yogavahakas of Kannada script

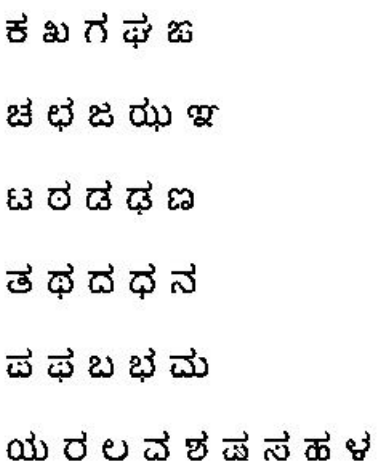

Fig 3: Consonants of Kannada script

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

B. Motivation

Offline handwritten word recognition is one of the most difficult task compared to online recognition approach. This is because of the difference in writing style from one person to another person, thickness of the pen, environment, depends on the situation of the writer etc. Many researches are going on in this particular field for efficient way to recognize the handwritten words. In olden days people used to write the documents, books in papers. But it's not lasting for many years in the original format.

Because of the improvement in the technology in the past few decades, the olden historical documents are stored in the form of digitalization. Hence it can be helpful for the feature generation for extracting, modifying and storing. Offline HWR is the better approach to achieve this goal.

III. LITERATURE SURVEY

Rodolfo Luna-Pérez, Pilar Gómez-Gil [3] described the combination of neural network method for identifying the handwritten words. Here they used three components: a Self Organizing Map (SOM) for non-supervised classes, a function which measures the probability of each segment and a Simple Recurrent Network (SRN) for classification, 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 method 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 for recognition. In this paper 85% of the recognition rate has been achieved.

Alex Graves and Jurgen Schmidhuber [7] presented an 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.

Naresh Kumar Garg et .al [11] described a offline handwritten Hindi text recognition using SVM method. The shape based features were extracted. 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 are collected total 500 genuine signatures. The accuracy rate achieved by the proposed Devanagari handwritten signature recognition system was 96.12 %.

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.

M.S. Patel et al [23] proposed a grid based approach to offline Kannada handwritten word recognition. Principal Component Analysis (PCA) is the best dimension reduction algorithm in the subspace learning approaches. Initially given image is divided into 4 sub parts. Then finding the average image of all the sub images. Normalize the sub

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

images by subtracting it from the mean. In the next step Eigen vector is determined. Using this Eigen vector feature vectors are computed. Results shows that proposed grid based approach is better than standard PCA method.

IV. PROPOSED METHOD

There are 5 stages in the proposed offline Kannada handwritten word recognition system

A. Data acquisition
B. Pre-processing
C. Feature extraction
D. Classification
E. Post processing

A. Data acquisition

Collecting the handwritten words is a challenging task. Based on this data acquisition, the experiments are conducted. Here we have collected our own dataset. Names of 30 districts and 174 taluks of Karnataka is written by 50 people. Then the collected handwritten words are scanned using scanner. This scanned document has been cropped and resized. Image is stored in standard format like tiff.

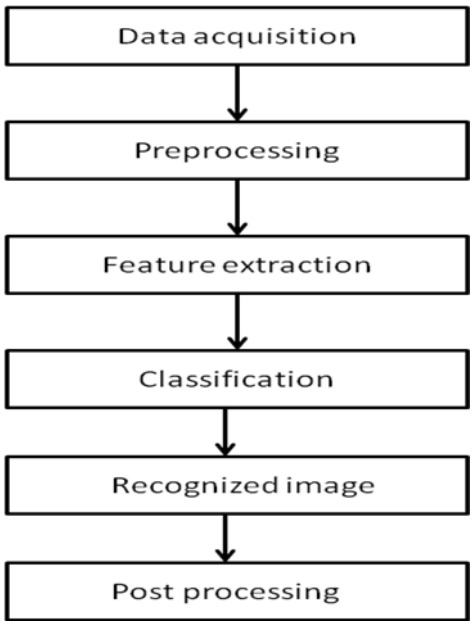

Fig 4: Proposed offline Kannada handwritten word recognition system.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

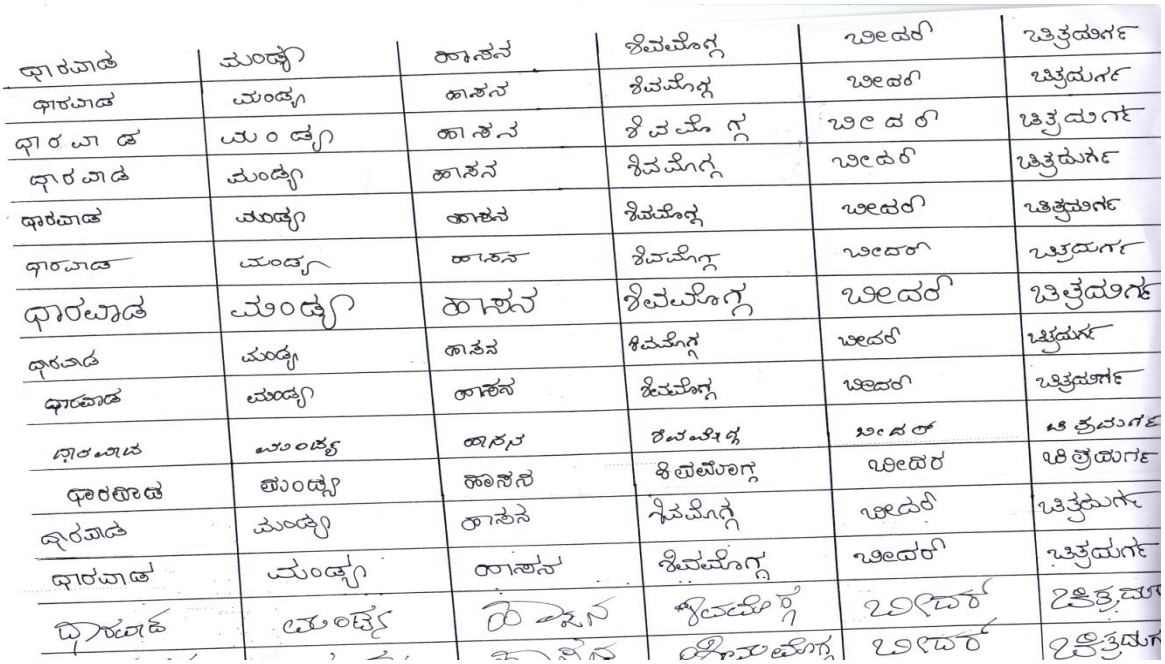

Fig 5: Scanned handwritten words

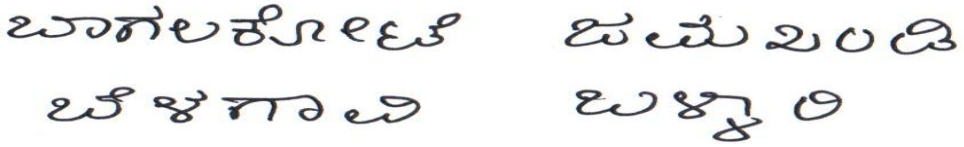

Fig 6: Cropped images

B. Preprocessing

The first step in the offline Kannada handwritten word recognition is preprocessing. Main aim of preprocessing is to enhance the image, so the recognition rate is improved. Following figure shows the steps involved in preprocessing.

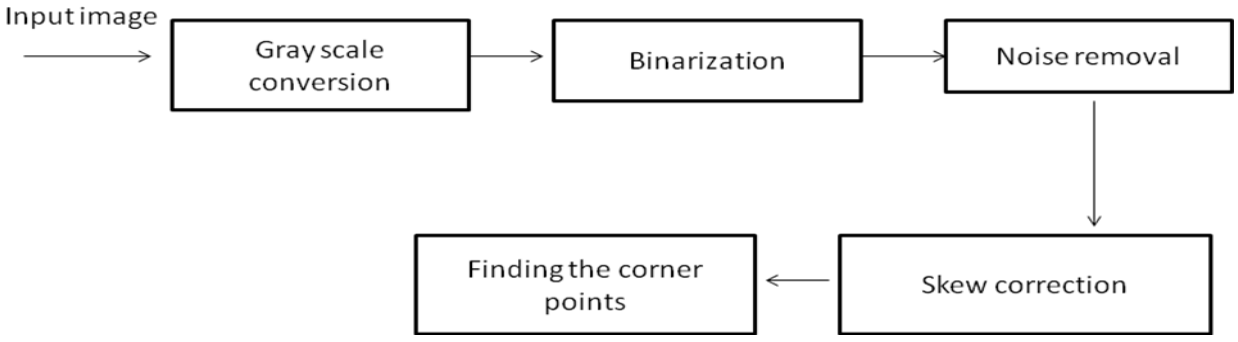

Fig 7: Preprocessing steps

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Gray scale conversion

Input image is converted into binary form. Supporting to this, image has to undergo gray scale conversion which includes gray shades in the middle of white and black pixels.

Binarization

Binarization is the process of converting gray scale image into binary form. It is useful for identifying the object of interest in the image. Otsu's approach is used for Binarization where based on the threshold value pixels are represented as white or black pixels.

Noise removal

Due to the low quality of scanner and degraded document, image might subject to noise which affects the recognition rate. Hence it is very important to remove the noise before the image is fed into next steps. Salt and pepper noise, Gaussian noise are the most common noise present in the document image. These noises can be removed by using median filter, linear filter and adaptive filters.

Skew correction

During document scanning skew is introduced in the image. Skew represents the angle in which the given image is tilted in the horizontal direction. Aim of the skew correction is to detect the skew angle and correct it. Hence it will be useful for the later stages.

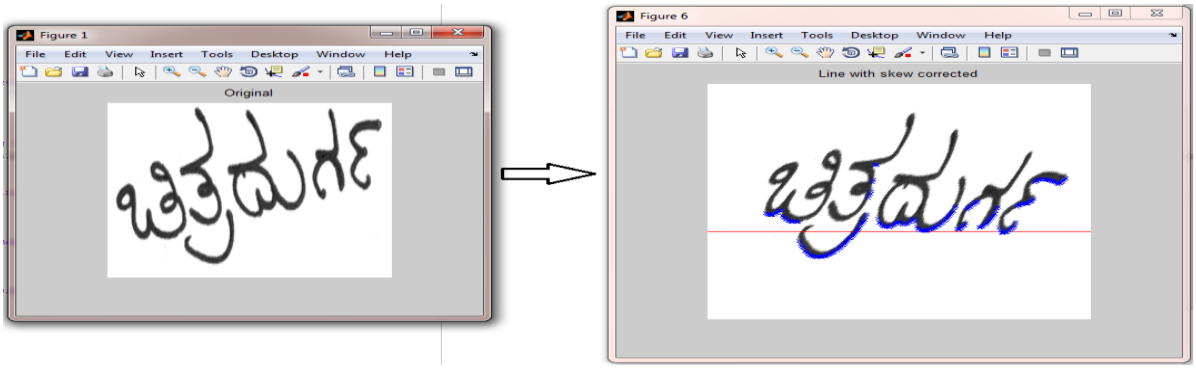

Fig 8: Input image after skew correction

Corner point detection

It is very important to differentiate the background pixels and foreground pixels. It is done using corner point detection. This helps to preserve the important structural properties of the image and eliminates useless information.

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. Here we have used Locality Preserving Projections (LPP) for the feature extraction. It is mainly used for the dimensionality reduction. LPP overcomes the disadvantages of Principal Component Analysis (PCA). LPP mainly focuses on the local structure of the image. But PCA focuses on the global structure.

Locality Preserving Projections (LPP) is a linear dimensionality reduction algorithm. [18] It is mainly used in face recognition and speech recognition. Here we are using this LPP to extract the features. Given the set of data points and local similarity matrix, we need to find the optimal projections by solving the minimization problem. Here we are comparing each data point with its neighborhood whereas in the PCA data points are compared with the average data

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

point values in the image. Obtained data features are stored in the database and compared with the features of testing image during the classification step.

D. *Cl*assification and recognition

Classification and recognition playa a major role in document image processing. During the training phase images are grouped into separate classes and each image is labeled. In the testing phase, features of test image are compared with the trained images. Matching word images are displayed later. this is done using classifiers. Here we have used Support Vector Machines (SVM) and the result is compared with the K-Means classifier.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is the most commonly used classifier for the offline word recognition. Here during the training phase feature vectors will be divided into 2 classes. In the testing phase unsupervised image will be classified with the help of hyper planes and support vectors.

K-Means classifier

K-means classifier creates the cluster of classes [21]. Recognition is done using 2 steps. In the first step trained images will be grouped into clusters. During the testing step distance between feature vector of test image and centroids of each cluster will be calculated. [22] Shorter distance will be picked up from the calculation and that clusters class label will be assigned to the test image.

E. Post processing

Post processing is the last step in proposed offline Kannada handwritten word recognition system. Message box with respect to the given test image is displaying after the classification and recognition. Referring to that, images of important places in the particular district is displayed. Sample result is shown in the figure 9.

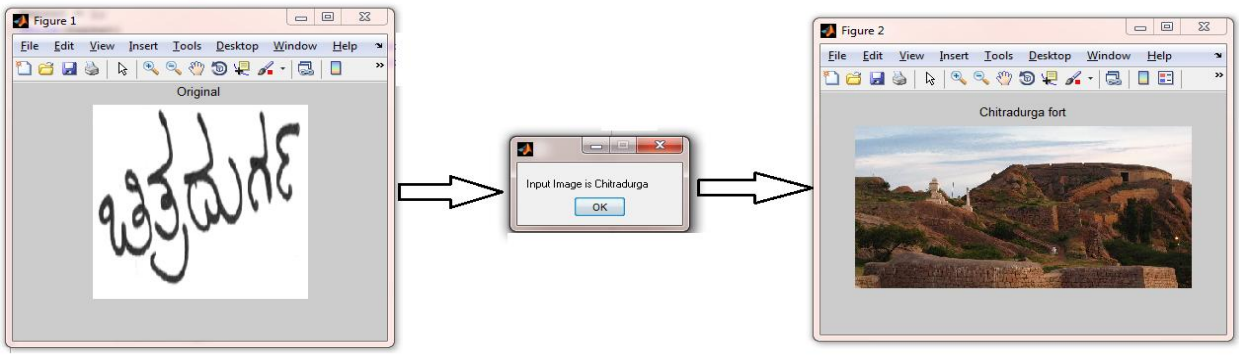

Fig 9: Output representation

V. EXPERIMENTAL RESULTS

This section evaluates the offline handwritten word recognition method by experimentally with a dataset containing handwritten words of 30 district and 174 taluk names of Karnataka state. The dataset is prepared with different people with geographical area of the Karnataka state. At the first stage the cropped word is resized to 320 x 200 size. For the experimental purpose randomly choose the names of the district and taluk in the prepared dataset. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

experiment is conducted by varying number of testing images with the interval of 10 like, 10, 20,30,40,50,60,70,80. Here two ways of conducting experiment taking different classifiers namely SVM and K-means algorithms respectively. In the training phase images are converted into vector format and features are extracted and stored in the database. During the testing phase test image is compared with the features of trained images database. The features are matched with trained image; the image is recognized and displayed. In the final step (post processing) the output recognized image is displayed, Instead of the recognized image display and also, we are following to display like, important place, renowned person, Historical place of the particular district. This is for visualization purpose.

Table 1: Experimental results

No. of test images	Recognized images (SVM)	Recognition rate (SVM)	Recognized images (K- Means classifier)	Recognition rate (K-Means)	Average recognition rate
20	18	90%	18	90%	SVM 85%
40	35	87.5%	34	85%	
60	51	85%	50	83.33%	K-Means 83%
80	66	82.5%	64	80%	

VI. CONCLUSION

Offline handwritten word recognition is the most challenging task in the field of image processing. Kannada language having its cursive nature and large number of alphabets. Lack of availability of standard data set challenges the researchers to do the research in Kannada language. We have used Locality Preserving Projections (LPP) to reduce the dimensionality and to extract the features .Support Vector Machines (SVM) and K-means is used to recognize the word images based on the extracted features. Experimental result shows the good accuracy rate towards Kannada handwritten words. In future, we explore different variants of subspace learning methods for better representation task.

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

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

[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
[18] Xiaofei He, Partha Niyogi, "Locality Preserving Projections (LPP)" Computer Science Department The University of Chicago Chicago, IL 60615 Chicago
[19] Vikramaditya Jakkula, "Tutorial on Support Vector Machine (SVM)", Schoo, of EECS, Washington State University, Pullman 99164
[20] Keshava Prasanna, P. Ramakanth Kumar "Handwriting Recognition of Kannada Characters and Context Free Grammar Based Syntax Analysis", International Journal of Science Research Volume 01, Issue 01, June 2012, pp. 24-29
[21] Shiv Ram Dubey, Pushkar Dixit, Nishant Singh, Jay Prakash Gupta, "Infected Fruit Part Detection using K-Means Clustering Segmentation Technique", International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, No.2.
[22] Sabna Sharma, Ratika Pradhan, "Classification Methods for Land use and Land Cover Pattern Analysis", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-4, Issue-1, June 2014