ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (936-942), Month: April - June 2015, Available at: www.researchpublish.com

Offline Kannada Handwritten Word Recognition Using Support Vector Machines (SVM)

[1]Rohith Kumar, [2]M S Patel

[1]M. Tech Student, [2]Research Scholar, VTU Belgaum, Department of Information Science and Engineering DSCE, Bangalore

***Abstract*: Offline Handwritten word Recognition (HWR) achieved increasing attention in the past few decades. Recognizing Indian language of Kannada is very difficult task not only because of variation in the handwriting, also because of number of alphabets present in Kannada languages, number of shapes, etc. To achieve the good recognition rate it is important to have good combination of feature extraction techniques and classifier. In this paper mainly we have used Gray Level Co-occurrence Matrix (GLCM) to extract the features and Support Vector Machines (SVM) for recognition. Experimental results show the good accuracy rate.**

***Keywords:* Offline Handwritten Word Recognition (HWR), feature extraction, classifier, Gray Level Co-occurrence Matrix (GLCM), Support Vector Machines (SVM).**

I. INTRODUCTION

In the era of digital world everyone wants to finish their work as early as possible. Supporting to that now a days so many innovations, researches are taking place. As we know in the earlier days people were writing the books, or any other kind of information by their own handwritings. That too on palm leaves, or papers. But it's not lasting for many years. Hence there was a need for new technology which can save the information for next generations also. Image processing technique helps to achieve this .Many researches are taking place in this field to recognize the handwritten words.

Image processing is the technique which modifies or manipulates the digitized image in order to enhance its quality. Different types of image processing are exist. Some of them are satellite image processing, medical image processing, document image processing, etc. In satellite image processing, images are captured by the satellites and processing it. In medical image processing medical images are analyzed to find the cancer tissues or other types of diseases. Document image processing is the process of analyzing the documents and the preparation of secondary information that can be helpful for the future use. There are mainly 2 approaches in the document image processing. Online and offline approach.

In Online HWR the route 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 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 cheque recognition.
- Signature Verification in banks.
- Historical documents.
- Identifying the words in inscriptions.
- Palm leaf manuscript

Several works have been taken place under the HWR. Still this field is a open problem for the research people. In the proposed method Gray Level Co-occurrence Matrix (GLCM) is used 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 south Indian state of Karnataka. More than 30 million people talk Kannada as the first language. Around 11 million people use Kannada as the second language. Kannada has got its own script derived from Brahmi script. Kannada has a base 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 [11]. In Kannada there exist more than 250 basic shapes. Hence it's not so easy to recognize the large collection of characters with similar shapes.

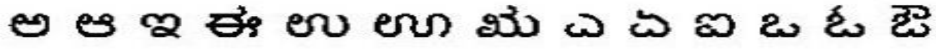


Fig 1: Vowels of Kannada script

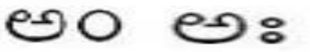


Fig 2: Yogavahakas of Kannada script

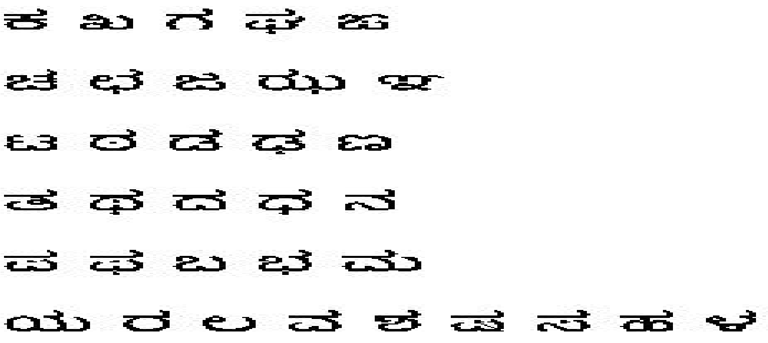


Fig 3: Consonants of Kannada script.

B. *Motivation:*

Nature of handwriting differs from person to person. Cursive nature of Kannada language makes it more difficult to identify the handwritten words. Number of alphabets, shapes of the letters plays a major role in achieving the good results. Compared to English like languages less work has been taken place in Kannada language. Some of the reasons are lack of availability of standard dataset and recognition rate obtained.

III. LITERATURE SURVEY

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%.

Ahlam MAQQOR et .al [3] 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 [4] 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.

Brijmohan Singh et .al [5] 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 [6] 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.

Thungamani.M et .al [7] 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 [8] 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.

IV. PROPOSED METHOD

There are 4 Stages in the offline Kannada handwritten word recognition system

- Data acquisition
- Pre-processing
- Feature extraction

- Classification

Initially database of words has to be prepared. In the next step by applying preprocessing step images are enhanced. Then features present in that image are extracted and stored in data file. In the testing phase features of trained images are compared with the features of test image. If both are matching then the word is recognized and displayed.

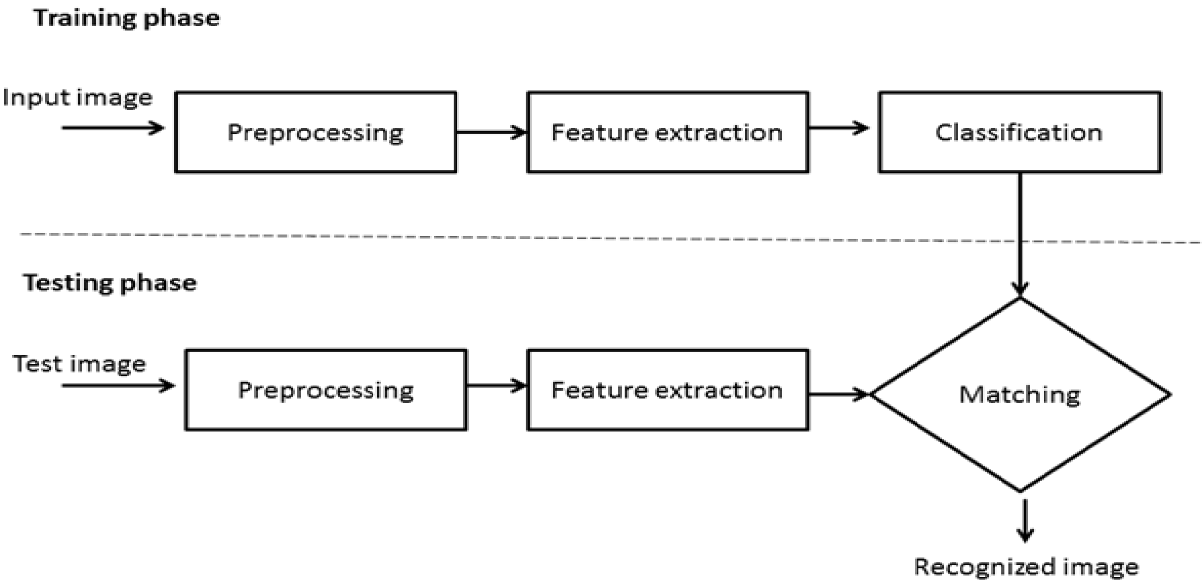
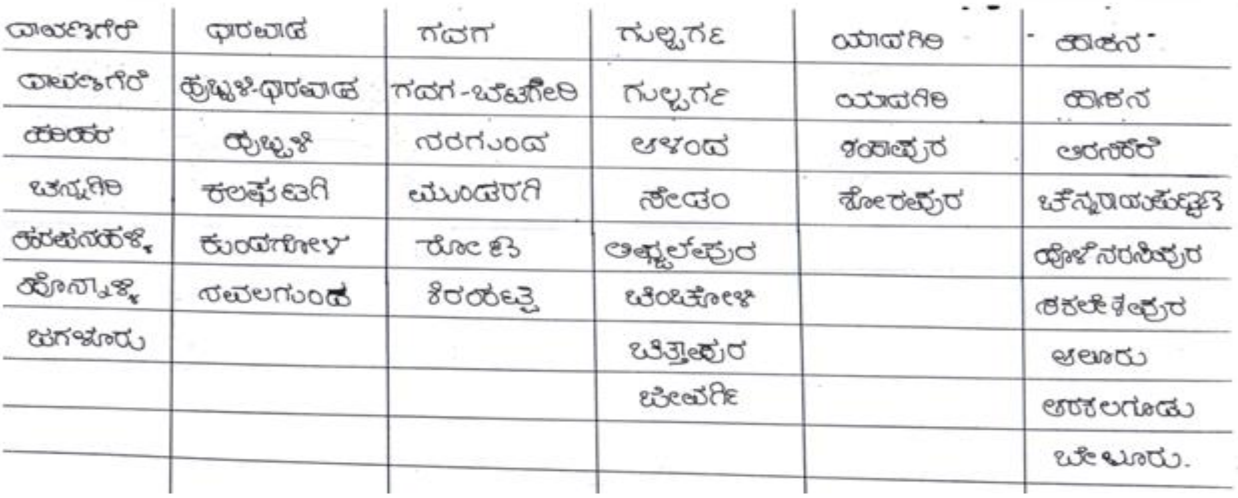


Fig 4: Proposed system

A. Data acquisition:

Data acquisition is the first step in the offline handwritten word recognition process. We have collected handwritten words from the persons of various ages, sex, education, and occupations. Names of 30 districts and 174 taluks of Karnataka from 50 people were obtained and stored as data set. The A4 size paper sheet having the data written by various writers is digitized using the scanner. The images are stored in tiff format and resized to 320*200 resolutions.



ದಾವಣಗೆರೆ	ಧಾರವಾಡ	ಗದಗ	ಗುಲ್ಬರ್ಗ	ಯಾದಗಿರಿ	ಹಾಸನ
ದಾವಣಗೆರೆ	ಹುಬ್ಬಳ್ಳಿ-ಧಾರವಾಡ	ಗದಗ-ಬೆಟಗೇರಿ	ಗುಲ್ಬರ್ಗ	ಯಾದಗಿರಿ	ಹಾಸನ
ಹರಿಹರ	ಹುಬ್ಬಳ್ಳಿ	ನರಗುಂದ	ಆಳಂದ	ಶಹಾಪುರ	ಅರಸೀಕೆರೆ
ಚನ್ನಗಿರಿ	ಕಲಘಟಗಿ	ಮುಂಡರಗಿ	ಸೇಡಂ	ಶೋರಾಪುರ	ಚನ್ನರಾಯಪಟ್ಟಣ
ಹರಪನಹಳ್ಳಿ	ಕುಂದಗೋಳ	ರೋಣ	ಅಫ್ಜಲ್‌ಪುರ		ಹೊಳೆನರಸೀಪುರ
ಹೊನ್ನಾಳಿ	ನವಲಗುಂದ	ಶಿರಹಟ್ಟಿ	ಚಿಂಚೋಳಿ		ಸಕಲೇಶಪುರ
ಜಗಳೂರು			ಚಿತ್ತಾಪುರ		ಆಲೂರು
			ಜೇವರ್ಗಿ		ಅರಕಲಗೂಡು
					ಬೇಲೂರು.

Fig 5: Sample dataset

B. Preprocessing:

In preprocessing step image is enhanced, Hence it’s very easy to extract the features and for the recognition. Initially the image is converted into gray scale. In the next step Binarization is performed. Now its required to remove the noise present in the image. Most commonly occurring noise in the scanned document is salt and pepper noise. [9] It has been removed using median filter.

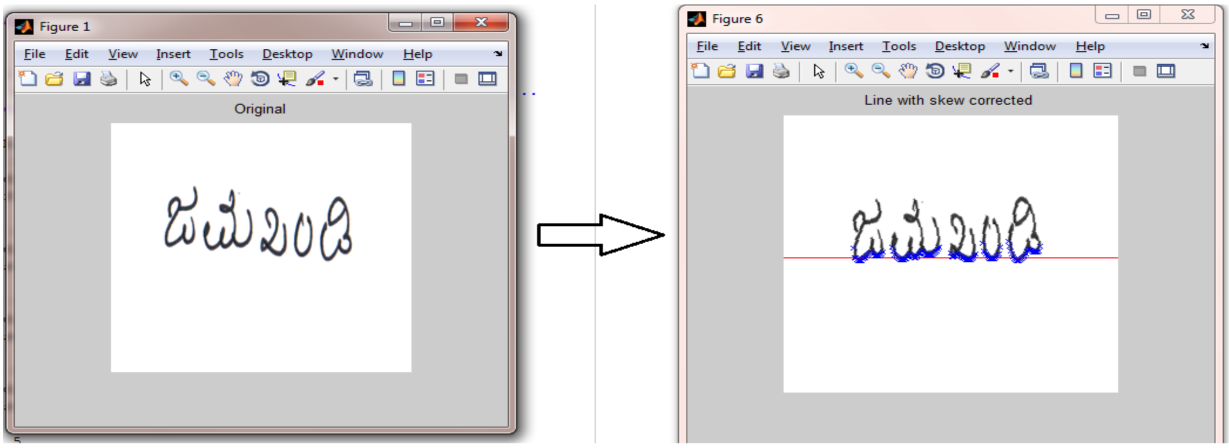


Figure 1
File Edit View Insert Tools Desktop Window Help
Original
Figure 6
File Edit View Insert Tools Desktop Window Help
Line with skew corrected

Fig 6: input image after skew correction with base line.

While scanning the document image there is a chance of tilting. This is what we call it as skew. To increase the recognition rate its mandatory to detect this skew and correct it. First we need to check how much angle it has been rotated. Once the skew angle is detected it has to be corrected as shown in the Fig 6.

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. Good combination of feature extraction method and classifier gives the high accuracy rate. We used mainly 3 feature extraction methods.

- Gray Level Co-occurrence Matrix (GLCM)
- Shape based features
- Graph based feature extraction method

Gray Level Co-occurrence Matrix (GLCM):

Gray Level Co-occurrence Matrix (GLCM) is mainly used for image texture analysis. [12] Here we are calculating the distance and angle between the sub regions of the image. Initially the image has to be converted into gray scale. GLCM is calculated using this gray scale image, based on how often the co occurring values are present in the image.

Shape based features:

Area, perimeter, form factor, major axis, minor axis, roundness, compactness, density – these are the features, calculated based on the shape of the image.

Graph based feature extraction method:

In graph based feature extraction method corners of the foreground image are recognized. Then Calculating the distance between each corner points using Euclidian distance algorithm.

D. *Classification:*

Extracted features in the feature extraction steps are used to classify the images by assigning labels to these features. For the classification and recognition we are using Support Vector Machine (SVM). It is the most commonly used classifier which has the ability to identify the patterns for which it has not been trained. [10]

Steps to be followed during classification using SVM

Training phase:

Step 1: Training images (X,Y) are represented in the vector format.

Step 2: Each vector has to be labeled and stored in a class.

Step 3: Create two classes to differentiate the patterns based on the equation

$$f(x) = W^{T}x+b$$

Where W is the normal vector, b is the bias value.

Testing phase:

In testing phase we need to check for which class testing image belongs.

Step 1: Initially create a hyper plane to distinguish the two classes based on the equation

$$wx+b= 0$$

Step 2: To support this hyper plane, create two more hyper planes using

$$wx+b= 1 \text{ and } wx+b= -1$$

Vectors which intersect at these hyper planes are called as support vectors.

Step 3: Using these hyper planes find the distance to the testing image and recognize under which class it will come.

V. EXPERIMENTAL RESULTS

We have collected data samples from various persons of different age, sex, education and stored in tiff format. For the recognition purpose we have used Support Vector Machine (SVM) classifier. Following table represents the experimental results. Then the result is compared with the K-Means classifier.

Following Table 1 shows the experimental results. Using SVM classifier 88.96% accuracy rate is achieved. 82.85% recognition is obtained using K-Means classifier. Graphical representation of the experimental results is shown in Fig 7.

Table 1: Experimental results

No. of test images	20	40	60	80	100	**Average recognition rate**
Recognized images (SVM)	19	36	53	70	84	88.96%
Recognition rate	95	90	88.33	87.5	84	
Recognized images (K- Means classifier)	17	33	51	67	78	82.85%
Recognition rate	85	82.5	85	83.75	78	

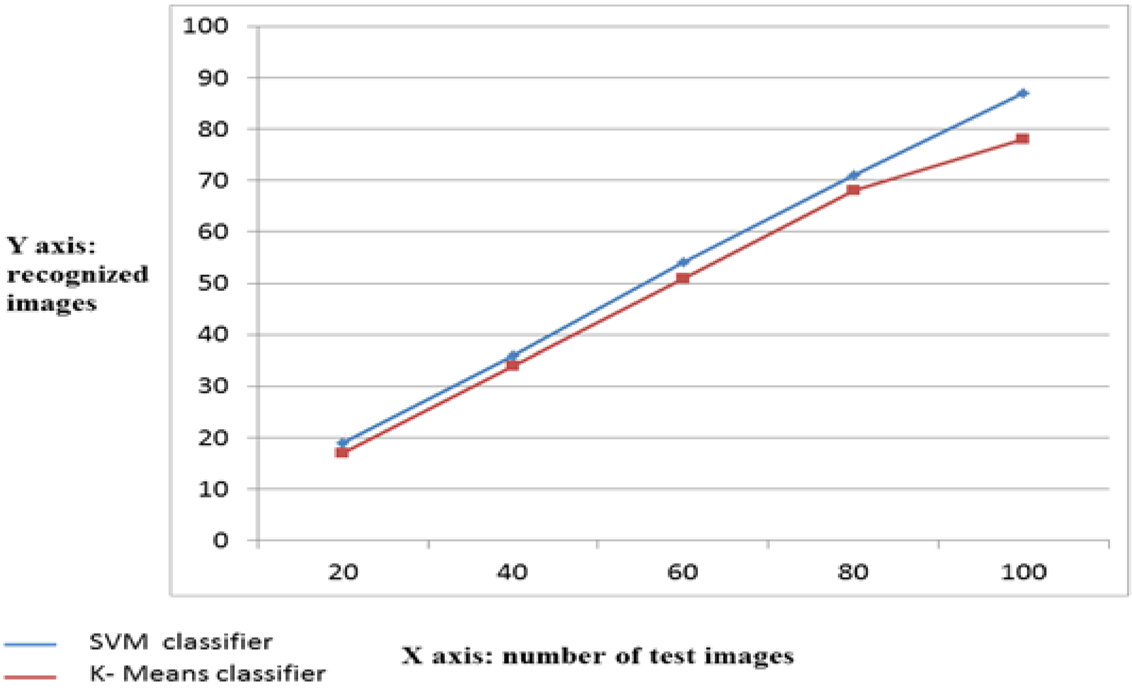


Fig 7: Comparison of SVM and K-means classifier

VI. CONCLUSION

Offline handwritten word recognition is one of the most difficult task in the field of document image processing. Recognizing the Kannada words is challenging one compared to English like languages. This is not only because of the different handwriting style, also because of the variation in the shape of the character, number of characters present in Kannada language, quality of the paper used, etc. In this paper we have GLCM, shape based features and graph based feature extraction techniques to extract the features. Support Vector Machine (SVM) classifier is used for the recognition of handwritten words and the result is compared with the K- Means classifier. From experimental results, SVM classifier shows better accuracy than K-Means classifier for our dataset.

REFERENCES

[1] 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] M. Manomathi, S. Chitrakala, “Skew Angel Estimation and Correction of Noisy Document Images”, ACC 2011, Part 3, CCIS 192, pp. 415-424, 2011.

[10] Vikramaditya Jakkula, “Tutorial on Support Vector Machine (SVM)”, Schoo, of EECS, Washington State University, Pullman 99164.

[11] 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.

[12] R.M. Haralick, K. Shanmungam, and I. Dinstein, “Textural Features of Image Classification,” IEEE, vol. 3, pp. 610-621, 1973.