Feature Extraction for Marathi Compound Character using KNN Classifier

[(1)]S. D. Bhosale [(2)] Dr. U. B. Shinde

[(1)] Research Scholar, National Institute of Electronics & Information Technology, Aurangabad, India.
[(2)]Department of Electronics and Telecommunication Engineering, CSMSS. Chh.. Shahu College of Engineering, Aurangabad, India.

[(1)] sdbhosale1968@gmail.com [(2)]drshindeulhas@gmail.com

Abstract: **Compound character recognition of Handwritten Devanagari is one of the challenging tasks due to its complexity as compare to many other scripts as the language is challenging for recognition. Compound characters itself complex in structure in Marathi compared to other languages. It is written with combination of two or more characters in alphabets. The character may be formed with different sequence of combinations of basic characters, such as vowels and consonants for Marathi alphabets. The recognition of compound characters makes this task more challenging for recognition. The frequency of occurrence of compound characters in Marathi language is more as compared to other languages derived from devanagari script as more compound letters are included in Marathi. The various researchers used different classification techniques such as Neural Network, Soft Computing, Seventh Central Moment, SVM Classifier, wavelet transformation etc and derived various results.**

The paper presents a novel approach for recognition of unconstrained Marathi compound characters. The recognition is carried out using multistage feature extraction and classification scheme. This character tends to touch each other in different forms and so the segmentation of compound characters is difficult task because of high error rate and higher complexity.

This paper, shows a similar investigation of Devanagari character recognition utilizing various feature techniques. In this paper we are using OCR method for recognition of compound text and then converting the recognized text.

Keyword: **Digital image processing, Marathi compound character, OCR, Segmentation, classification**

I. INTRODUCTION

Data processing by computer has many advantages in the viewpoints of cost, convenience. Now days people are trying to process data by computer in many fields. To realize computer data processing, it is essential to input all data to process into computer for safety. But there is a bottleneck on the input process because it requires much time, cost, and labor efforts. As a solution of this problem, new automatic character recognition is very important technology. This is the way the character recognition research is performed in many universities, companies, and research centers. There was many advances in this field that printed character recognition technology is being commercialized now. But recognition technology for handwritten character needs much improvement for commercialization.. There are several mechanism proposed in the direction of Marathi character recognition which includes design of feature vectors, classifier, cascading of the classifiers, new preprocessing techniques and so on various techniques. Off late the trends has shown that support vector machines produces best results in the direction of character recognition.

The recognition systems developed so far were for simple characters comprising of consonants and vowels in alphabets. But there is one more category of characters called as compound characters in Marathi script which is very difficult for recognition. These characters are formed by joining two or more consonant. The recognition of compound characters is still more difficult task due to their features in Marathi script. In this paper, we propose system for unconstrained Marathi compound character recognition without separation of the characters in the compound character. Very less work for Marathi compound character recognition is found so far to the best of my knowledge. This paper discusses the design of various models, simple and hybrid for the recognition of handwritten Marathi compound characters.

The compound character can have two or more characters joined together in various methods. One way of forming compound character is by removing the vertical line of a character and then joining it to the other on its left hand side. This type of joining is more common in Marathi. Second way of connection of characters in the compound character is by just joining the characters side by side or one above the other alphabet. In Marathi more than two consonants also join in various ways to form a compound character. In another way, one of the characters completely changes its form and then gets attached with other to form a compound character. The compound characters not only exhibit a variation in the shape of the character but also in the aspect ratio as per the joining strategy of two alphabets. One might get tempted to use the features like aspect ratio or number of end points, but the various joining strategies limits the use of these features to achieve acceptable recognition accuracy in compound words. All these challenges cannot be met by just a single feature extractor or using a single classifier. A feature extractor and classifier combination may recognize a character which may not be recognized by the other feature extractor and classifier combination. So a multistage system is needed that can recognize the characters over a wide range of varying conditions.

It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting line by line and words by word. Segmentation is to separate line, word and character from the image of Marathi script. There are two types of contextual segmentation depending on signal discontinuity and signal similarity of alphabets. Cluster, compression based methods, histograms, edge detection are widely used in contextual segmentation for Marathi.

II. Proposed Model for Marathi Compound Character Recognition using KNN Classifier:

Algorithm and description

1. Image Acquisition

Image acquisition in image processing can be broadly defined as the action of obtaining an image from some source, usually a hardware-based source. It can be passed through whatever processes need to occur afterward. Performing image acquisition in image processing is always the first step in the workflow sequence.

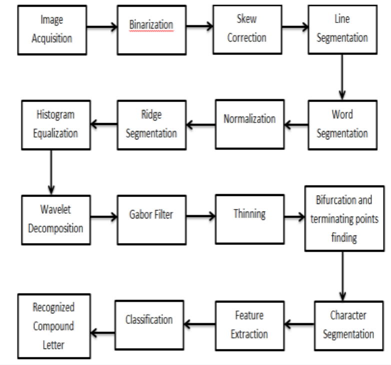


Fig1: Block Diagram of proposed method

The Input Image that we get is completely unprocessed and same as original one. One of the ultimate goals of this process is to have a source of input that operates within such controlled and measured guidelines that the same image can reproduced under the same conditions so anomalous factors is easy to locate and eliminate.

2. Binarization

Binarization means digitization of image. binaries an Image based on the threshold value chosen. Thresholding is an image processing technique for converting a color or gray scale image to a binary image based upon a threshold value chosen. If a pixel in the image has an intensity value less than the threshold value, set the corresponding pixel in the resultant image to black. Other-wise, if the pixel intensity value is greater than or equal to the threshold intensity, the resulting pixel is set to white. So used an image with only 2 colors, black (0) and white (255).

Most document analysis algorithms are built on taking advantage of the binarized image data. The use of information decreases the computational load and enables the utilization of the simplified analysis methods compared to 256 levels of grey-scale or colour image information. Document image understanding methods require logical and semantic content preservation during thresholding process. For example letter connectivity should be maintained for optical character recognition and textual compression. This requirement narrows down the use of a global threshold in many cases. Binarization has been a subject of intense research interest during the last few years. Most of the developed algorithms rely on statistical methods and not considering the special nature of document images. However, recent developments on document types, for example documents with mixed text and graph.

Many binarization techniques used in processing tasks are aimed at simplifying and unifying the image data at hand. The simplification is performed to benefit the oncoming processing characteristics, such as computational load, algorithm complexity and real-time requirements in industry. One of the key reasons when the binarization step fails to provide the subsequent processing a high-quality data is caused by the different types and degrees of degradation introduced to the acquired source image. The reasons for the degradation may vary from poor source type, the image acquisition process to the environment that causes problems for the image quality.

Since the degradation is one of the main reasons for processing to fail, it is very important to design the binarization technique to detect and possible imperfections from becoming the subject for processing and potential cause of errors for post-processing steps. Most degradation types in document images affect both physical and semantic understandability in the document analysis tasks, such as page segmentation, classification, optical character recognition. Therefore, the result after all the desired processing steps can be entirely unacceptable, because of the poorly performed binarization.

3. Skew Correction

When image is captured by camera or scanned by scanner a few degrees skew is unavoidable. Skew angle is the angle that text lines in digital image makes with the horizontal direction. Skew estimation and correction are important pre-processing steps of document layout analysis and OCR approaches in this. The horizontal and vertical projection profile is histogram of the number of black pixels along horizontal and vertical scan lines. For a script with horizontal text lines the horizontal projection profile will have peaks at text line positions and thorough at positions in between successive text lines of data. To determine the skew in document, the projection profile is used at a number of angles and for every angle, a measure of difference of peak and through height is made for that. The maximum amount of difference corresponds to the best alignment with text line direction which, in turn, determines the skew angle more correctly.

4. *Line Segmentation*

Segmentation process is segmenting the text documents into lines, also called as line segmentation. Header lines are those with maximum number of black pixels and base lines are rows with minimum number of black pixels in the image. Finding header line is a challenge task because of skew in header line. Now a day's most of researchers are detecting the header line by finding the row with maximum pixel density, but it can't work for skew variable text in image. This method gives good results for uniform and non-uniform skewed lines easily.

Determining Location of Text Line

1) For each scan line the proposed system will check all pixels on the scan line.
2) If for particular pixel intensity value is 1, then system will store scan line number.
3) For the stored scan line position the system will check subsequent scan lines till scan line containing no black pixels is obtained.
4) Then the dimension of the text line will be found from stored scan line positions in image.

5. *Word Segmentation*

Word segmentation is easier task as compared to line segmentation and character segmentation. Space between two words is generally more than two or three pixels normally. Words segmentation is done by the projection based method. For word segmentation use the following algorithm used.

Determining Location of Word in Text Line

1) For each vertical scan line all horizontal pixels will be checked.
2) If any pixel having intensity value 1 is found, note position of that.
3) Subsequent scan lines are checked till we get a scan line with all pixels having intensity value 0. Position of scan line will be noted.
4) Position 1 and 2 noted in above two steps find the location and boundary of word to be segmented.

6. *Normalization*

Normalization is one of basic step for pre-processing factors of character recognition in image. Normally in normalization the character image is linearly mapped on to a standard plane by interpolation or extrapolation. The position, size of character is controlled such that the width and length of normalized plane are filled. By linear mapping, the character shape is not only deformed but the aspect ratio changes.

7. *Ridge Segmentation*

The need of ridge segmentation is for finding the break points in characters. In mathematics the ridges of a smooth function of two variables are a set of curves whose points are local maxima of the function in one dimension. For a function of N variables its ridges are a set of curves whose points are local maxima in N-1dimensions. The notion of valleys for a function can be defined by replacing the condition of a local maximum with the condition of a local minimum. The union of valley sets and ridge sets, with a related set of points called the connector set form a connected set of curves that partition or meet at the critical points of function. Histogram Equalization Apply the Histogram equalization for adjusting image intensities to enhance contrast for more better results.

8. *Wavelet Decomposition*

The wavelet decomposition is merely done for the loss less reduction of the size of image. DWT decomposes the signal into orthogonal set of wavelets. DWT decomposition is used for finding the pixels at horizontal and vertical as well as in diagonal direction. In CWT, the wavelets are not orthogonal and the data obtained by this method are highly correlated. The four divisions in the image is the same in all level of the decomposition. These four divisions of the decomposition level shave their own characteristics and preserved the values of the image as

LL: Smoothing of original image
LH : Preserves edge at horizontal side
HL : Preserves edge at Vertical side
HH : Preserves edge at diagonal side

The original image is decomposed into many levels for that used the dwt function in the MATLAB. The db1 function is used in MATLAB for wavelet transforms and four levels decomposition has carried out for the characters.

9. *Gabor Filter*

There are several image noise removal methods which are applied in transform, or time-frequency domains. In the spatial domain a small mask is convolved with the image. The mask can be mean, Gaussian or average filter. In the transform domain, first the image is translated, then is multiplied by a low pass filter and at the end performs inverse transformation to enhance the image. In the transform domain, noise in the grey levels of an image contributes heavily to the high frequency components and the most of the image energy is concentrated in low frequency components. Although, applying a low pass filter to a noisy image in the transform domain reduces the noise, also it could eliminate some high frequency components that are not related to noise and weaken sharp transitions like edges. Furthermore, the transforms which perform on the whole image, do not show any spatial information where the frequency components come. Therefore, noise reduction by low pass filtering in these domains does not preserve the local information of the image. Time-frequency transforms combine time-domain and frequency-domain analysis and allow obtaining revealing picture of the temporal localization of the signal's spectral components. Due to this problem we consider the Gabor filter as a noise reduction technique.

10. *Thinning*

The next step is to thin the processed binary image using morphological thinning operation. The thinning algorithm removes pixels from ridges until ridges are one pixel wide. The thinning of an image I by a structuring element J = (J1, J2) is given by

$$Thin(I, J) = I - (I * J)$$

Where, the subtraction is the logical subtraction defined by:

$$X - Y = X \bigcap NotY$$

Finding termination and bifurcation point using Hit or Transform After thinning apply hit or Miss transform to find termination and bifurcation points. Termination is those pixels in an image which have only one neighbor in 3X3 neighborhood. The terminations are given by applying Hit or Miss Transform on I by Jas follows:

$$M1 = (I * J)$$

Where, I is thinned image and J is the sequence of structuring element pairs (J1, J2)

$$I * J = (I\Theta J1) \bigcap (I_C \Theta J_2);$$

11. Feature Extraction

The main aim of feature extraction is to make improvement in the accuracy and speed of classifier for the pattern recognition. The extraction of features of the characters is done in such a way that complete portion of binary image covered and there is a distinct property associated with the every position. Feature extraction method categories in these three types.

1. Structural 2. Statistical 3.Hybrid

12. Feature Extraction

KNN is used for feature extraction. Features are extracted using single level wavelet decomposition as discusses earlier discussion. The approximation coefficients obtained for every character after single level decomposition is considered for that. The single level decomposition leads to approximation features. The modified KNN features are also generated in order to improve the recognition results.

III. Results & Conclusion:

Compound character is one of the features of the Marathi script and commonly used. This paper presents a system for compound character recognition for Marathi script. For this we used Marathi language OCR and used KNN for better results. The recognition of characters is done using KNN recognition scheme.

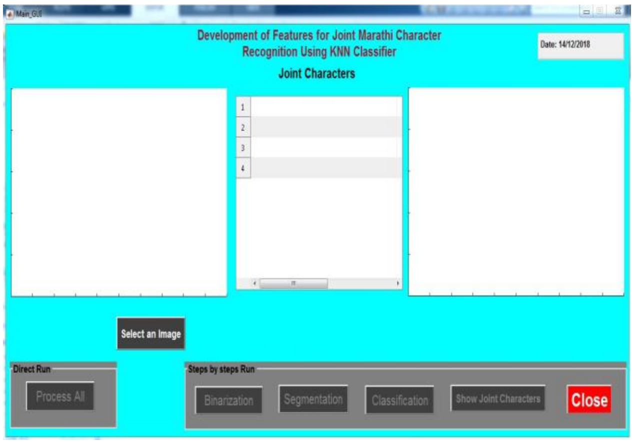
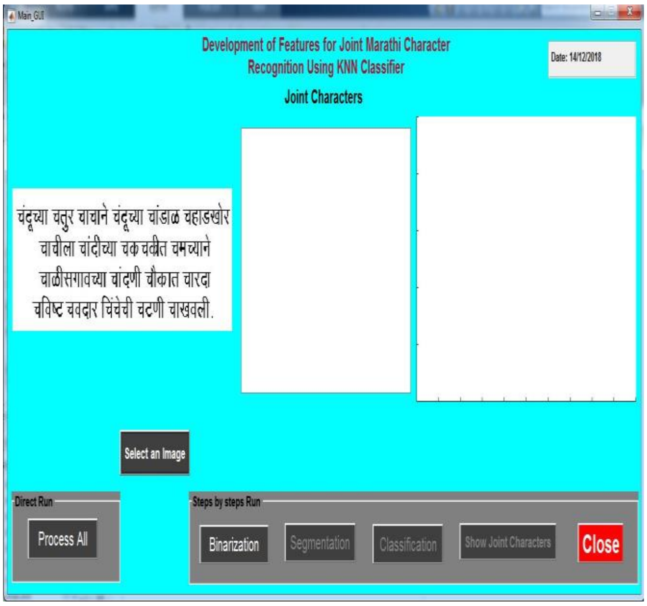
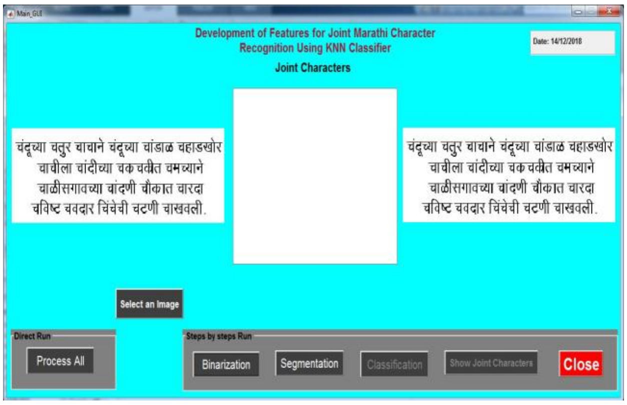


Fig2: GUI design for project work



चंदूच्या चतुर चाचाने चंदूच्या चांडाळ चहाडखोर
चाचीला चांदीच्या चकचकीत चमच्याने
चाळीसगावच्या चांदणी चौकात चारदा
चविष्ट चवदार चिंचेची चटणी चाखवली.

Fig3: Importing an image



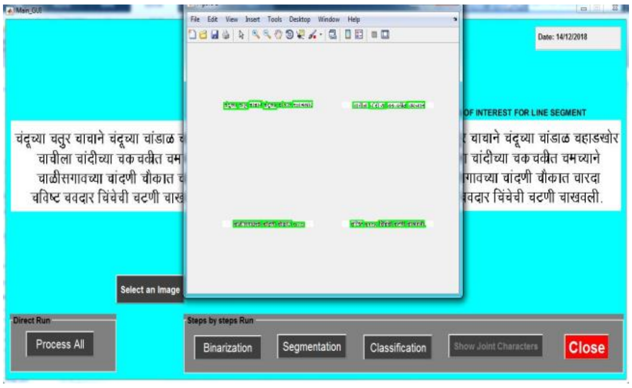
Development of Features for Joint Marathi Character Recognition Using KNN Classifier

Joint Characters

Date: 14/12/2018

चंदूच्या चतुर चाचाने चंदूच्या चांडाळ चहाडखोर
चाचीला चांदीच्या चकचकीत चमच्याने
चाळीसगावच्या चांदणी चौकात चारदा
चविष्ट चवदार चिंचेची चटणी चाखवली.

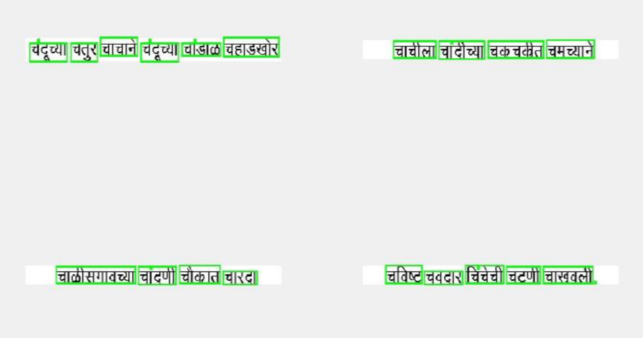
Fig4: Binarization Process



Date: 14/12/2018

OF INTEREST FOR LINE SEGMENT

Fig5: Line by line segmentation



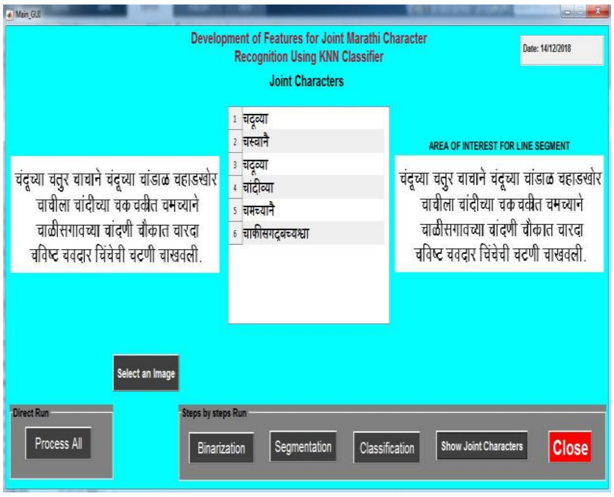
चंदूच्या चतुर चाचाने चंदूच्या चांडाळ चहाडखोर

चाचीला चांदीच्या चकचकीत चमच्याने

चाळीसगावच्या चांदणी चौकात चारदा

चविष्ट चवदार चिंचेची चटणी चाखवली.

Fig6: Word to word segmentation



Development of Features for Joint Marathi Character Recognition Using KNN Classifier

Joint Characters

Date: 14/12/2018

1	चदूव्या
2	चस्वानै
3	चदूव्या
4	चांदीव्या
5	चमच्यानै
6	चाकीसगदूबच्यश्वा

AREA OF INTEREST FOR LINE SEGMENT

चंदूच्या चतुर चाचाने चंदूच्या चांडाळ चहाडखोर
चाचीला चांदीच्या चकचकीत चमच्याने
चाळीसगावच्या चांदणी चौकात चारदा
चविष्ट चवदार चिंचेची चटणी चाखवली.

Fig7: Recognition and displaying of compound characters

This approach is KNN approach is presented for the recognition of compound Marathi character. So using KNN we get better results as compared to other methods.

REFERENCES

[1] H. S. Baird," Anatomy of a versatile page reader",. *Proc. of the IEEE, 80(7):1059-1065, 1992.*

[2] G. Nagy. "Twenty years of document image analysis", *PAMI. IEEE Trans. On Pattern Analysis and Machine Intelligence, 22(1):38-62, 2000.*

[3] C. Y. Suen, S. Mori, S. H. Kim, and C. H. Leung.," Analysis and recognition of Asian scripts - The state of the art.", *Proc. of the 6th Int. Conf. on Document Analysis and Recognition (ICDAR), pages 866-878, 2003.*

[4] U. Pal and BB Chaudhuri. "Indian script character recognition: A survey. Pattern Recognition," *37(9):1887-1899, 2004*.

[5] V. K. Govindan and A. P. Shivprasad, "Character Recognition - A Review,"Pattern Recognition, vol.23 no.7,pp 671-683, 1990.

[6] SuryaPrakash Kompalli, Srirangaraj Setlur, Venugopal Govindaraju, Ramanaprasad Vemulapati ,"Creation of data resources and design of an evaluation test bed for Devanagari script recognition.",*13th International Workshop on Research Issues on Data Engineering: Multi-lingual Information Management*

[7] SuryaPrakash Kompalli, Srirangaraj Setlur, Venugopal Govindaraju, Ramanaprasad Vemulapati "Creation of data resources and evaluation tool for multi-lingual OCR.",.*Symposium on Document Image Understanding Technology - 2003* .

[8] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition - A Survey", *Pattern recognition, vol.29, no.4, pp.641-662, 1996.*

[9] Huang YS, Suen CY. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1995; 17(1): 90-94*

[10] R.M.K. Sinha, H. Mahabala,,"Machine recognition of Devanagri script", *IEEE Trans. System, Man Cybern. 9(1979) 435-441*.

[11] Plamondon, R. Srihari, S.N. ,Ecole Polytech.,Montreal, Que.; Online and Offline HandwritingRecognition : A comprehensive Survey,1EEE Transactions On Pattern Analysis And Machine Intelligence. *VOL. 22, NO. 1. JANUARY 2000 63*

[12] U. Pal , B.B. Chaudhuri , "Printed Devanagri script OCR system", *Vivek 10 (1997) 12-24.*

[13] S. Palit, B.B. Chaudhuri,,"A feature-based scheme for the machine recognition of printed Devanagri script", *P.P. Das, B.N. Chatterjee* (Eda.) Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India 1995, pp. 163-168.

[14] I.K. Sethi, B. Chatterjee, "Machine recognition of constrained hand-printed Devanagri numerals", *J. Inst.Electron. Telecom. Eng. 22 (1976) 532-535.*

[15] R.M..K. Sinha, "A syntactic pattern analysis system and its application to Devanagri script recognition", *Ph.D. Thesis , Electrical Engineering Department, Indian Institute of Technology, India, 1973.*

[16] V. Bansal, R.M.K. Sinha, "Partitioning and searching dictionary for correction of optically read Devanagri characters strings", *Proceedings of the Fifth International Conference on Document Analysis and Recognition , 1999, pp. 653-656.*

[17] S. Arora, D.Bhattacharya, M. Nasipuri, L.Malik, "A Novel Approach for Handwritten Devanagari Character Recognition" *in IEEE – International Conference on Signal And Image Processing, Hubli, Karnataka, Dec 7-9, 2006.*

[18] M. Hanmandlu and O.V. Ramana Murthy, *"Fuzzy Model Based Recognition of Handwritten Hindi Numerals", In Proc. Intl. Conf. on Cognition and Recognition, pp. 490-496, 2005.*

[19] R. Bajaj, L. Dey, and S. Chaudhury, "Devanagri numeral recognition by combining decision of multiple connectionist classifiers", *Sadhana, Vol.27, pp.-59-72, 2002.*

[20] U. Bhattacharya, S. K .Parui, B. Shaw, K. Bhattacharya, "Neural combination of ANN and HMM for handwritten Devanagri Numeral Recognition", *In Proc. 10th IWFHR, pp.613-618, 2006.*

[21] S. Kumar and C. Singh, "A Study of Zernike Moments and its use in Devanagri Handwritten Character Recognition", *In Proc. Intl. Conf. on Cognition and Recognition, pp. 514-520, 2005.*

[22] N. Sharma, U. Pal, F. Kimura and S. Pal, "Recognition of Offline Handwritten Devanagri Characters using Quadratic Classifier", *In Proc. Indian Conference on Computer Vision Graphics and Image Processing, pp- 805-816, 2006.*