International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 07, July-2019

Handwritten Marathi Compound Character Recognition

Amol A. Kadam
M. Tech Student
Department of Electronics and Telecommunication
Shri Guru Gobind Singhji Institute of Engineering &
Technology Nanded, India

Dr. Milind V. Bhalerao
Assistant Professor,
Department of Electronics and Telecommunication
Shri Guru Gobind Singhji Institute of Engineering &
Technology Nanded, India

Mohit N. Tanurkar
M. Tech Student
Department of Electronics and Telecommunication
Shri Guru Gobind Singhji Institute of Engineering & Technology
Nanded, India

***Abstract*- Handwritten Marathi compound character recognition is an intelligent activity of a pattern recognition system. Sometimes the human brain is also confused to identify handwritten characters in a certain language. In an image of handwritten Marathi compound characters, feature extraction techniques are playing a significant role to extract special features of the image. For handwritten characters, zoning is the most popular method to extract the features. The main aim of feature extraction is to extract the relevant information of an object or image. In this system, the zoning feature extraction technique is used to extract features. Besides this, the statistical feature extraction method is also proposed. To classify the handwritten Marathi compound characters, the zoning and statistical features are stored in the form of the feature vector. The overall accuracy of handwritten compound characters recognition by applying SVM and K-NN classifier is 96.49% and 95.67% respectively.**

***Keywords- Handwritten Marathi compound character recognition, image zoning, zoning feature, statistical feature, feature extraction method, classification, character recognition*.**

I. INTRODUCTION

Recognition of handwritten characters in any script is a difficult task in the research field. Now a day's popularity of handwritten Marathi character recognition is growing very fast. The handwritten Marathi character recognition systems are mainly focused on On-line and Off-line techniques. The On-line character recognition method captures an image by a different sensor at the time of the writing process. The information of On-line handwritten character is available dynamically according to the shape and strokes of the character. Off-line handwritten character recognition takes place in the rest form of data. After writing handwritten Marathi compound characters on plain white paper, they are scanned and stored in the form of images or documents. On-line and Off-line techniques are difficult to automate for handwritten character detection or recognition. Especially in off-line handwritten character recognition system, various difficulties occur such as variation in strokes, the shape of the character, pen width, pen ink and occurs due to the effect of a mental and physical situation of person on writing style. Due to all these reasons, its effects on the accuracy of the device to recognize the characters. This paper focuses on the Off-line handwritten Marathi compound character recognition. Generally, Marathi script consists of 52 alphabets including 16 vowels (swar) and 36 consonant (vyajnas) [10]. The writing styles of Marathi compound characters is horizontal, and it is written from left to right.

II. LITERATURE SURVEY

There is a lot of work done in character recognition or pattern recognition. The optical character recognition work was started in 1970. The Devanagari character recognition includes basic simple characters, numerical and compound character done by V. Bansal [17]. The work on Offline handwritten Devanagari script segmentation by using line segmentation and word segmentation using the histogram and classifies by using a support vector machine by Ashwin S. Ramteke [14]. The work on numerals digits in Kannada and its classification by using a multi-layer neural network was proposed by R.S. Hegadi [16]. The work was done in handwritten Marathi basic character recognition proposed by P.M. Kamble and R.S Hegadi. In this paper distance classifier was used to the classification of character and statistical feature extraction techniques are used to extract the features [1]. The work was done on handwritten Devanagari compound character recognition proposed by Karbhari V. Kale, by using Zernike moment feature extraction techniques for extracting the features. After that support vector machine and the K-NN classifier is used to classify characters [12]. The work was done on zoning-based Devanagari character recognition by O.V. Ramana Murthy and M. Hanmandlu, by using the zoning-based feature extraction method, by calculating the pixel density of each zone. After, that give it to the support vector machine classifier to classify the Devanagari printed script [2]. The handwritten Marathi compound character recognition method was proposed by

Minakshi Bhandare and A.S. Kakade, by using a support vector machine classifier to classify character [10]. In the below table 1 shows the literature survey which was done earlier.

Table 1. Literature Survey

Method	*Feature*	*Classifier*	*Accuracy (%)*
N. Sharma [9]	Chain code	Quadratic	80.36
Deshpande [4]	Chain code	RE & MED	82.00
S. Arora [8]	Structural	FFNN	89.12
O.V. Ramana [2]	Zoning	SVM	88.9
Hanmandlu [5]	Vector Distance	Fuzzy set	90.64
U. Pal [7]	Gradient and Gaussian filter	Quadratic	94.24
P.M. Kamble [1]	Statistical	Minimum distance classifier	94.38

III. DATABASE DESIGNING

The handwritten Marathi compound character database is created for this work. The database is created on an A4 size of white blank paper sheets written from the different people of different age groups, at different times with different writing styles. We select randomly 35 different compound characters and written each character for 100 times by 100 different persons. The total number of images is 3500. The recorded compound character is scanned by canon scanner with the high resolution at 500 dpi with the size of 100✕100 pixels and stored the image in JPEG file format and these datasets are used for our work. Below table 2 shows the properties of the database.

Table 2. Properties of database

Property	*Description*
No. of compound characters	35
Total No. of characters	3500
Resolution	500 dpi
Image size	100*100

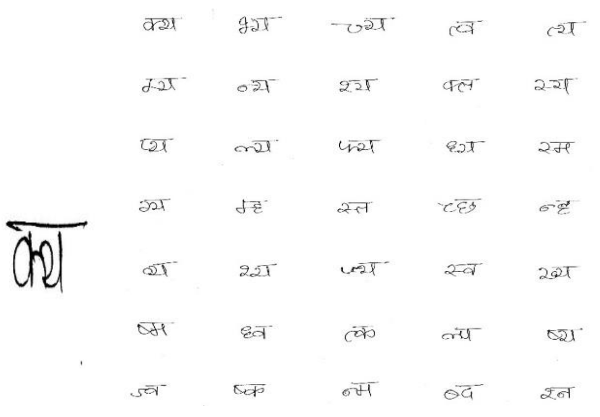


Fig 1. Sample of Handwritten Marathi compound character database image

IV. PROPOSED METHOD

In this proposed method we aim to recognize the handwritten Marathi compound characters. This is done by using the zoning-based feature extraction technique and the statistical feature extraction technique. These features are save in feature vector and this feature vector gives to support vector machine and K-NN classifier. These classifiers are used to classify the characters and recognize the characters. The basic block diagram of the proposed method is shown below in figure 2.

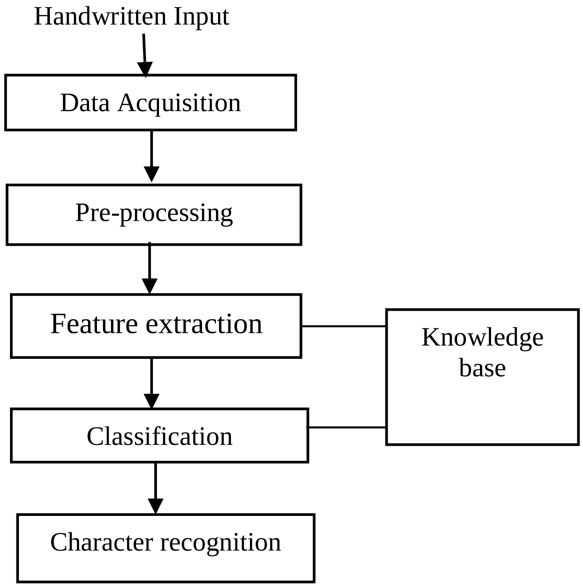


Fig 2. System architecture

The above figure shows the basic block diagram of handwritten Marathi compound character recognition. It consists of data acquisition, knowledgebase, pre-processing, feature extraction, classification and character recognition.

1. Data collection

For this research work, 100 sets of handwritten Marathi compound character samples data are collected randomly from different people of different age groups, at different times with different writing styles. This data was scanned by the canon scanner with the high resolution at 500 dpi with the size of 100✕100 pixels and store image in the JPEG file format. This data set is used to extract the features and categorize the handwritten Marathi compound characters.

2. Pre-Processing

Pre-processing plays a vital role in any pattern recognition and a handwritten character recognition system. The collected dataset at the time of scanning process documents is not clearly scanned. So, due to this generates small dots and shaded parts in the scanned image. These shaded parts and dots must be filtered by using the filtering method. The first step is image resizing into a standard dimension. The input images are in a color form and it converts into a grayscale image. Then grayscale images are again converted into black and white images or binary by using threshold value obtained by thresholding method. Resize followed by some noise to eliminate the noise which uses the median filter. Thus, after the step of pre-processing the pre-processed image of the handwritten Marathi compound characters is achieved. The basic block diagram of the pre-processing of the image is shown in figure3.

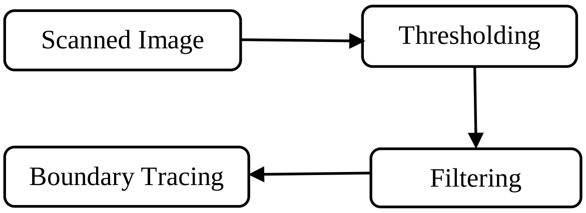


Fig 3. Block diagram of basic pre-processing steps

a) RGB to Gray image

The database contains a color character image. In the pre-processing method, the color images are converted into grayscale images. The output of the grayscale image is applied to the thresholding step.

b) Thresholding

The pre-processing method converts the grayscale image into a binary image by using the thresholding method. In this process, the boundary of the pixel is determined. The pixels above boundary level are considered as white and those below boundary level are treated as black. So, the resultant images are in binary form

c) Filtering

At the time of the thresholding process, noise is occurring in a binarized image. To apply a median filter for extracting a noise and the black shade which is appearing at edges of the image and small block spot which is present in the image. Further thresholding and filtering step of the image often resulted in some broken parts of the character in images. To rejoin the broken character of the image some morphological operations were applied on the filtered images like dilation operation has been performed on the images.

d) Boundary Tracing

The main purpose of this pre-processing step is to trace the boundary of the object to identify the connected component of handwritten character in the output of filtered image and save in array.

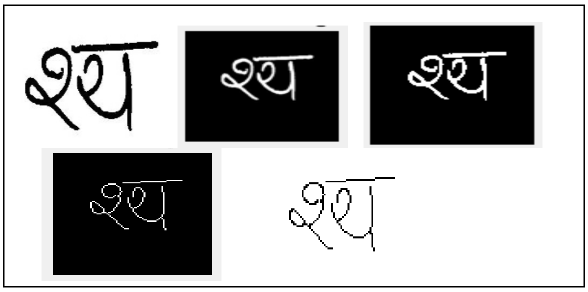


Fig 4. Pre-processing step resultant images

V. FEATURE EXTRACTION

In the system of handwritten character identification or any pattern recognition, features play a vital role. Now a day's number of feature extraction techniques are available for the handwritten Marathi or any character identification or recognition system. Chain code feature extraction was proposed by Sharma and Deshpande [9][4], Structural feature extraction was proposed by Arora [26], Vector Distance feature extraction was proposed by Hanmandlu [23], Gradient and Gaussian filter feature extraction was proposed by U. Pal [24] and Statistical feature extraction techniques proposed by P.M. Kamble [24]. For this paper two feature extraction techniques are used to extract the features that are zoning and statistical.

A) Zoning Based Feature Extraction Method

In this feature extraction method after pre-processing handwritten input image is applied to the zoning-based method to divide the image into the predefined size of the zone. The size of the processed image was (60×30) and it divides into a total of nine zones. The size of each zone is (20×10). It means that each zone consists of 20 columns and 10 rows. After the zoning process of the image, the image consists of the black pixel as a background and white pixel as a foreground. But our area of interest lies in the foreground of the handwritten character of an image to count only white pixel of each zone and store into some another array. Below figure 5 shows the zoning of the image.

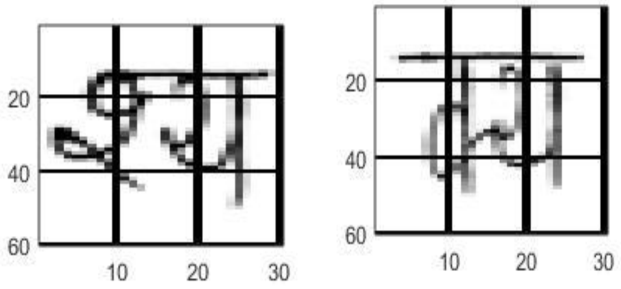


Fig 5. Zoning of image

The zone-based feature extraction method is shown to count the white pixel which is present in each zone of the image.

Zoning algorithm

```
Input: Binary image (size [60×30])
New zone = 1
For i = 1: 20
    For j = 1: 10
        read 20×10 pixels in a matrix format
        Zone [New zone] = read matrix
        New zone = New zone + 1
For new zone from 1: 9
read zone [New zone]
Count = number of white pixels in the zone
Signature array [New zone] = Count
Result: Signature array of the image
```

B) Statistical Based feature extraction techniques

In this feature extraction technique after pre-processing handwritten input image is applied to the statistically-based feature extraction technique. The basic feature like centroid, area of an object, perimeter, eccentricity, equivalent diameter and roundness of object are considered [1].

A] Centroid:

In the centroid, two coordinates X and Y specify the center mass of the handwritten character region. The first element and second element of the centroid is the horizontal and vertical coordinates of the handwritten character region. The region consists of the white pixels and the red dot is the centroid of the character.

$$X = \frac{\text{size of } M^{th} \text{ element of the character}}{2}$$

$$Y = \frac{\text{size of } N^{th} \text{ element of the character}}{2}$$

B] Perimeter:

It has been calculated by using the distance between each neighbouring pair of the pixels across the border of the region of the handwritten character [1]. It is the distance across the boundary of the handwritten character region

C] Eccentricity:

In Marathi handwritten compound character shape, size, orientation is heterogynous [1]. Eccentricity is defined as the ratio of the major axis and minor axis of the ellipse which covers the whole handwritten character. The below equation shows

$$E = \frac{PQ}{RS} \qquad 1$$

Where E= Eccentricity, PQ = major axis, and RS= minor axis.

D] Area of character:

The mass of handwritten character is the total number of white pixels which are present in binarized character images. In character count the total no. of the white pixel in the handwritten image of the character to obtain the mass of handwritten character value. This is all done after the pre-processing step of the handwritten character and normalized in standard size and then calculate the area of the handwritten character.

E] The roundness of object:

Roundness is closely related to the shape of the object or character of the images. In the roundness of the object, the shape is the dominant feature of the object rather than edges and corners of the object or character. It does not describe the radial displacement of the shape of the object or character form middle points, but it describes the overall shape of the object or character.

VI. CLASSIFICATION

In this work, two classifiers are used to classify the handwritten Marathi compound characters, support vector machine (SVM) and K-NN classifier to get better accuracy of the handwritten character recognition. In this method, first SVM classifiers are used then the K-NN classifier was applied for the handwritten compound character recognition system.

1) Support vector machine (SVM)

The support vector machine is a supervised learning classifier. It is capable to get better performance of the proposed model for handwritten Devanagari Marathi compound characters identification. SVM has a better capability to reach error-free identification or recognition. SVM has nonlinear mapped lower level input data into high dimensional feature space and determine separate hyper-plane with the large margin between the two or more class. The margin space between the two classes was determined using kernel function in input space [2]. Support vector machine (SVM) generating a model based on training dataset which indicates target values of the test image features or test data feature and gives the result that was available in training dataset of sample label pairs($x_{j,}\ \ y_j$), j =1,2,3,…..l where $x_{j,}\epsilon R^n$and y $\epsilon\{1,-1\}^l$, SVM need a solution of bellow optimize problem

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + c \sum_{i=1}^{n} \varepsilon_i \qquad 2$$

Dependent on$y_i \left(w^T \emptyset(x_i) + b\right) \geq 1 - \varepsilon_i, \quad \varepsilon_i > 0$

Here train dataset vectors x_j are map into a high dimensional vector space by the function of *Φ*. The support vector machine finds a perfect hyperplane which magnifies the distance, or more particularly the margin between the closest example of both classes. These closest examples are known as support vector. C > 0 is the fine criterion of error terms. In addition to K ($x_{j,}$ x_k) ≡$\Phi(x_j)^T \Phi(x_k)$ is known as kernel function [2]. In our system, we used the radial basis function kernel (RBF). The RBF is given by

K ($x_{j,}$ x_k) ≡ $e^{(-\gamma \|x_{ij} - x_k\|)}, \ \gamma > 0$

A search is applied to find the value ofγ which is the criterion of radial basis function. The value of variance and parameter both are select in the range of (0, 1) for gamma (γ) and (0, 1000) for the cost (C) for the support vector and examines the identification rate or recognition rate of the handwritten compound character.

2) K-Nearest Neighbor Classifier

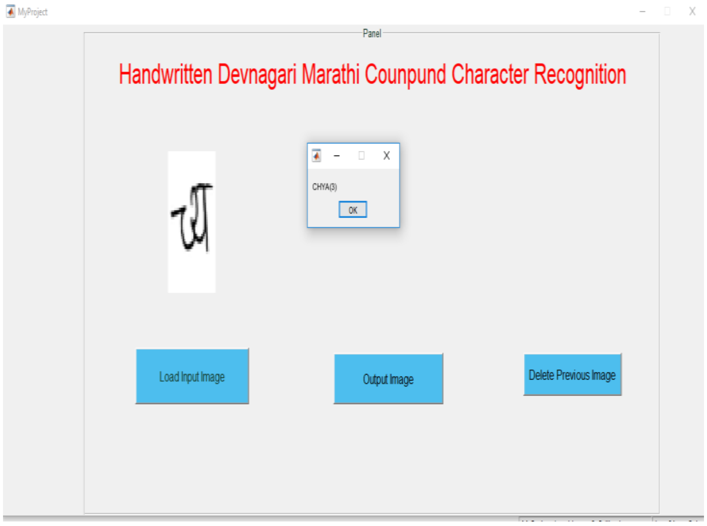
In the K-Nearest Neighbor classifier, classifies based on the similarity between the test data and train data then whatever observation is that belong to the same classes. The test numeral feature vector of the character image is classifying based on the nearest neighbor distance of the train numeral feature vector. Depending upon the Euclidean distance between the train feature set and test feature set of the handwritten character was used to decide the class of the handwritten character. The most common similarity measure for K-NN classification is the Euclidian distance metric, define between the feature vector as

$$\textbf{euclidean}\ (\mathbf{x,y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_j)^2} \qquad 3$$

Where **n** = number of features. The less distance values represent greater similarity [12].

VII. RESULTS AND DISCUSSION

In the handwritten Marathi compound character recognition system, a database of 35 distinct compound characters is created which is randomly selected from the set of compound characters. Each character written by 100 times, means 3500 databases were created for this work. This database was divided on 50% for training and 50% for testing from each handwritten Marathi compound character of the image. Zoning and statistical feature extraction techniques are used to extract the features. For this system, two classifiers are used to classify the handwritten Marathi compound characters SVM and K-NN. The comparative analysis is shown below table 3. The GUI of the handwritten Marathi compound character recognition is shown below figure 6.



CHYA(3)

Fig 6. GUI of the Result

Table 3. Comparative analysis

Method	*Feature*	*Classifier*	*Accuracy (%)*
N. Sharma [9]	Chain code	Quadratic	80.36
Deshpande [4]	Chain code	RE & MED	82.00
S. Arora [8]	Structural	FFNN	89.12
O.V. Ramana [2]	Zoning	SVM	88.9
Hanmandlu [5]	Vector Distance	Fuzzy set	90.64
U. Pal [7]	Gradient and Gaussian filter	Quadratic	94.24
P.M. Kamble [1]	Statistical	Minimum distance classifier	94.38
Proposed method	Zoning and Statistical	SVM and K-NN	96.49 and 95.67

VIII. CONCLUSION

This proposed offline handwritten Marathi compound character recognition system, Marathi script was derived from the Devanagari script. We have created handwritten compound characters database form various age group persons, at a different time with different writing styles. This database further utilizes for the classification and recognition purpose for the handwritten compound character recognition system. From the database, various features of compound characters are created using the zoning-based feature extraction method and statistical feature extraction methods are implemented successfully. We have applied a two-stage classification for improving the resultant accuracy of the handwritten character recognition. The two classifiers were used SVM and k-NN to classifies the handwritten compound characters. We tested our proposed system on 3500 images for handwritten Marathi compound character and obtained an average accuracy of 96.49% form an SVM classifier and the average accuracy of k-NN was 95.67%. In the future, we would like to extract more features and implement this proposed system on various languages of a printed script as well as the handwritten script for improving the recognition rate of this proposed system. Table 3 shows some resultant accuracy from the literature survey.

REFERENCES

[1]. Kamble, Parshuram M., and Ravindra S. Hegadi "Handwritten Marathi basic character recognition using the statistical method.". *Emerging Research in Computing, Information, Communication, and Applications* 3 (2014): 28-33.

[2]. Murthy, OV Ramana, and M. Hanmandlu."Zoning based Devanagari character recognition." *International Journal of Computer Applications* 27.4 (2011): 21-25.

[3]. Arora, Sandhya, et al " A two-stage classification approach for handwritten Devnagari characters." *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*. Vol. 2. IEEE, 2007.

[4]. Deshpande, Parag S., Latesh G. Malik, and Sandhya Arora. "Fine Classification and Recognition of Hand Written Devnagari Characters with Regular Expressions and Minimum Edit Distance Method." *JCP* 3.5 (2008): 11-17.

[5]. Hanmandlu, Madasu, OV Ramana Murthy, and Vamsi Krishna Madasu "Fuzzy Model-based recognition of handwritten Hindi characters.". *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA 2007)*. IEEE, 2007.

[6]. Borse, Sunita B., Madhuri Bhalekar, and M. U. Kharat. "Handwritten character recognition with optimal zoning using GA."

[7]. Pal, Umapada, et "Off-line handwritten character recognition of Devanagari script "al*Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Vol. 1. IEEE, 2007.

[8]. Arora, Sandhya, et al."Recognition of non-compound handwritten Devanagari characters using a combination of mlp and minimum edit distance." *arXiv preprint arXiv:1006.5908*(2010).

[9]. Sharma, Nabin, et al."Recognition of off-line handwritten Devanagari characters using the quadratic classifier." *Computer Vision, Graphics and Image Processing*. Springer, Berlin, Heidelberg, 2006. 805-816.

[10]. Minakshi Sanjay Bhandare and Anuradha Sopan Kakade."Handwritten (Marathi) compound character recognition." *2015 International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS)*. IEEE, 2015.

[11]. ." Kadhm, Mustafa S., and Asst Prof Dr. Alia Karim Abdul. "Handwriting word recognition based on SVM classifier*International Journal of Advanced Computer Science and Applications* 1 (2015): 64-68.

[12]. Kale, Karbhari V. "Zernike moment feature extraction for handwritten Devanagari (Marathi) compound character recognition." (2014).

[13]. Vijayaraghavan, Prashanth, and Misha Sra."Handwritten tamil recognition using a convolutional neural network." (2014).

[14]. Ramteke, Ashwin S., and Milind E. Rane. "Offline handwritten Devanagari script segmentation." *International Journal of Scientific and Technology Research* 1.4 (2012): 142-145.

[15]. Dhandra, B. V., Gururaj Mukarambi, and Mallikarjun Hangarge."Zone-based features for handwritten and printed mixed Kannada digits recognition." *IJCA Proceedings on International Conference on VLSI, Communications, and Instrumentation (ICVCI)*. No. 7. 2011.

[16]. Hegadi, Ravindra S "Classification of kannada numerals using a multi-layer neural network." *Proceedings of International Conference on Advances in Computing*. Springer, New Delhi, 2013.

[17]. Bansal, Veena, and R. M. K. Sinha "Integrating knowledge sources in Devanagari text recognition system." . *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30.4 (2000): 500-505.

[18]. Bhalerao, Milind, et al. "Combined classifier approach for offline handwritten Devanagari character recognition using multiple features." *Computational Vision and Bio Inspired Computing*. Springer, Cham, 2018. 45-54.

[19]. Bhalerao, Milind, Sanjiv Bonde, and Madhav Vaidya."Frequently Used Devanagari Words in Marathi and Pali Language Documents." *Advances in Computer Communication and Computational Sciences*. Springer, Singapore, 2019. 97-110.

[20]. Vaidya, Madhav, et al."Discrete Cosine Transform-Based Feature Selection for Marathi Numeral Recognition System." *Advances in Computer Communication and Computational Sciences*. Springer, Singapore, 2019. 347-359.

[21]. Vaidya, Madhav V., and Yashwant V. Joshi. "Handwritten numeral identification system using pixel level distribution features." *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, Cham, 2017.

[22]. Vaidya, Madhav, Y. V. Joshi, and Milind Bhalerao."Marathi numeral identification system in Devanagari script using discrete cosine transform." *Int. J. Intell. Eng. Syst* 10.6 (2017): 78-86.

[23]. Vaidya, Madhav V., and Y. V. Joshi. "Marathi numeral recognition using statistical distribution features." *2015 International Conference on Information Processing (ICIP)*. IEEE, 2015.

[24]. Shelke, Sushama, and Shaila Apte."A multistage handwritten Marathi compound character recognition scheme using neural networks and wavelet features." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 4.1 (2011): 81-94.

[25]. Ajmire, P. E., R. V. Dharaskar, and V. M. Thakare."Handwritten Devanagari (Marathi) compound character recognition using seventh central moment." *International Journal of Innovative Research in Computer and Communication Engineering* 3.6 (2015).