Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 5470-5479

An Approach for OCR detection and Classification for Devanagari Printed Text using Deep Learning

Mr. Prashant Sopanrao Kolhe[1], Dr. Ulhas Shiurkar [2]

[1]Research Scholar: Department of Electronics and Telecommunication Engineering TPCT'S College of Engineering, Osmanabad, India,
[1] prashantkolhe4005@gmail.com
[2]Directot: Deogiri Institute of Engineering and Management Studies, Aurangabad, India
[2]shiurkar@gmail.com

Abstract: Optical Character Recognition (OCR) is the method of interpreting text from digital documents automatically. It is a large area of signal processing science. In several languages of India, such as Hindi, Nepali, Marathi, Sindhi etc., Marathi has used. The Marathi language is used by more than 300 million people worldwide. This pattern forms the basis of the communities of India. It plays a significant role in the production of manuscripts and literature. OCR research has been done for in banks, post offices, defence organizations, library modernization, etc., because of its potential applications. Several techniques are available for the character segmentation of handprint Gujrati, Bangla, Tamil, Hindi, etc., with these methodologies. However, much work is done for both the given material, but for the laboratories, it is only limited.. In this paper, proposed a Deep Learning based Convolutional Neural Network (CNN) classifier technique has used for OCR System of printed as well as scanned newsprint Marathi script. This research system deals with various feature extraction and feature selection techniques with CNN classification. In experimental analysis we train system around more than 51 characters that produces better detection accuracy for test images.

Keywords: **n**Character Recognition; Marathi Characters, binarization, pre-processing, classification, segmentation, feature extraction, deep learning.

1. Introduction

In computer science demands to make the machines work and function as humans do. Tremendous progress has been made in the fields of Artificial Intelligence, Machine Learning, and Deep Learning which have attracted people towards computer vision and human-computer interaction. This has made machine possible to think, see and process things around it in the same way as human do. This is the era of machine learning. The natural way of interacting with computer is through Character Recognition. There is great scope of research available for researchers and scientists in the field of Character Recognition. In Character Re-cognition process input is provided in the form of text image from which character is detected and recognized by the machine. During this process data is converted into code which is understood by the machine. The modes of input may vary depending upon the application. Input can be taken in the form of image, file, csv, etc. There are several applications where character recognition is used namely Bank automation system , Handwritten character recognition for Indian scripts , postal automation system , car number plate characters, form filling ,etc. There are 5 common phases involved in character recognition. These phases are Pre-processing, Segmentation, Feature Extraction, Classification and Post-Processing. Pre-processing includes Binarization, Normalization, Sampling, De-noising and Thinning .The Segmentation phase includes horizontal which includes upper, lower and main segmentation and vertical includes line, word, and character segmentation .There are various methods for feature extraction and classifier such as CNN, SVM, RNN, Hough Transform, HMM, MLP and many more. It ' s very to decide which method to be implemented according to requirement. Dataset play crucial role in driving the accuracy for these methods .Accuracy varies depending upon the strokes and structure of character dataset .Accuracy also depends upon the volume of trained dataset.

The organization of paper in section 2; we proposed various existing methodologies of literature, in section 3; we demonstrates research methodology of proposed work with numerous machine learning and deep learning techniques. Section 4; describes about results and discussion of system while section 5 shows the conclusion and future work of system.

2. Literature Survey

Compound characters of Marathi language are derived from Devanagri. A method by Shelke et al [1] is proposed for recognition of these characters using multistage feature extraction and classification scheme. Feature extraction stage is based upon the structural features and the characters are classified into 24 classes according the structural parameters. The final stage of feature extraction uses wave transform. An approach based on invariant moments for the recognition of isolated Handwritten Numerals and their divisions was described by Ramteke et al [2]. This projected technique is independent of dimension, slant, direction, translation and other differences in handwritten characters. Recognition of handwritten Marathi Characters are more complex than corresponding handwritten English characters due to variations in order, number, direction and shape of the essential strokes. The Gaussian Distribution Function has been accepted for classification purpose. The success rate of the process is found to be 87%. Pankaj et al [3] presents a methodology to recognize the unconstrained handwritten Marathi characters. 500 handwritten characters are used for experimentation which is collected from 10 people. Array based feature extraction and artificial neural network classifier is used for recognition. The handwritten characters are scanned and preprocessed. Feature extraction is applied on every individual character. These features form the feature vectors which are used as an input for training the back propagation neural network. Then the testing is carried using individual characters and sentences. The accuracy obtained for recognition of individual handwritten characters is 92% and for handwritten sentences is 88.25%.

Compound character in Devnagari script is the combination of one or more characters. Recognition of such characters is one of the challenging tasks. These characters are complex in structure because they are treated as fusion of two or more characters. Sanskrit, Hindi, Marathi and Nepali languages are written using Devnagari script. The compound characters are available in all these languages. The combination of basic characters such as vowels and consonants are used to form the compound characters. Marathi language contains large number of compound characters and recognition of such characters is one of the challenging tasks for researchers. A method which uses seventh central moment is proposed by Ajmire et al [4]. The SVM classifier is used and the overall performance obtained is 93.87%. U. Pal et al [5] extracted the features based upon the directional and curvature information in the characters and used the combination of two classifiers SVM and modified quadratic discriminate function for classification of characters. Kamble et al [6] calculated the correlation coefficient between the characters available in the database and every extracted characters. Then characters are recognized using template matching method. The high value of correlation coefficient between any two characters indicate successful match of character. The proposed algorithm will yield encouraging results.

The OCR can likewise be exceptionally helpful for postal location acknowledgment, cheque acknowledgment, video text identification, re-establishing of old and authentic reports [7] etc. In the event that somebody needs to change or modify some portion of document, the entire report should be physically arranged with the assistance of human composing administrators. With the scanned document images of various reports, we can't utilize the search tool to find/locate a specific word or expression. Another advantage of digitizing the archive is the space required to store the records is less when contrasted with direct putting away the scanned report pictures. Devanagari is the most fundamental script for the huge number of dialects, like Marathi, Hindi, Gujarati, Nepali, Konkani, Bhojpuri, and so forth. Devanagari lipi is utilized in the establishment of 12 well-known dialects in India. Devanagari characters hang down from a flat straight line (header line or Shirorekha [8]) composed at the head of the each character. The main stroke, or strokes, in a character, are composed from the left of page to right side and are then trailed by any down strokes and lastly the head strokes are included. It ordinarily takes somewhere in the range of three and five strokes to compose a Devanagari character.

Numerous Indian Government as well as private workplaces are utilizing Devanagari Script for correspondence as well as keeping the records [9]. Most of the time printed or handwritten paper reports/documents are captured by digital means and put away as computerized records. Through the use of DIP strategy, paper documents are stored in a computerized but processable format. Best in class Optical Character Recognition (OCR) strategies are equipped for giving adequately high correctness on printed archives. Very less work has been done in [10] to grow automatic OCRs for handwritten Marathi language. Some work was reported earlier to recognize and reproduce the Marathi characters without any modifiers as well as with few modifiers. The greater part of the ongoing Indic OCR frameworks Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) to categorize letters in the picture

The Hindi offline study[11] explored the use of the Artificial Neural Network (ANN), Fuzzy Logy, GA, SVM, KNN to define the Secret Markov Model (HMM), Bacterial Foraging (BF), Clonal Selection (CSA), etc. Besides,[12] recognized the Hindi text words offline by ignoring upper/below modifications and half characters, extracting characteristics and identifying areas of character and middle words. In [13], pre-processing binarization removal were performed by Hindi OCR; features are removed by an unsupervised clustering method called as K-Means; and finally, kernel-based linear SVM classification. The next Hindi OCR[14] was the Rough probability-based Fuzzy classification techniques, like fuzzy logic with SVM, with various similarity functions. The Hindi OCR[14] Binarization different soft computing and supervised classification algorithms have been used to detect and classify. The Devanagari OCR[15] removed chain coding characteristics, gradients and directional edge detection features, decreased LDA characteristics and, finally, categorized SVM characters.

Another Hindi OCR [16] used effective multi-classifier feature extraction techniques. Simultaneously, the offline Devanagari OCR [17] extracted directional features based on gradient strength, angle and histogram (SOG, AOG and HOG), classifying features by combining various data split techniques as well as cross-validation classification. It then used the Gaussian filter, followed by HOG features, to sample SOG and AOG. The handwritten Marathi OCR[18] study discussed the recognition steps: elimination of morphology and threshold noise; correction of Hough-transform skew; bounding segmentation; normalization; connected feature extraction based on pixels; and finally, five-fold validation classification. Also, [19] the handwritten OCR screened by Marathi was discussed, and multi-oriented features were recognized by using invariable scales and rotation to obtain background and foreground information where characters were first split into multi-circular regions. Three centroid features have calculated for all areas; each character segment was divided into two clusters and evaluated the system's accuracy.

3. Research Methodology

The proposed method starts with scanning of handwritten Marathi text, followed by pre-processing steps. The pre-processed images were utilized for line and word segmentation by finding blank row and column respectively. The segmented words are feed to CNN network. The CNN network will extract the features and make the judgement to produce the recognized result. The workflow of proposed method is mentioned in figure 1.

The Figure 1 shows the execution of proposed system with each dependant layers, in the first data collection has done from various devanagari printed document images that has for validation. For module training around 55 characters have been used for entire word set.

Pre-processing

In that phase few steps like data scanning, pre-processing of data, normalization, feature extraction, feature selection etc. these are describes in below.

Data Scanning : Data scanning The document image obtained by scanning a hard copy news document as a black & white photograph using a flatbed scanner is represented as a two dimensional array.

Text Pre-processing: Pre-processing stage consists of compression and binarization steps. Binarization:- It is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by thresholding.

Document Segmentation: Lines & Terms Segmentation: The preliminary segmentation consists of the steps that follow:

We calculate the vertical projection of both the document picture box separately from of the given text. Develop one matrix that contains binary image for all the column in a row. So from that number, the row line is isolated from the text. We also determine the horizontal representation of the text picture box by finding the threshold value and deleting it. The row that contains the highest number of black points is known to be the line of both the heading and will be omitted The separate element boxes below the left margin: To do it too, we create a vertical object projections starting from the location of the left margin to the bottom row of the word image box. Columns without black pixels are viewed as boundaries for the extraction of character-corresponding image boxes.

To do this, we calculate the vertical projection of the image, starting from the top row of the image to the header line location, using separate symbols from the top strip. For removing top modifier symbol boxes, column with no black pixels have been used as delimiters..

Feature Extraction and Feature Selection: One of its most significant steps in designing a grading system is image segmentation. This phase explains the different characteristics that have been chosen to identify the identified participants. For the identification of Marathi characters, several elements have been collected. Consider this function as the measurement of histograms and Gary Level Co-occurrence Matrix (GLCM). The Histogram block computes this same graphical representation of the components in the source. The secondary part is GLCM; the gray-level co-occurrence matrix (GLCM), also known as a spatial dependency matrix, is a mathematical method of analysing textures that consider pixels' functional relations. The GLCM performance abilities a picture's surface by measuring how frequently vector pairs occur in an image with unique values and in a given spatial association, generating a GLCM, and then obtaining statistical measurements from the whole matrix..

Classification: For handwritten Devanagari word recognition, a simple 16 layer CNN model is proposed. The handwritten structure for OCR. The first layer in the OCR system is an input layer that consists of 32 x 32 binary representations of the raw pixel values. C1 has 96 filters with a scale of 3 x 3 for the initial convolution layer. The feature map that appears as 2D planes in the figure is extracted by C1. All the units in a feature map share a common weight structure, so they are initiated at different locations by similar features. Each unit in a layer obtains its contribution from the earlier layer from a small neighborhood in the same place. Since all the units are run independently from the data taken from a local neighborhood, they recognize nearby features such as corners, edges, and end points. C1 has 1 output and 1 input. Batch normalization and the ReLU layer are accompanied by the convolution layer. Activations and gradients received from the convolution layer are normalized by the batch normalization layer. The ReLU is a nonlinear activation function commonly used by the thresholding technique to minimize the characteristics. After the ReLU layer, the pooling layer (S1) follows. By averaging the features in the local region for the greatest benefit, the pooling layer reduces feature map components obtained from the convolution layer. Because the basic location of the feature varies for different images of a similar term, it is desirable that the structure does not learn the characteristics from the absolute position, but that it should acquire knowledge of the characteristics' relative position.

The pooling layer achieves this purpose and makes moving and twisting the classifier increasingly safe. The next convolution layer (C2) has 128 3x 3 filters trailing this layer of sub-sampling (pooling). By taking a contribution from the previous layer S1, each feature map in the C2 layer is generated. From the 3 x 3 neighborhood, the units in C2 obtain their contribution at the indistinguishable location of some layers in S1 and not all. Trimming the trainable parameters is the reason behind not connecting C2 parameters with all parameters of S1. The output of this layer of convolution is sub-sampled, transformed and sent to the next layer. Here, following batch normalization and ReLU layers, we used a total of 4 convolution layers. Before the completely linked layer, a dropout layer with a probability of 0.5 is used. The completely linked layer has 104 groups of output. The network depth and the size of different layers to be used within the network structure are incredibly dependent on the dataset of training. Thus, we tested different designs during experimentation by adjusting the number of layers and parameters of each layer (number of filters and filter size, etc. and suggested the simple CNN architecture of 20 layers to recognize the handwritten Devanagari words that reproduce the words recognized in the printed form. The findings of the experiment are listed in the results section.

Another of the crucial things to note is that there are gray-scale frames that we have made. The measures used to construct the image region areas continue to follow:

1. Generating a 300x300 pixel image.

2. Reignited the idea to even a pixel of 28x28.

3. Transform into a 1D graphics processing array.

4. Place the data collection in an excel sheet (.csv).

The first step towards making our illustration was to select the original protagonist's right size so that it was easier to create images and less space for errors to be completed. If we take the initial aspect ratio to be 720x720 pixels, then there are odds of making mistakes when making straight lines and making some curves that we know are by far the essential part of this script. Or imagine if we take very little of the original image scale, then it will be nearly difficult to carry care of all the personality characteristics. We, therefore, created the old image size to be placed to make sure pixels.

Here we look at an approach of the CNN use and workflow description using our proposed process. The backward computational order determination system considers the data size of the layer with the vector size for each layer that contains parameters, then determines the order of differential computation of the mentioned back - propagation algorithm. The configuration parameters has updates with tensorFlow environment.

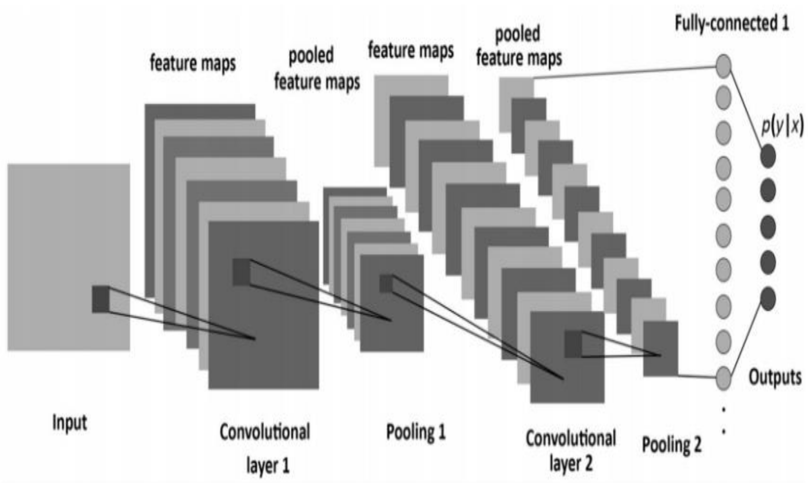

Figure 2. Proposed Convolution Neural Network (AlexNet).

Result and Discussion

We demonstrated the deep learning CNN model. Again, here before designing and performing the actual model, we first grounded our dataset to absolutely fit for our system. The augmentation approach has been utilized to boost the dataset. We prepared our own dataset in this work. In this part, we illustrated the working of the convolutional network with its distinct levels. We utilized the Convolutional level, max pooling and Dense i.e. fully associated level to put this model into practice. Here we discussed about the parameters utilized and the outcome of our system. In below figures we demonstrate a various experimental results.

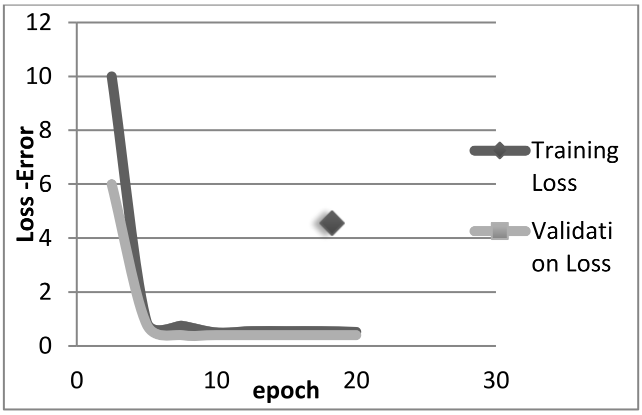

Figure 3 : Accuracy of system for training and testing

The Figure 3 describes loss of data during the module training and validation of system, various epochs has given for training and testing respectively. When epochs has given 5 or more it provides consistence loss for both functions.

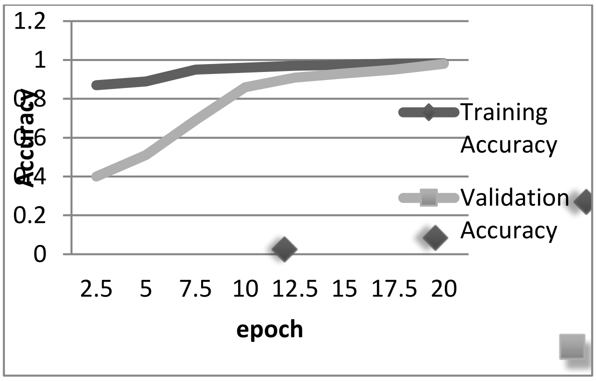

Figure 3 : Accuracy of system for training and testing

The above Figure 4 describes detection accuracy for training and validation with various epochs. It provides good accuracy for both validations, moreover when it evaluated with numerous machines learning algorithm it better than existing approaches.

Table 1: Performance evaluation of proposed system with various existing systems

Methodology	Accuracy	Error Rate
Devenagari Character	88.60	2.65

recognition using SVM and MQDF [5]		
Marathi Vowels Recognition using Correlation Coefficient [6]	84.80	6.52
OCR Postediting in Historical Documents [7]	90.00	3.95
OCR for Sanskrit using CNN [8]	92.40	4.12
Hybrid Deep Architecture for printed documents [10]	88.50	7.2
OCR with Soft Computing [14]	95.30	2.75
Proposed with CNN	98.23	0.14

The above Table 1 describes the accuracy of proposed system with various exiting machine learning and soft computing techniques. System provides better efficiency then [8] and [6] where used deep learning classifiers for text detection. However we evaluate the system on large devanagari text documents that occurs very less error rate for detection as well as classification.

6. Conclusion

This paper presents a system for printed documents Marathi character recognition. The system has been evaluated on a large amount of document images of Marathi characters. Many researchers used the word as well as character segmentation approach for recognizing the printed Marathi letters. With the proposed methodology, the requirement for character segmentation is totally eliminated. Moreover, the issues identified with upper and lower modifiers and combined/fused characters are resolved to great extent. For multiple experiments we have proposed Convolutional Neural Network (CNN) based deep learning approach to recognition of printed Marathi text. The result indicates that the model's overall efficiency is complicated but can also t improved by growing the observations for network training. However, the CNN architecture is more precise. The fundamental explanation for errors in identifying written Marathi characters is the inaccurate stratification of the affected or broken symbols. In India, a large number for better access, share and indexing of ancient writings and textbooks remain digitalized. It will undoubtedly benefit social sciences, economy and linguistics for those other research communities in India. Specific tests on the limited fonts and sizes are currently carried out. In contrast with current software and learning techniques, the suggested methodology achieves improved detection exactness. Work is underway to increase efficiency, speed and large-scale research. However, the approaches described here, like the Hybrid CNN algorithm, seem very successful.

References

[1] Sushma Shelke, Shaila Apte, “A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features”, International Journal of Signal Processing and Pattern Recognition Vol. 4 Mar. 2011.

[2] R. J. Ramteke, P. D. Borkar, S. C. Mehrotra ―Recognition of Isolated Marathi Handwritten Numerals: An Invariant Moments Approach‖, Proceedings of the International Conference on Cognition and Recognition.

[3] Pankaj Kale, Arti V. Bang, Devashree Josh, “Recognition of Handwritten Devanagari Characters using Machine Learning Approach”, International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-3, Issue-9, Sept.-2015.

[4] P.E.Ajmire, R V Dharaskar, V M Thakare, “Handwritten Devanagari (Marathi) Compound Character Recognition using Seventh Central Moment”, International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 6, June2015.

[5] U. Pal, S. Chanda, T.Wakabayashi, F. Kimura, “Accuracy Improvement of Devanagari Character Recognition Combining SVM and MQDF” , Proc. 11th Int . Conf. Frontiers in Handwriting Recognition , Qu ́ ebec, Canada, Aug. 19-21, pp. 367-372,2008.

[6] Parshuram M. Kamble, Balasaheb J. Kshirsagar, “Handwritten Marathi Vowels Recognition using Correlation Coefficient”, International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 298-301

[7] A. Poncelas, M. Aboomar, J. Buts, J. Hadley, and A. Way, “A Tool for Facilitating OCR Postediting in Historical Documents”, Workshop on Language Technologies for Historical and Ancient Languages, LT4HALA (2020), arXiv:2004.11471

[8] M. Avadesh and N. Goyal, “Optical Character Recognition for Sanskrit using Convolution Neural Networks,” 13th IAPR Int. Work. Doc. Anal. Syst., pp. 447–452, 2018

[9] N. Babu and A. Soumya, “Character Recognition in Historical Handwritten Documents – A Survey,” 2019 Int. Conf. Commun. Signal Process., pp. 299–304, 2019.

[10] C. Biswas, P. S. Mukherjee, K. Ghosh, U. Bhattacharya, and S. K. Parui, “A Hybrid Deep Architecture for Robust Recognition of Text Lines of Degraded Printed Documents,” 24th Int. Conf. Pattern Recognit., pp. 3174–3179, 2018

[11] Indian, A., and Bhatia, K. (2017) “A survey of offline handwritten Hindi character recognition”, in Third International Conference on Advances in Computing, Communication & Automation, IEEE Press, pp. 1–6.

[12] Garg, N. K., Kaur, L., and Jndal, M. (2015) “Recognition of offline handwritten Hindi text using middle zone of the words”, in Fourteenth International Conference on Computer and Information Science, IEEE Press, pp. 325–328.

[13] Gaur, A., and Yadav, S. (2015) “Handwritten Hindi character recognition using K-means clustering and SVM”, in Fourth International Symposium on Emerging Trends and Technologies in Libraries and Information Services, IEEE Press, pp. 65–70.

[14] Chaudhuri, A., Mandaviya, K., Badelia, P., and Ghosh, S. K. (2017) “Optical character recognition systems for Hindi language”, in Optical Character Recognition Systems for Different Languages with Soft Computing: Studies in Fuzziness and Soft Computing, vol. 352, Springer, Cham, pp. 193–216. [15] Shitole, S., and Jadhav, S. (2018) “Recognition of handwritten Devanagari characters using linear discriminant analysis”, in Second International Conference on Inventive Systems and Control, IEEE Press, pp. 100–103.

[16] Yadav, M., and Purwar, R. (2017) “Hindi handwritten character recognition using multiple classifiers”, in Seventh International Conference on Cloud Computing, Data Science & Engineering – Confluence, IEEE Press, pp. 149–154.

[17] Bhalerao, M., Bonde, S., Nandedkar, A., and Pilawan, S. (2018) “Combined classifier approach for offline handwritten Devanagari character recognition using multiple features”, in Hemanth D. and Smys S. (eds) Computational Vision and Bio Inspired Computing: Lecture Notes in Computational Vision and Biomechanics, vol. 28, Springer, Cham, pp. 45–54.

[18] Kamble, P. M., and Hegadi, R. S. (2016) “Comparative study of Handwritten Marathi characters recognition based on KNN and SVM classifier”, in Santosh K., Hangarge M., Bevilacqua V., and Negi A. (eds) International Conference on Recent Trends in Image Processing and Pattern Recognition: Communications in Computer and Information Science, vol. 709, Springer, Singapore, pp. 93–101.

[19] Bhandare, M. S., and Kakade, A. S. (2015) “Handwritten (Marathi) compound character recognition”, in International Conference on Innovations in Information, Embedded and Communication Systems, IEEE Press, pp. 1–4.

[20] Tripathy, N., Chakraborti, T., Nasipuri, M., and Pal, U. (2016) “A scale and rotation invariant scheme for multi-oriented character recognition”, in 23rd International Conference on Pattern Recognition, IEEE Press, pp. 4041–4046