# A REVIEW: CHARACTER SEGMENTATION AND RECOGNITION OFMARATHI LANGUAGE

**Saurabh Ravindra Nikam**[1], **Dr. D. L. Bhuyar**[2], **Mr. B. R. Guru**[3]

[1]*PG Student, Department of Electronics and Telecommunication CSMSS's CHH. Shahu College of Engineering, Aurangabad, India.*

[2]*Associate Prof. , Department of Electronics and Telecommunication CSMSS's CHH. Shahu College of Engineering,Aurangabad, India.*

[3]*Assistant Prof., Department of Electronics and Telecommunication CSMSS's CHH. ShahuCollege of Engineering,Aurangabad, India.*

**Abstract—** The number of distance dimension styles like the nearest neighbor, similarity, Linear-correlation, cross-correlation, and hamming were used to find the distance between characters. In this paper, we propose a statistically grounded point birth approach for the recognition of handwritten Marathi characters. These features are dependent on area, shape, exposure, border, and other variations in handwritten characters. 200 samples of each character from different pens have been collected and the database is reimbursed. 100 samples of each character were treated as training samples and average curiosity, exposure, and center mass of graveness features were estimated for these training samples. A distance- grounded approach is used to classify the remaining 100testing samples of each character. The results show a satisfactory performance rate.

**Keywords**: Character recognition, Bracket, Perimeter, Eccentricity, Mass of character, Exposure.

## 1. INTRODUCTION

The Indian Devanagari character segmentation and recognition system, which defines the capability of a machine to dissect and identify the script characters. Over the last many decades, machine reading has grown day by day. Optic character recognition has come one of the most successful operations of technology in the field of image processing and artificial intelligence. The Optical Character Recognition system, delved by maintaining focus on the Marathi language Character, is overcome through the limitations of the being methodologies and it has been enhancing day by day. Its bracket is grounded upon two major criteria the process of data accession and the type of textbook type ( Noise/ Noise reduced). The thing state is to cost the character of the Marathi language into digital form after identification. The introductory Character set of the Marathi language is called Akshara's. The different words in the Marathi language correspond of Akshara's which is farther nominated as Samyuktaksharas. The Samyuktaksharas Correspond of vowels, consonants, and conjoint characters. It also includes composite characters, which correspond of two or further introductory characters. The Marathi language consists of 36 consonants and 12 vowels in addition to 14 vowel modifiers. Besides consonants and vowels, it also contains a modifier called Kana, a slating line placed at the top of the character, and Matra's which are placed at the leftor right part of the character.

The complexity of the script increases with the presence of partial characters. Another type of modifier present in the textbook is upper and lower modifiers which makes the script indeed more complex. It has no conception of upper or lowercase characters. The jotting style in the Marathi language is from left to right. It's a phonetic script. The Marathi language is phonetic as words are written exactly as they're pronounced. It's also syllabic script, which means that textbook is written using consonants and vowels that together form syllables. Character segmentation is an operation that seeks to putrefy an image of a sequence of characters into sub-images of individual symbols. It's one of the decision processes in a system for optic character recognition (OCR). Its decision, that a pattern insulated from the image is that of a character, can be right or wrong. It's wrong sufficiently frequently to make a major donation to the error rate of the system. For this recognition, we need to maintain a data set of checkup images of documents.

Later several operations like preprocessing point birth and its separate bracket are done. The technology of Marathi language character recognition had been led to more transfigure development. Optic Character Recognition (OCR) in the field of exploration in character recognition of Marathi language as well as in Artificial Intelligence and digitization also. Image Processing forms the core exploration area within engineering and computer wisdom disciplines too. It's among fleetly growing technologies moment, with its operations in colorful aspects of services.

## 2. LITERATURE REVIEW

In India, there are 18 functionary (Indian constitution accepted) languages. Utmost of these languages has been written in Marathi language. The colorful inquiries have been enforced on Marathi language character recognition. Image of written document in Marathi language is given as an input to the system and the affair, as a machine editable train, has been enforced. This affair train is compatible with utmost typesetting software. The proposed system excerpts words from the image of the document. The segmented words discerned into the sub character position part. The structure of the script was used in the proposed scheme for segmentation. A unique set of features for the recognition problem, which are computationally simple to prize, is proposed. The final recognition is fulfilled by employing with classifiers, which is grounded on the Support Vector Machine (SVM) system (1).

Patterns initiate segmentation fashion for optic character recognition that contributes to document structure analysis. Connected Element pattern are uprooted and spatial interrelations between factors. It measured and grouped into meaningful character patterns. Stroke shapes are anatomized. An extended form of pattern acquainted segmentation is considered. An effective and computationally focused system for segmenting character and plate’s part of scrutinized images grounded on textural cues is used. It's assumed that the plates part have different geometric parcels than thenon-graphics (textbook) part. For segmentation, perpendicular and vertical protuberance fashion is used and for point birth, convex housing fashion is used (2).

The data set is maintained for medication and character bracket. Dataset medication part can be divided into three subsystems; preprocessing, character birth, and separation of Training and Testing set. Character birth deals with checkup documents, cropping single characters and labeling them. Preprocessing subsystem deals with preprocessing on the character images and the last subsystem proportionally splits the dataset into testing and training set. Character bracket phase includes testing and training on dataset. Deep Complication Neural Network- coach was run-on the Training set and the delicacy of the trained model is tested using Testing set of Devanagari Character Data set (3).

A neural network approach was introduced to perform high delicacy recognition system has been developed using MATLAB. The scrutinized document image taken as input and feed forward armature was used. The neural network structure of includes an input subcaste with each input be resized to inputs, two retired layers each with 33 neurons and an affair subcaste with 33 neurons. Neural network has been trained using given dataset. After the network training, the Recognition system was tested using several unknown dataset and the results attained are presented shows neural network training state (4) (5).

The new approach had been used with a combination for the bracket. Different bracket styles have their own advantages and limitations. So numerous times multiple classifiers are combined together to break a given bracket problem. Distinct classifiers trained on the same data can't only differ in their performances encyclopedically, but they also show strong original variations. Some neural network classifiers show different results with different initializations due to the randomness essential in the training procedure. We cancombine colorful networks rather of opting the stylish network and discarding the others, by taking advantage of all the attempts to learn from the data (6).

India is a country where numerous different languages and scripts are used. In Marathi language, substantially Marathi language is used. National language of India is Marathi and after Chinese and English, the third most spoken language of the world. Substantially Marathi is used in attestation in

Rajasthan, New Delhi, Madhya Pradesh, Uttar Pradesh, Himachal Pradesh, Chattisgarh, Uttarakand, Bihar and Haryana. Thus, Marathi language is substantially used in colorful documents like operation forms, bank cheques, envelops, answer wastes, road reservation forms etc and also numerous websites are hosted in devanagari decreasingly. There are numerous marketable systems available for reading and searching english scripts, but still Marathi language for similar are in development stage. Bansal etal. (2010) (8)This paper developed the segmentation of different irregular textbook words of Gurumukhi script. The segmentation of words containing disposed, irregular caption, broken, touching and lapped characters are bandied in this paper. Some new ways similar as counter tracing styles are developed with the help of vertical and perpendicular protrusions. Kumar etal. (2010) (9)The segmentation of the colorful scrutinized textbook image is bandied in this paper. The full image is known as a large window in this fashion. The large window is resolve into small windows as giving lines and once the lines are honored also fete a word that's was in a line and at the end character is honored. The variable-sized window conception is also developed in this paper. GargN. etal. (2011) (10) In the OCR system, the recognition rate is dropped due to the touching of the partial character along with full characters, the analysis of the actuality of partial character is a veritably complicated task. In this paper, a new algorithm of structural parcels of documents is proposed to member half characters of handwritten Marathi textbook. The results are concluded for both handwritten Marathi textbook and for published Marathi textbook. The proposed algorithm acquires delicacy in segmentation as83.02 with partial characters in the handwritten textbook and87.5 in the published textbook. Kumar and Singh (2011) (11)

Numerous tests were conducted on different documents, the results were attained with great delicacy. Some characters of lines in the lower zone were observed nearly rightly. The equals of the detected lines and words are used to get the character. The character segmentation process was distributed in two corridor (i) to acquire the segmented region R (ii) to corroborate that R has a meaningful symbol or not. If R is meaningful also it's accepted else rejected. Rhead etal. (2012) (12) this paper developed aspects of the applicable legislation and norms, after applying them on earth range plates.

## 3. CONCLUSION

This review is proposed to many regional languages throughout world have different writing styles which can be recognized with this system using proper algorithm and strategies. We have learning for recognition of Indian Marathi language characters. It has been found that recognition of Indian Marathi language character becomes difficult due to presence of odd characters or similarity in shapes for multiple characters. Scanned image is pre-processed to get a cleaned image and the characters are isolated into individual characters.

## REFERENCE

[1] Bansal V, Sinha R. M. K., "Integrating Knowledge Resources in Devanagari. Text recognition system", IEEE Trans. on Sys. Man & Cybernetics Part A: Sys. & Humans. Vol. 3, No. 4, pp. 500-505, July 2000. 669 | P a ge

[2] P. S. Deshpande, Latesh Malik, "Fine Classification & Recognition of Hand Written Devnagari Characters with Regular Expressions & Minimum Edit Distance Method", Journal of Computers, Vol. 3, No. 5, pp. 11-17, May 2008.

[3] Nafiz Arica, Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics— Part C: Applications and Reviews, Vo. 31, issue 2, May 2001.

[4] Agnihotri, Ved Prakash, "Offline Handwritten Devanagari Script Recognition", International Journal of Information Technology and Computer Science, 2012, Vol. 4, No. 8, pp. 37-42, 2012.

[5] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal, "Offline Recognition of Devanagari Script:

[6] A Survey", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Volume 41, Issue 6, November 2011.

[7] Mahesh Jangid, "Devanagari Isolated Character Recognition by using Statistical features", International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No 3, pp. 2400-2407, 2011.

[8] S.Kompalli,S. Setlur, and V. Govindaraju,"Devanagari OCR using a recognition driven segmentation framework and stochastic language models." Int. J. on Document Analysis and Recognition, Vol. 12, no. 2, pp. 123-138, 2009.

[9] V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, Setlur and Vemulapati, "Tools for enabling digital access to multi-lingual Indic documents", In Proc. 1st International Workshop on Document Image Analysis for Libraries, pp. 122-133, 2004.

[10] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 32, Issue 4, pp. 449–459,2002.

[11] U. Pal, P. P. Roy, N. Tripathy and J. Llados, —Multi-oriented Bangla and Devnagari text recognition", Pattern Recognition., Volume 43, Issue 12, pp. 4124–4136, 2010.

[12] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition", 3rd IEEE International conference on Industrial and InformationSystems, pp. 1-6, 2008.

[13] M. Vaidya and Y. Joshi, "Marathi Numeral Recognition using statistical distribution features", In: Proc. IEEE conference on Information processing, pp.586-591, 2015.

[14] M. Vaidya and Y. Joshi, "Handwritten Numeral Identification System Using Pixel Level Distribution

[15] Features", In: Proc. 2nd International Conference on Information and Communication Technology for Intelligent Systems, Vol. 2, pp. 307-315, Springer, Cham,2017.

[16] C. C. Tappert, C. Y. Suen, and T. Wakahara, ‘‘The state of the art in online handwriting recognition,’’ IEEE

.

[17] M. Kumar, S. R. Jindal, M. K. Jindal, and G. S. Lehal, ‘‘Improved recognition results of medieval handwritten Gurmukhi manuscripts using boosting and bagging methodologies,’’ Neural Process. Lett., vol. 50,pp. 43–56, Sep. 2018.

[18] M. A. Radwan, M. I. Khalil, and H. M. Abbas, ‘‘Neural networks pipeline for offline machine printed Arabic OCR,’’ Neural Process. Lett., vol. 48,no. 2, pp. 769–787, Oct. 2018.

[19] P. Thompson, R. T. Batista-Navarro, G. Kontonatsios, J. Carter, E. Toon, J. McNaught, C. Timmermann, M. Worboys, and S. Ananiadou, ‘‘Text mining the history of medicine,’’ PLoS ONE, vol. 11, no. 1, pp. 1–33, Jan. 2016.

[20] K. D. Ashley and W. Bridewell, ‘‘Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings,’’ Artif. Intell. Law, vol. 18, no. 4, pp. 311–320,

[21] R. Zanibbi and D. Blostein, ‘‘Recognition and retrieval of mathematical expressions,’’ Int. J. Document Anal. Recognit., vol. 15, no. 4,pp. 331–357, Dec. 2012,

[22] I. K. Pathan, A. A. Ali, and R. J. Ramteke, ‘‘Recognition of offline handwritten isolated Urdu character,’’ Adv. Comput. Res., vol. 4, no. 1, pp. 117–121, 2012.