Journal of Engineering Sciences
Vol 11, Issue 6,June/2020
ISSN NO:0377-9254

Comparison of Feature Extraction methods for the study of Cursive Handwritten Marathi Compound Character Recognition

R. N. Khobragade[1], Nitin A. Koli[2], Vrushali T. Lanjewar[3]

[1, 2, 3] Sant Gadge Baba Amravati University, Amravati, India

[1]rnkhobragade@gmail.com, [2]nitinkoli@sgbau.ac.in,

[3]vrushali.lanjewar@gmail.com

***Abstract*— The objective of this paper is to make a comparative study of feature extraction methods in the new era to focus on problems in Handwritten Marathi Characters. We discussed some issues that arise due to diversity in writing handwritten characters of Marathi. The quality of images must be improved to learn the features by giving a variety of manuscript letters better classification results. We prepared formulation of problem definition Marathi compound characters and manuscript. The cloud of line distribution, distribution feature, concavity features, shearlet transform analyzed to extract features that can be evaluated to give improved results.**

***Keywords*—** Handwritten Characters, Marathi, Concavity feature

I. INTRODUCTION

Today the study and description of handwritten Marathi heritage are one of the imperative tasks of Marathi studies. It should be noted that the codicological methods used in the study of manuscripts in Indian and some Oriental languages are not yet fully applied in the study of their Marathi counterparts. Despite the fact that the tradition of studying Marathi manuscripts as a whole has a long history, the methods for describing and studying manuscripts have not changed much. Even a preliminary acquaintance with the works devoted to the study of the Marathi manuscript heritage shows that the principles of formal, as well as textual and, in a narrower sense of the word, linguistic analysis of the manuscript vary slightly and are based on a clearly verified and time-tested paradigm for describing manuscripts. Such a paradigm provides for their chronologization, lexical and terminological analysis, correlation with one of the manuscript schools, the study of the historical or linguistic context associated with their creation or correspondence.

Typically, studies of this kind consider the features of the language of the manuscript, handwriting, ink properties, type of binding, subject matter, and paratext elements-colophons (information about the author, place and time of correspondence, name of the copyist / customer at the end of the manuscript) and subscription (the same at the beginning of the manuscript), incipits and expressions (formulas for the beginning and end of the text), the presence of seals. Moreover, the analysis of Marathi manuscripts is based largely on the subjective assessments of the researcher, which, in turn, depend on his qualifications, experience and knowledge (Nag et al.).

The hypothesis includes the following tasks:

1. Study of the grammatical and graphological features of Marathi texts in the light of the recognition problem.
2. Development of a methodology for preparing a text image to highlight informative features.
3. Development of a method for isolating the skeleton, providing the required quality.
4. Development of methods for highlighting lines, characters, symbols, words and sub-words. 5. Development of a text model that displays the topology, geometric parameters and relationships of the elements of the Marathi text.
6. Development of a method for recognizing handwritten and typewritten texts of the Marathi language.
7. A comparative analysis of the reliability and effectiveness of the proposed algorithms.

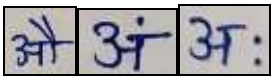
II. PROBLEMS IN MARATHI CHARACTER

While writing 'ढ' and ' द' character may be confused in OCR system if it is filled with noise or due to writing style. In 'ष', the diagonal bar should be bold, otherwise it can be confused with character 'प'; if the bar is

removed or broken while printing or scanning in binarization process. Compound containing two half characters such as ‘न्स्पे’, ‘त्स्य’, ‘त्त्व’ ‘ब्ल्यू’, ‘म्प्यु’ are difficult to recognize and need more preprocessing further for classification.

A. **Modifiers in character**

Printed: अ आ अः औ अँ अं

Handwritten: 

Considering ‘अ’ is a basic (base) character अँ, अं, अः containing angular with dot, single dot, double dot these modifiers can be removed while processing and removal of noise. Often same characters can be written differently by the same writer depending upon speed of writing.

B. **Detection of Joining Character:**

The two or more characters join together called as compound character. To detect join characters firstly detect half characters in it. There are two types of compound characters as follows:

Same character join in sequence:

न्+न =न्न, म्+म=म्म ,प्+प =प्प, ल्+ल=ल्ल

Different character joined at left side:

स्+ त= स्त, ल्+फ= ल्फ, स्+क=स्क, द्+ य=द्य, ष्+ ठ=ष्ठ, ष्+ ट=ष्ट, च्+छा=च्छा

Handwritten characters

Figure 1. Compound Characters in Marathi

The above Marathi handwritten characters are taken from different documents, letters by anonymous writers. These characters depict the diversity in same and different combination of consonant characters.

C. **Detection of half character:** In Compound character, half character (fig. 2) is joined to traditional vowel or consonant character. So, to detect compound character half-character needs to be identified.

Figure 2. Half Characters

Sometimes it is not possible to identify half-characters after preprocessing because key features and pixel information is lost during the binarization.

D. **Extra ligaments in some parts of the character**:

The features cannot be determined due to extra ligaments(fig. 3) in some parts of the character. These ligaments may be due to pen thickness or due to segmentation of joint characters.

Figure 3. extra ligaments with characters

The some of the characters showing some extra ligaments with the characters ‘श्री, ऊ ,क , अ ,बा, चं ’ etc. formed due to different writing style and speed of writing.

E. **Mixed or Overlapped Characters**

It is difficult to obtain compound character like ‘ड्डु’, ‘द्ध’, ‘क्ड’,‘न्या’, ‘ट्ट’ in fig. 4, when two characters in words are closely spaced. Due to non- uniform spacing two characters is overlapped or mixed.

Figure 4. Overlapped Characters

The characters are mixed or overlapped on each other due to less spacing between them hence cannot be recognized properly.

Sample of Half Characters:

Half form of Consonant with Vertical Bar and combination of half Consonant and Consonant is depicted in fig. 5. There are different combined/compound characters are used in Marathi. The sample of the compound character form is as shown in fig. 6.

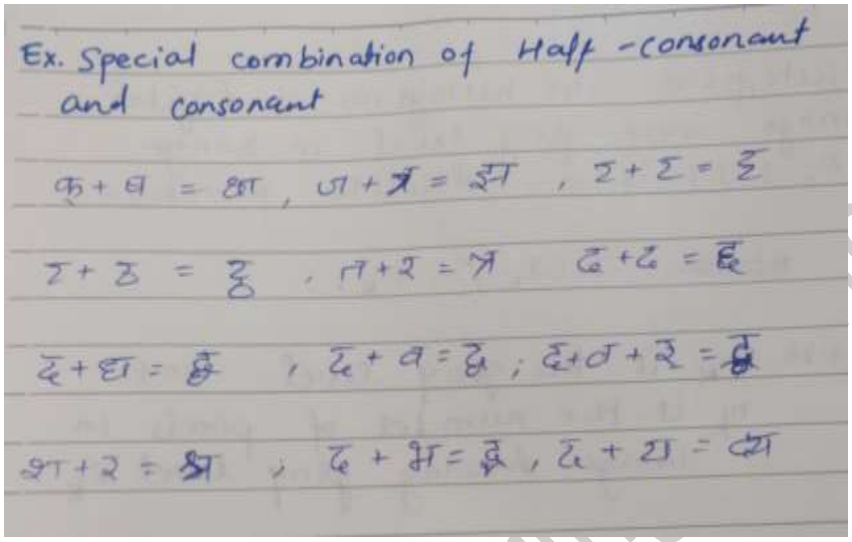


Figure 5. Formation of character in Marathi

Some of the compound characters are depicted in Fig.5 and Fig.6

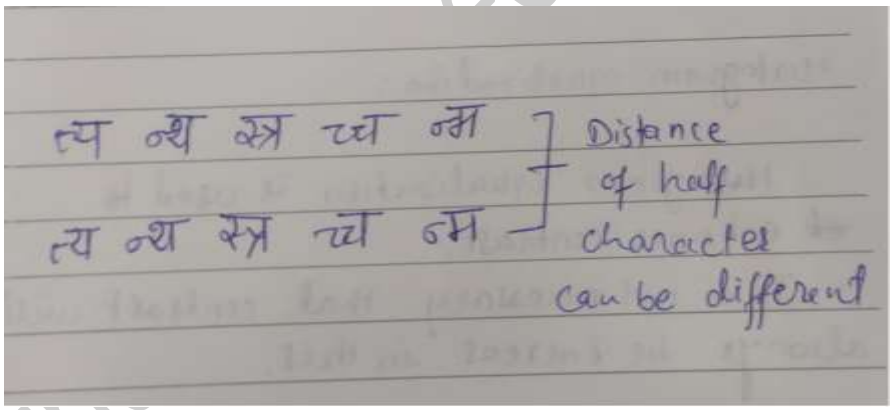


Figure 6. Style of writing depicting same characters.

F. **Non-uniform distance of haft-character in compound Character**

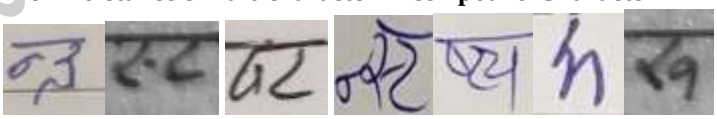


Figure 7. Non-uniform distance

Due to non-uniform spacing between half-character and base character(fig.7), it is difficult to identify the whole compound character after preprocessing.  The compound character has broken ligament which then not useful in training.

As the writing style differs the characters depicts different features as the shape of character changes as shown in above Fig. 6. One writer can write the same character showing various shapes, orientation and strokes of a character may change due to writing speed. For example, last character image of 'रू' in fig.7 written by two different writers shows how it differs from each other.

III. RELATED WORK

Theoretical studies were performed using set theory, mathematical logic, and pattern recognition theory. Experimental studies were performed using the developed software system and some well-known products in real Marathi texts, Chinese characters, as well as real and artificial drawings (Mukherjee *et al.* 2018).

The use of information and communication technologies in creating electronic handwritten catalogs entails the emergence of new terminology to describe this process. Recently, the term "digitalization", or "digitization" of manuscripts, which is interpreted as digital computer processing of manuscripts, has become widespread. The digitalization of the Marathi manuscript heritage can make significant adjustments to the process of describing manuscripts and help in the development of electronic on-line catalogs and, to a certain extent, optical character recognition.

The most significant scientific results of the study include: 1. The method of thinning sections of the object, characterized in that in order to improve the quality of skeletal lines, the concept of "indefinite" color is introduced into the binary image and, in order to increase the reliability of text recognition, calculates the width parameter of the source lines. 2. A text model in the form of a flat graph that displays the topology, geometric parameters, and relationships of symbols, signs, specific elements of the language syntax, and the methodology for the formation (extracted from the image) of this model.

Creating an effective program for recognizing Marathi printed, not to mention handwritten, text is a more difficult problem than developing a similar product designed to work with Latin-based texts (Karki *et al.*2018). With optical recognition, it should be borne in mind that the type of writing of handwritten documents varies depending on the traditions of a particular handwritten school, and may also include individual characteristics of the authors' handwriting. The recognition problem becomes even more complicated when you have to deal with additional “noises”, such as comments from scribes, writing defects, damage to the material used, as well as gaps and omissions, as well as later additions to the original text. All this makes the adequate identification of Marathi written texts extremely difficult (Gagaoua *et al.* 2017). Translation of Marathi handwritten text into a digital format with the possibility of its subsequent processing and the implementation of the search function in the content array is done mainly in manual mode, and, due to the high complexity, this method of digitization is very rare. In this context, the term “digitization” in relation to existing electronic collections of manuscripts has a limited number of examples (Rizky *et al.*2018).

The effectiveness of existing recognition products for the Marathi language largely depends on the structure of the analyzed materials, they operate with a minimum number of errors only if they work with the "ideal" text (typed using one of the most common fonts, without voices, etc.). Obviously, the problem of character recognition in Marathi text is more complicated than in English texts. This circumstance is largely due to problems of both a linguistic and technological nature. The difficulty of recognizing Marathi graphics is also due to the large number of derivatives, the "fused" nature of the letter, allowing for different lengths of connecting lines, the possibility of realizing points aside from the letter, the presence of ligatures, and fused spelling of a number of prepositions and particles. The urgent need to use methods of computer analysis of manuscripts is also due to such features of the handwritten text as the presence of filigree, or even palm script and various types of paper used in it. So, with digital processing, you can determine the exact number of lines in the text, at what angle the text is placed between the grid lines (shorter) on a particular page, as well as in the entire manuscript as a whole.

In compound characters, the half-character is attached upper side or lower side of a base character mentioned by (Khobragade *et al.* 2020). Handwritten characters contains unique illumination, size, orientation and occlusion so they need to be carefully processed. The author also suggest that proper segmentation should be done to remove complexity of compound character.

IV. ANALYSIS OF FEATURES

In this section, we discussed the different feature extraction methods by various authors that plays a major role in character recognition. M. Gagaoua *et al.*, discussed and analyzed in detailed four features including Distribution, Directional, structural and concavity features.

Distribution features:

Marty and Bunke *et al.*, proposed the distribution features, then researchers use these features to classify handwritten text using HMM. The binary pixel distribution of the image is shown in eqn. 1. To determine the properties, nine geometric dimensions are calculated from the pixels at the beginning of each column of the image. The three features are the number of black pixels, the center of gravity, and the second numbered moment. To extract many features from connected components (CC), first CC images are divided into four horizontal blocks that allow to capture the spatial distribution of information and then use a window that slides from left to right.

$$P(x,y)=\begin{cases} 1 \text{ if pixel black} \\ 0 \text{ if pixel white} \end{cases} \quad \text{.....(1)}$$

The extra feature gives more information about the composition. The fourth and fifth features give the position of the upper and lower profiles, the sixth, seventh, eighth and ninth give the position of the closest visible pixels in the upper and lower profiles, the proportions and inclination of the front of the foundation changes. Upper and lower profiles in the previous section, which gives dynamic data. To separate the groups of features from the connected component, the pictures of the first CCs are divided into four flat squares, which allow to capture the space vehicle of the data, at which time a sliding window from left to right is used.

Directional features:

The slope being separated for each phase of the connected component and the image is partitioned into smaller blocks. The features are registered with the help of two 3x3 Sobel operators with attached images. The inclination of the center pixel is determined by the details of its eight closest neighbors. The gradient direction is used to compute the feature vectors of gradient that includes feature maps. The direction is divided into 12-non-overlapping region. In all sampling regions, approximately 12 gradient features are for each image section, and a histogram of the gradient directions is evaluated. It compares every gradient direction in that neighborhood. The counters are attached to represent the main features of the connected components.

Structural features

Favata *et al.*, proposed the structural features and Mahmud, S.A. *et al.*, used structural features for an Arabic manuscript. It obtain mini strokes of the image. A set of 12 rules applied to each pixel. Each rule checked for a specific pattern of neighboring pixels to allow gradient ranges. These features represent structural feature vectors. The feature vector has 12 types of structural features corresponding to the rules.

Concavity features

Xiang Dong *et al.*, defined the concavity feature. This feature is used to calculate the average distance scanning pixel by pixel from the cell border to the first met stroke (Fig.8).

The four concavity features are shown in fig. 8. Let ri, li, ui and di respectively, the distance from the right side, left side, upper and bottom cell border to the first meet stroke of character along left side, right side, upside and down directions.

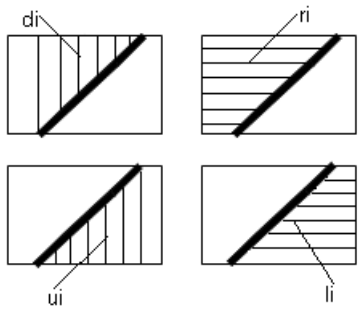


Figure 8. The four types of concavity features (M.Gagaoua et al.)

Online pen trajectory features

Mukherjee *et al.* mentioned online features that include curvature at each point on the pen trajectory of the character. These online features are computed based on the above window around a point *Pt* with coordinates $x(t)$, $y(t)$. Curvature at a point *Pt* of the trajectory is computed based on a window of seven consecutive points of the trajectory with *Pt* as the center of the window. The curvature (κt)at *Pt* is calculated by considering a discrete curve through the $2m+1$ and the formula for the same is given by eqn. 2

$$\kappa = \frac{|f''(P_j)|}{\left(1+[f'(P_j)]^2\right)^{\frac{3}{2}}} \ldots\ldots\ldots\ldots(2)$$

The Vicinity Aspect (*Va*) is computed by using the formula,

$$Va = \frac{\Delta y(t)-\Delta x(t)}{\Delta y(t)+\Delta x(t)} \ldots\ldots\ldots\ldots(3);$$ where $\Delta y(t) = y(End) - y(0)$, $\Delta x(t) = x(End) - x(0)$.

The length of the trajectory is sum of Euclidean distance between each two consecutive points in the vicinity divided by maximum of $\Delta y(t)$ and $\Delta x(t)$.
The average square distance of each surrounding point from the straight line connecting *P0* and *Pend* shown in Fig.9. The X coordinate of *Pt* after subtracting the average of *x*-values of all points in its vicinity. All the online features calculated on the basis of its vicinity is given above.

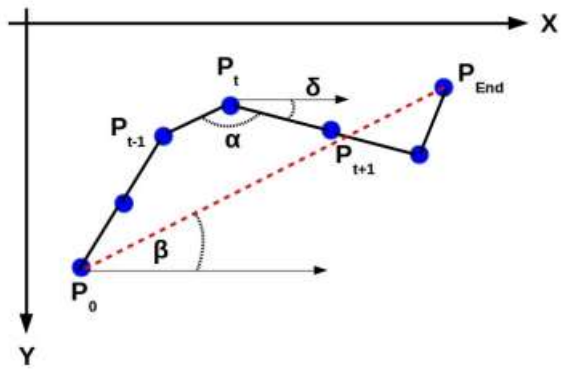


Figure 9. Sample of computation of Online feature (Mukherjee *et al.*)

Pretrained CNN as features

Karki *et al.* used pretrained CNN as features for reconstructing the characters, removing noise via pixel level denoiser to map the responses. Each pixel is considered as a single data point, and the total number of images is reconstructed according to the classification of each pixel on that image. Each input image pixel corresponds to the pixel of each feature map. The values of the pixels in the corresponding map responses are used as features that when associated together form hypercolumns. ImageNet, which is used to pretrain CNN, has 1000 object classes and over a million images. It is used to train objects, animals, scenes and something geometric shape.

Cloud of line distribution (COLD) features

Sauradip Nag *et al.* explores Cloud of line distribution (COLD) features for writer identification. COLD changes shape according to the behavior of Line segment of character elements. If the character is too cursive then, one should expect more small line segment. When the lines are short, the COL distribution in the middle appears denser. If the lines are long, they can be distributed to anyone. If there is a regular curvature of the elements in the character, a central COLD distribution at a certain angle is expected. If the character elements have irregular curvature, the cold distribution is scattered at arbitrary angles. To calculate, the straightness and curvature of the line segments in the polar coordinated domain, to find the angle, θ for the line segments of the contour using x and y coordinates as defined in eqn. (4). The length, *r* of line segments is determined as defined in eqn. (5).

$$\theta = \tan^{-1}\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right) \ldots\ldots\ldots\ldots(4)$$

$r = abs\sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2}$……..(5)

(xi, yi) and $(xi+1, yi+1)$ be the coordinates of a dominant pair. A line segment shown using θ and r as a point (θ, r) in polar domain and draw points for all the line segments in polar domain results in a distribution. Polar coordinates produced with the use of distance between dominant points of the contours of edge components which gives COLD distribution in polar domain.

Shearlet Transform

Rizky *et al.* performs shearlet transform to extract features of Mbojo character. A shearlet framework gives the ability to create a directional depiction of images with anisotropic features, Guo *et al.* The complex shearlet framework calculations are done with different inputs, for example, columns size of image, row size of image, wavelet, Gaussian scales per octave, shear level, alpha, octaves, scales etc.

- Steps to obtain features using shearlet transform framework:

1. In shearlet phase, using multiplication of wavelet frequency and Gaussian frequency obtain an output of the matrix .
2. Execute 2-D Inverse Fast Fourier transform process.
3. Carry out Hilbert process with several conditions.
   1st Condition: shearlet=Hilbert(-shearlet) ;
   shearlet=circsift(shearlet,[1 0])
   2nd Condition : shearlet=((Hilbert(-shearlet)’)’);
   shearlet=circsift(shearlet,[0 1])
   3rd Condition : shearlet=((Hilbert(-shearlet)’)’);
4. Execute the 2-D Fast Fourier transform.
5. Save the shearlet value on a 3D matrix.

- Steps to obtain edge detection:

1. In edge detection, insert the image, complex shearlet system, minimal contrast, and offset.
2. Gained coefficient by multiplying the image with complex shearlet system.
3. Execute element shift on coefficient with several conditions.
   1st Condition = sift(-coefficient(:,:,ori,:),[-1,0,0,0])
   2nd Condition = sift(-coefficient(:,:,ori,:),[0,1,0,0])
   3rd Condition = sift(-coefficient(:,:,ori,:),[0,-1,0,0])
4. Then, calculate coefficient image (CI) and coefficient real (CR).
5. Compute CiPivot and MaxPivot
6. Compute pivotoris(pivot orientation) and pivotscales
7. Get the edge orientation

Shearlets provide a practically sparse estimation of the anisotropic features of multivitate data. The Shearlet Transform is different from Wavelet Transform as it can detect directionality.

V. ANALYSIS AND DISCUSSION

The analysis of respective methods are depicted in Table 1., which gives the different features taken from various classifier with the improvement in accuracy of assorted networks. Karki *et al.*, provides a pixel level classifier that used to improve the effectiveness of probabilistic quadtrees to extract the character pixels and remove noise from handwritten character images. The pixel level denoiser (a deep belief network) use the map responses obtained from a pre-trained CNN as features to reconstruct the characters removing noise. The experiment shows that proposed approach to reconstruct and classify a noisy version of handwritten Bangla Numeral and Basic Character datasets are effective. Quadtrees used to represent learning from sparse features. The efficiency of probabilistic quad-trees, improved by using a pixel level classifier to reconstruct noisy handwritten character images by segmenting out the noisy pixels. The pixel level denoiser (a deep belief network) use the map responses obtained from a retrained CNN trained on Imagenet as features for reconstructing the characters eliminating noise.

M. Rizky *et al.*, shows in the pre-processing stage that comparison of recognition process completed by adding centering region method. A recognition system that adds a centering region method can increase the system's accuracy. By using 50 data tested, the system is able to recognize 45 data and failed to recognize 5 data. The accuracy of system is 90% based on the recognition result.

Table 1: Comparison of different features for different scripts

Sr No	Author	Script	Text type	Features	Method	Database	Sample Size	Classifier	Accuracy
1.	Gagaoua, M. et al. (2017)	Arabic	Handwritten	Distribution, Directional, structural and concavity features	Hidden Markov models	Iben Sina database	60 pages, 25,000 Arabic subword shapes written in the Naskh style.	HMM	Gradient and concavity features give 45.36% and best Recognition rate 63.88%
2.	Manohar Karki et al. (2018)	Bangla	Handwritten Character	Pretrained CNN as features, averages of pixel values of decomposed blocks, awgn contrast, motion	Quadtree decomposi-tion	N-MNIST and noisy Bangla Basic Character,	Not mentioned	Deep belief network	Reduced contrast images had the worst recognition rates
3.	Mahathir Rizky (2018)	Mbojo	Printed (image of mbojo words)	Shearlet Transform for Isotropic features	Centering region method, Shearlet Transform	bimambojo.otf font	150-word images; size of 50x50 pixels.	Multi-class support vector machine (MC-SVM)	90%
4.	Mukherjee, P. S., (2018)	Bangla and Devanagari	Handwritten character sequence	Vicinity, Curvature, trajectory	Hybrid deep neural network architecture CNN, BLSTM,	Bangla: ICBOHR-W2. Devanagari: ICDOHR	ICBOHR-W2: Words :771, Character: 58, ICDOHR	Bi-Directional Recurrent Neural Network with LSTM	LER in % **CNN-BLSTM-CTC** ICBOHR-W2:86.73 ICDOHR:

					CTC		-Words :1959 Character:79	nodes (BLSTM), Connectionist temporal classification (CTC),	78.31 **BLSTM-CTC** ICBOHR-W2:70.29 ICDOHR: 61.84
5.	Nag, Sauradip et al. (2018)	English	Handwritten	Cloud of Line Distribution (COLD) features	Principal axis with Principal Component Analysis (PCA),	English handwriting images from ASIAN countries (QUWI)	500 text line images	SVM classifier	Overall rate is 75%

Mukherjee, P. S. *et al.*, shows that the proposed hybrid architecture recognizes online handwriting more efficiently than a BLSTM network alone. It comprises first and second order discrete derivatives which evaluate geometric properties of discrete curves along the pen trajectory of the input online handwritten word sample. The segmentation free recognition of unconstrained cursive online handwriting of Devanagari and Bangla is studied. The proposed feature vector is obtained at each point on the trajectory of the pre-processed word sample based on a window (called vicinity) centered at the point. An efficient algorithm developed to obtain the core region of a handwritten word sample and compute a robust feature set which includes measures based on the knowledge of the core region.

Nag *et al.* presents a method to identify ethnicity based on cloud of line distribution (COLD) feature of handwriting. The tangent angle for contour is to find out in each row and then mean the intensity values of each row in text line is segmented. The tangent angle and direction of the baseline used for segmented text line to eliminate rule lines in the image. For contour of edge part, the polygonal approximation is used to find dominant point. The proposed method connects the nearest dominant point of each dominant. The point which result in the division of dominant point pair. For each segment the proposed method estimates the angle and points which point to dense point in polar domain. For all line method produces thicker point which results in COLD distribution.

VI. CONCLUSION

We have analyzed the feature extraction methods used up till now. In Marathi handwriting, the compound character comprising half character and base character in which distance of half-characters depend on writing speed of writer that can change the features of the same character. Thus, we modify the quality of images to get better results and demonstrated results for the different feature extraction methods.

VII. FUTURE WORK

This work provides the basis for further research Marathi Handwritten Character analysis. Still a lot of possible combination of feature information methods In addition to testing, it requires a large dataset which provide more samples for training so that the system is enabled to identify more characters.

ACKNOWLEDGEMENT

The authors acknowledge the financial assistance of the Department of Science & Technology-Science & Engineering Research Board (DST-SERB) New Delhi, Government of India under the grant EEQ/2017/000102.

The research held at P.G. Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra.

REFERENCES

Gagaoua, Meriem, et al. "Distribution, Directional, structural and concavity features for historical Marathi handwritten recognition: a comparative study." *Proceedings of the International Conference on Computing for Engineering and Sciences*. 2017.

Mukherjee, Partha Sarathi, Ujjwal Bhattacharya, and Swapan Kumar Parui. "An Efficient Feature Vector for Segmentation-Free Recognition of Online Cursive Handwriting Based on a Hybrid Deep Neural Network." *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018.

Nag, Sauradip, et al. "New COLD Feature Based Handwriting Analysis for Enthnicity/Nationality Identification." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018.

Karki, Manohar, et al. "Pixel-level reconstruction and classification for noisy handwritten bangla characters." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018.

Rizky, Mahathir, Ingrid Nurtanio, and Intan Sari Areni. "Mbojo Character Recognition Using Shearlet Transform and Support Vector Machine." *2018 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2018.

Khobragade R.N., Koli N.A., Lanjewar V.T. (2020) Challenges in Recognition of Online and Off-line Compound Handwritten Characters: A Review. In: Zhang YD., Mandal J., So-In C., Thakur N. (eds) *Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies*, vol 165. Springer, Singapore.

Guo, Kanghui, Gitta Kutyniok, and Demetrio Labate. "Sparse multidimensional representations using anisotropic dilation and shear operators." Wavelets and Splines (Athens, GA, 2005), G. Chen and MJ Lai, eds., Nashboro Press, Nashville, TN (2006): 189–201.

U., V.M. and H. Bunke, Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system,. International Journal of Pattern Recognition and Artificial Intelligence, 2001. 15(01): p. 65-90.

Mahmoud, S.A. and M.A. Sameh Recognition of off-line handwritten arabic (indian) numerals using multi-scale features and support vector machines vs. hidden markov models. The Arabian Journal for Science and Engineering, 2009. 34(2B): p. 429-444