A Survey of Marathi Character Recognition

Rajivkumar Mente
Department of Computer Science,
Solapur University, Solapur, Maharashtra, India

Abstract: Marathi is the main official language in Maharashtra. In recent years, handwritten Marathi character recognition has captured a lot of attention as it has broad application in areas like passport, railways, postal address reading etc. Recognition of handwritten characters would decrease the task of data entry and save the time in case of Form filling, Banking and Postal Automation etc. This paper presents the various steps involved in recognition of handwritten characters and review of methods reported by various researchers for handwritten character recognition.

Keywords: Marathi Characters, OCR, Feature Extraction, Invariant Moments **,** Devnagari

I. INTRODUCTION

Optical character recognition (OCR) means the recognition of printed or handwritten manuscript by a computer. This involves forming image of the text by scanning, analyzing the scanned-in image, and then translating the character image into character codes which is useful for data processing. From the past few decades recognition of handwritten characters is a frontier research area. There is a large requirement of optical character recognition on handwritten documents. Due to the large character set in Indian languages and the presence of vowel modifiers and compound characters, there is no complete hand written character recognition system available in Indian scenario. Some reports have appeared for isolated handwritten characters and numerals recognition of a few Indian languages [1]. Few challenges in developing a system for handwritten character recognition are

- Variation in the size and shape of the characters which depends on the writer
- The physical and mental condition of the writer
- Width of the pen, the color of ink used in the pen
- The acquisition device used and many other

Recognition of handwritten Marathi characters is more difficult due to the complex structure and shape of characters and presence of modifiers.

II. STEPS INVOLVED IN HANDWRITTEN CHARACTER RECOGNITION

The script used for Marathi language the Devnagari script. Marathi language contains 13 vowels and 36 consonants. A Marathi language script is put in writing from left side to right. Vowels in Marathi language are the soul of speech and they are usually written in abbreviated symbols.

Fig.1, fig.2, and fig.3 show the vowels, consonants and some examples of combination of vowels with consonants.

Fig. 1. Vowels in Marathi Language 1

Fig. 3. Examples of Vowels combined with Consonants
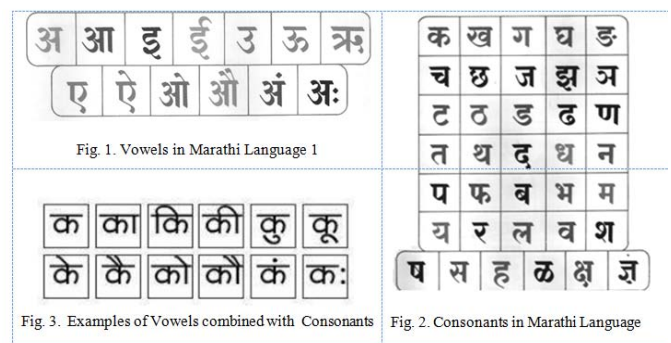
Fig. 2. Consonants in Marathi Language

Fig.4 shows the basic components of the character recognition system.
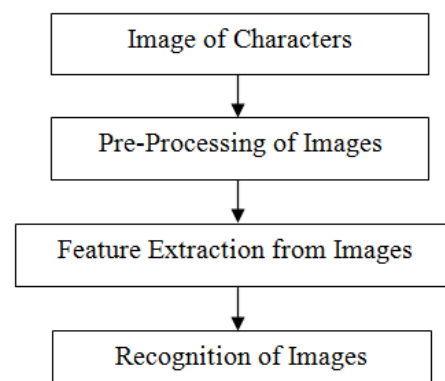

Fig.4. Components of Character Recognition System

Acquisition devices like flatbed scanner are used to acquire the images of characters. The data cleaning tasks which may include reduction of noise, normalization etc. is carried by the pre-processing component. The representatives features require to classify the characters are extracted using the feature extraction component. It also reduces the redundancies in data. Features extracted in the previous stage are used by the recognition component for making the decision i.e. recognition of character.

III. LITERATURE SURVEY

There are many script and Languages in the world. Many researchers have done work on the languages like English, Latin, Chinese, Japanese, Arabic, and Devanagari. Eleven scripts and several languages are used in India. The work done towards recognition of handwritten Indian scripts is less as compared to other languages. Recognition of handwritten characters is important because of its applicability to a number of problems, like postal code recognition and information extraction from fields of different forms. Many statistical techniques and models techniques has come forward by developments in the area of pattern recognition such as recognition of characters [2, 3, 4, 5], finger print recognition, face recognition etc.

Mitchell [6] and Zafar et al [7] extracted a variety of features such as primitives, profiles etc. Then these measured features are used to train the Multi Layer (ML) algorithm. This trained ML algorithm is used to map the input features onto a class among certain predefined classes. The classifier is then used to determine the class of unknown testing samples.

Character recognition task has been attempted by Pankaj et al [8] through many different approaches like template matching, statistical techniques like HMM, NN, Quadratic Discriminant function (QDF) etc. Template matching works effectively for standard font recognition. Its performance found poor when applied to handwritten characters as the dataset grows in size.

A method to recognize handwritten Marathi characters based on their feature extraction and adaptive smoothing technique is proposed by Archana et al [9]. The recognition accuracy of 98% is achieved provided the number of training samples images per standard Marathi images should be the maximum.

Compound characters of Marathi language are derived from Devanagri. A method by Shelke et al [10] 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 [11]. 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 [12] 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%.

The recognition Devnagari script is popular and open area of research due to its complex nature. Based on positional properties of each pixel in the character Bombade et al [13] proposed method for feature extraction. They used Devnagari and Kannada characters database and achieved 82.89% and 85.62% of accuracy respectively.

Wavelet-based kernel recognition for handwritten Marathi compound characters is proposed and given to classifier by Apte et al [14]. Wavelet kernels are obtained by using wavelet transforms. For 16 * 16 resize factor the accuracy obtained is 95.89% and for 32 * 32 resize factor the accuracy obtained is 96%.

A fuzzy technique for the recognition of handwritten Hindi numerals and characters is proposed by Hanmandlu et al [15] and obtained 92.67% and 90.65% recognition accuracy for numerals and characters respectively.

For the offline Devanagri handwritten character recognition Bikash Shaw et al [16] proposed a Hidden Markov Model (HMM). A continuous density HMM is constructed for each word. The conditional probability is calculated for each word. The class having highest probability is selected and accuracy obtained is 80.2%.

Holambe et al [17] proposed a nearest-neighbor classifier for devanagri character recognition. They extracted gradient and curvatures based features. These features are used for classification. The classifiers NN, kNN and Euclidean distance Neighbor based Kohonen are used for the analysis of results. Maximum accuracy is obtained is 94.98%.

The historical documents available in libraries and in many National Archives need to be converted in to digital form. It is important to convert the digital document into text lines. In 2006 Laurence Likforman – Sulem et al [18] presented a survey on existing methods and dedicated to documents of historical interest.

The segmentation of handwritten words into characters is a critical task because of complexity of structural features and varieties in writing styles. Due to this without segmentation these touching characters [19], it is difficult to recognize the individual characters.

Ajmire et al [20] proposed a method for recognition of handwritten Marathi vowel characters using seven invariant central moments of the image and the average success rate was compatible.

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 [21]. The SVM classifier is used and the overall performance obtained is 93.87%.

U. Pal et al [22] 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 [23] 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.

IV. CONCLUSION

In this paper different methods of character recognition proposed by the researchers were discussed. A lot of research is also required to handle the challenges in handwritten Marathi characters. The detailed discussion in this paper will be beneficial to understand various concepts involved and boost further advances in the area. The references enlisted in this paper will help to understand in detail the described approaches.

V. REFERENCES

[1] Ratnashil N Khobragade, Dr Nitin A Koli and Mahendra S Makesar, "Zoning Density Based Feature Extraction For Recognition of Marathi Handwritten Character", International Research Journal of Engineering and Technology, IRJET Volume 02 Issue 04, July-2015.

[2] 2. U. Pal, T. Wakabayashi and F. Kimura. "Comparative study of Devanagari Handwritten character recognition using different Features and Classifier", 978-0-7695-3725-2, IEEE, ICDAR, 2009.

[3] Chavan S V , Kale K.V, Kazi M M and Rode Y S "Recognition of Handwritten Devanagari Compound Character A Moment Feature Based approach, International Journal of Machinne Intelligence, Vol.5, Issue 1,pp 421-425,2013.

[4] V. S. Tapkir, S, D. Shelke "OCR For handwritten Marathi Script",International Journal of Scientific & Engineering Research,Vol.3,Issue 8, 2012.

[5] Priyanka Kamboj and Versha Rani "A Briff Study of Various Noise Model and Filtering Techniques" ,Journal of Global Research in Computer Science, Vol.4, Issue. 4,2013.

[6] T.M. Mitchell, "Machine Learning", McGraw-Hill Education (ISE Editions), December 1997.

[7] M. Zafar, D. Mohamad, M.M. Anwar, "Recognition of Online Isolated Handwritten Characters by Back propagation Neural Nets Using Sub-Character Primitive Features", IEEE Multitopic Conference (INMIC), 2006, pp. 157 –162

[8] J.R. Prasad, U.V. Kulkarni, R.S. Prasad, "Offline handwritten character recognition of Guajarati script using pattern matching", IEEE 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, 2009, pp. 611-615.

[9] Mrs. Archana P. Jane and Mukesh A. Pund, "Recognition of Similar Shaped Handwritten Marathi Characters Using Artificial Neural Network", Global Journal of Computer Science and Technology Neural and Artificial Intelligence, Vol.2, Issue 11, 2012.

[10] 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.

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

[12] 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.

[13] Swapnil A. Vaidya, Balaji R. Bombade," A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction", International Journal of Computer Science and Mobile Computing, IJCSMC, Volume 2, Issue.6, pg.179 –186, June 2013.

[14] Shelke,Apte,"Multistage handwritten marathi compound character recognition using neural networks"Journals of pattern recognition research 2, 253-268, 2011.

[15] M.Hanmandalu and O.V. Ramana Murthy,"Fuzzy model based recognition for handwritten hindi numerals",International conference on recognition , pp.490-496, 2005.

[16] Bikash Shaw, Swapan Kumar Parui, Malayappan Shridhar, "Offline Handwritten Devanagari Word Recognition: A Holistic Approach Based on Directional Chain Code Feature and HMM,"pp. 203-208, International Conference on Information Technology, 2008.

[17] Holambe, Thool,"Comparative Study of devanagri handwritten & printed character & Numerals recognition using nearest neighbour classifiers ", ninth International Conference on ICT and knowledge Engineering 2011.

[18] Laurence Likforman - Sulem, Abderrazak Zahour, Bruno Taconet, "Text Line Segmentation of Historical Documents: a Survey", Special Issue on Analysis of Historical Documents, International Journal on Document Analysis and Recognition, Springer, 2006.

[19] Vikas J Dongre, Vijay H Mankar, "Segmentation of Printed Devnagari Documents", DC Wyld et al. (Eds) ACIT Y 2011, CCIS 198, pp. 211–218, 2011. © Springer-Verlag Berlin Heidelberg 2011.

[20] Ajmire P. E., Warkhede S. E., "Handwritten Marathi Characters (vowel) Recognition", Advances in information Mining, Vol. 2Issue-2, pp. 11-13, 2010.

[21] 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.

[22] 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 , Qúebec, Canada, Aug. 19-21, pp. 367-372,2008.

[23] 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