A Comprehensive Survey on Handwritten Gujarati Character and Its Modifier Recognition Methods

Priyank D. Doshi and Pratik A. Vanjara

Abstract In India, handwritten character recognition is becoming necessity region-alwise due to new education policy 2020. Various technologies are applied to solve the problem in this area like statistical or probability model, support vector machine, Bayes probability model, deterministic finite automaton (DFA), hidden Markov model, and many more which are used. Due to the advancement in machine learning, convolutional neural network is a good solution of HCR which gives more promising results but any new algorithm in machine learning that depends on training data, mathematical function, loss function, and method of evaluation of model. Focusing on past research of handwritten Gujarati character recognition is found that sufficient research is required for modifier level called "Barakshari". Results obtained in past are limited to character level only. In this paper, our effort is to analyze and summarize previous contributions in the handwritten character recognition for several Indian languages.

Keywords Support vector machine · Bayes probability model · Deterministic finite automaton · Hidden Markov model · Convolutional neural network

1 Introduction

The regional languages processing is still in the growing phase in India where various approaches and techniques contributed by the researchers and many more required. Internet is widely used as a platform for multiple Indian languages in the area of education, e-commerce, documentation, information sharing, etc. Therefore, regional language processing is becoming a challenge day by day. In this context, demand of offline handwritten character recognition research for Indian regional languages is rising rapidly. Focusing on Gujarati and other Indian languages problem

P. D. Doshi (✉)
Atmiya University, Rajkot, Gujarat 360005, India

P. A. Vanjara
Shree M. & N. Virani Science College, Rajkot, Gujarat 360005, India

A. Joshi et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, Lecture Notes in Networks and Systems 191,
https://doi.org/10.1007/978-981-16-0739-4_79

of handwritten character recognition in the previous work is reviewed in this paper, and the challenges to solve Gujarati character recognition and its similar works in other languages are discussed. In this paper, we included HCR and OCR methods for Indian and foreign Languages to promote the work of Gujarati language handwritten character and vowels recognition. Offline handwritten character recognition for Indian regional languages is very much beneficial in number of ways.

Possible Advantages. Data should be taken from source so subsequently it avoids errors and unwanted changes which can be generated by retyping. Additionally, National Education Policy 2020 of India that depicts the medium of instruction will be student's mother tongue or regional language at least up to Grade 5. So Gujarati handwritten character recognition with vowels (character Modifiers) can be more advantageous to students and teachers. HCR can digitize old handwritten documents, postal address, and many more improvisation needed till it becomes easy and popular to use.

This paper is organized in five sections starting with general model (Table 1) which has mainly six different stages of handwritten character recognition given as

Table 1 Stages and task used in text recognition system

Stage	(Tasks)—description
Image acquirement	(Digitalization resizing, compression)-all images are converted into required format and size to prepare dataset
Preprocessing	(Noise removal, filtering, skew, thinning, edge detection and correction, morphological operation)-processing images to remove unwanted background to get region of interest. Image used to detect canny edge, bounded box, etc. Skewness detected and corrected
Segmentation	(Character based, word based, sentence based)-from the image, we required to separate lines with upper and lower boundaries. Then words are separated and then characters. A character made up of one or more connected components of pixels. Pixels intensity value can be used
Feature extraction	(Statistical—geometrical features)-curves and corners of connected components of pixels can be found to form characters. This is the area which is helpful in next stage of classification
Character recognition (classification)	The segmented content is fed to the classifier which gives result that input image corresponds to which character and modifiers
Post-processing	(Confusion matrix, contextual approaches dictionary-based approaches)—grouping characters and modifiers based on their location. Finding errors and correction. One to one correspondence between characters and unicode dictionary letters. Output obtained as a text file

under. Subsequently, it includes challenges motivation, efforts reviewed areas of their work, discussion findings, and conclusion in the last.

1.1 General Methodology

Scanned handwritten images are processed to get text file in target language. As shown in Table 1, processing of it divided into the phases like image acquirement, preprocessing, segmentation, feature extraction, classification, and post-processing.

2 Challenges and Motivation

Gujarati language character modifier recognition is required more efforts as it would be one step ahead since character recognition almost done so vowels recognition requires more research work. Upper line as it is in Devanagari language is missing in Gujarati, and many half characters are joined with another which is displayed differently. Skewed characters and modifiers are also giving variety to dataset. A character connected with another half character (conjunct consonants) shown as under and in Table 2 (counting them we get 12 so it is called "Barakshari") is also a challenging task. Following type of characters is varied very frequently person to person when written by their hands so it becomes very difficult to recognize by machine [1].

ક	ખ્ર	ગ્ર	ઘ્ર	ચ્ર	જ્ઞ
ચ્ય	જ્ય	ખ્ય	ઘ્ય	ચ્ય	જ્ઞ્ય
હ્ય	ધ્ય	ત્ત	હ્ન	ષ્મ	શ્ચ

Basically word formation is not possible without character modifiers. They are vowels attached with consonants. Even if character can be recognized, its combination with vowels (as shown in first column of Table 2) becomes difficult to recognize. Sometimes characters written by person may be vertically straight, forward slant, or backward slant. So it produces difficulty in character recognition.

3 Efforts Reviewed and Their Areas of Work

Various surveys in this context lead us to understand various [2] merits and demerits of techniques involved in each stage of text recognition. Language processing techniques [3] like chatbot, text-to-speech conversion, language identification, spell

Table 2 Basic shapes of geometry useful to recognize character modifier (vowels)

Character modifier	Detail of basic geometrical shapes observed in consonant modifier
અ–ક	No modifier present
આ–કા	Vertical line in right zone
ઈ – કિ	Vertical line in left zone, oval arc or circular arc in upper zone, connected or disconnected components
ઈ – કી	Vertical line in right zone, oval arc or circular arc in upper zone, connected or disconnected components
ઉ - કુ	Oval arc or circular arc in lower zone, connected components
ઊ - કૂ	Slanting line in lower zone, oval arc or circular arc in lower zone, connected components
એ - કે	Slanting line in upper zone
ઐ - કૈ	Two slanting line upper zone
ઓ - કો	Vertical line in right zone, slanting line in upper zone, disconnected components
ઔ - કૌ	Vertical line in right zone, Two Slanting line in upper zone, disconnected Components
અં - કં	One dot in upper zone
અઃ - કઃ	Two disconnected dots in right zone

check, medical record processing for real-time needs focused. We also found conclusion that [4] the selection of the classification as well as the feature extraction techniques needs to be proper in order to attain good rate in recognizing the character. Segmentation work in Gujarati character modifier still remains left. Over fitting problem in deep learning occurs when result is good fit on our model on the training data, but it is not generalized well enough on new data. It means model is very specific to the training data and inappropriate for other data. We can solve over fitting problem by adding more training data. So in deep learning [5] for image data augmentation result found that over fitting problem can overcome by improving the size and quality of training datasets. Augmentation like flipping, rotation, and zoom in and zoom out can be used to enrich the dataset in terms of size and quality. Comparatively it is found that offline character recognition is high accuracy and reliable system required [6], and also fuzzy membership functions could significantly outperform standard zoning methods [7]. For general stage of text processing having different techniques especially in classification stage including neural network, it is found that feature extraction and classification technique play an important role [8]

in the performance of character recognition rate and including neural network and data mining concepts merits and demerits [9] which are given for each stage.

3.1 Segmentation and Preprocessing

As there are six steps in general model of offline character recognition, and in each step, variety of scientific techniques of models can be applied segmentation and preprocessing can include histogram [10] projection and equalization technique [11], sliding window method, Hough transform technique [12] with preprocessing like dilation, erosion, and thinning. Also strip-based projection [13] and mixture of smearing and contour tracing for line segmentation are used in offline Gurumukhi language character recognition. Research has been done for segmenting modifiers as intact not subdividing further [14] from printed Bangla Text. They continued their work for overlapping, touching, and compound characters. Segmenting the upper and lower modifiers with characters is done by fuzzy functions in Hindi language [15].

3.2 Neural Network Used as a Classifier

Nowadays using deep learning [16] offline handwritten recognition system proposed based on convolutional neural networks (CNN) for Gujarati language with different accuracy and diversified data. Similarly by analyzing character's shape [17] and then neural network applied to recognize characters. For English language [18] applying neural network and KBS built for character recognition. Using ANN [19] an online multilanguage handwriting recognition system developed based on Bézier curves for 102 language including Gujarati using IAM-OnDB dataset. Paired adversarial learning (PAL) [20] method along with DNN is applied to recognize handwritten mathematical expression using CNN and RNN-based feature extractor.

Similarly for digit recognition [21] CNN and pooling layer for data reduction and fully connected layer and output layer applied as a classifier. For Urdu [22], HCR and digit recognition pioneer dataset were prepared with CNN. For HCR in English language shallow network based on the Fukunaga Koontz transform (FKT) [23] model used with neural network and to recognize characters from ancient Geez document [24] a deep CNN used with dataset. A model for English character recognition developed using DNN as classifier [25] and feature extraction by local Gabor pattern, Haar wavelet transform, histogram-oriented gradients, and grid level. Similarly, chain code and image centroid extraction method use along with feedforward ANN [26] as a classifier.

For classification of handwritten Bengali numerals [27], a model proposed with extra layers like zero padding, dropout, and max-pooling, and the number of filters

enhanced using CNN and for Gujarati numeral recognition [28] and multilayer feed-forward neural network and naive Bayes classifier used for handwritten Gurumukhi numeral [29] using backpropagation neural network with wavelength transform.

3.3 Support Vector Machine as a Classifier (SVM)

Using SVM [30–34] as a classifier many research work recorded for different languages like Gujarati and Devanagari. Along with SVM hybrid feature spaces like aspect ratio, image subdivision, polynomial kernel, Gaussian kernel are used. Also for Gujarati numeral recognition creating four features set of various size preprocessing and segmentation has been done. Distance profile, gradient profile, and wavelet form are used. Fourier descriptors as feature vectors and lexicon were used for post-processing.

3.4 K-Nearest Neighbor as a Classifier

K-nearest neighbor classifier [35] with distance transform method for zone identification and for segmentation projection profile and morphological operations used. Comparative analysis between KNN and principal component analysis (PCA) for Gujarati numerical classifier [2] developed. Noise removal, and thinning as preprocessing and stroke-based directional feature [36] also applied with this classifier. With this [37], normalization and interpolation used in preprocessing.

3.5 Other Method Used as Classifier

Decision tree approach for HCR for Gujarati characters [38] focuses on feature extraction of three types, i.e., connected or disconnected component, number of end point and number of closed loop. Similar concept like adjacent pixel connectivity and curvature-based pattern matching and classification used [39] in HCR method for Sinhala language.

For Gujarati language combined approach of Freeman Chain Code [40], Hu' Invariant Moment (4 order), center of mass applied with gradient feature. A deterministic finite automaton (DFA) [41] and fuzzy system [42] for intelligent word recognition also introduced in past using a regular grammar by labeling each pixel as vertical or as horizontal to group strokes for feature extraction.

A model using weighted integral image method, Bayes classifier [43] and statistical-structural features overcoming drawbacks of classic binarization method. For machine printed, Devanagari (Nepali) has been introduced in past.

Maximum mutual information [44], composite image and block-based PCA methods for character recognition are used in HCR. For English, writer independent character recognition HMM [45] used for each character with global and local processing features of images. For upper and lower modifiers and half characters are ignored [46] and topological features, heuristics for middle zone are used.

3.6 Dataset Creation

Including Gujarati language work noticed in dataset creation like page level handwritten document image dataset "PHDIndic_11" [47] of 11 official Indic scripts. Dataset for English language alphabets and numerals [48] is publically available and provides isolated characters and digits free of cost. A novel database [49] is consisting of 26,000 images of Hindi handwritten characters, and modifiers for offline recognition by segmentation and augmentation process were developed. Such dataset creation for Gujarati language is required to create for effective implementation HCR system.

4 Discussions and Findings

Due to emerging trends of machine learning, especially convolutional neural network can solve this problem more promisingly. As we can categorize machine learning algorithm in supervised or unsupervised in terms of label is present or absent. Label corresponds to prediction (output) which is based on features in training dataset. Further depending on labels, our problem of HCR lies in discrete classification problem. Along with sufficient training dataset, any new algorithm in machine learning depends on mathematical function like logistic, linear regression, loss function like mean square error (MSE), mean absolute error (MAE), and other classifications losses between predictions and actual observations. Methods of model evaluation also affect on this. So possibilities of new directions and improvements are always expected.

Unavailability or scarcity of dataset in HCR for Gujarati character modifiers (vowels) recognition is one of the necessary areas of work. Here we need to apply a holistic approach in which all stages should be integrated such a way that they should contribute effectively for central objective of the system. So successful integration of all phases of machine learning starting with features extraction (topological or geometrical) to classification is very important. Vowels sometimes are extremely cursive or artistic in nature. Using convolutional neural network, we can extract shapes (Table 2) by feature extraction and co-relate it with character to print and fed to classifier.

5 Conclusions

Offline handwritten Gujarati character modifier recognition ("Barakshari") is an interesting and challenging area of research. There are several classification and feature extraction techniques for handwritten character recognition problem. Various techniques can be applicable to complete the task in each steps of general model of recognition. Due to advancement in technology, machine learning attracts us toward selection of appropriate neural network model and it will be more effective if we can include good training dataset in case of Gujarati character modifier recognition. Any new algorithm in machine learning depends on how model has been trained with batch or mini batchwise of data, type of model or mathematical function has been used, its loss function, and method of evaluation. We can adjoin the effectiveness of each steps of model with good feature extraction of characters and modifiers using good augmentation techniques.

References

1. Patel, C., Desai, A.: Zone identification for Gujarati handwritten word. In: Proceedings of the 2nd International Conference on Emerging Applications of Information Technology. EAIT 2011, pp. 194–197 (2011)
2. Mj, B., Kv, K., Me, J.: Comparison of classifiers for gujarati numeral recognition. Int. J. Mach. Intell. **3**, 160–163 (2011)
3. Harish, B.S., Rangan, R.K.: A comprehensive survey on Indian regional language processing. SN Appl. Sci. **2** (2020)
4. Purohit, A., Chauhan, S.S.: A literature survey on handwritten character recognition. IARJSET **7**, 1–5 (2016)
5. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6** (2019)
6. Priya, A., Mishra, S., Raj, S., Mandal, S., Datta, S.: Online and offline character recognition: a survey. In: International Conference on Communication and Signal Processing. ICCSP 2016, pp. 967–970 (2016)
7. Impedovo, D., Pirlo, G.: Zoning methods for handwritten character recognition: a survey. Pattern Recognit. **47**, 969–981 (2014)
8. Sahu, V.L., Kubde, B.: Offline handwritten character recognition techniques using neural network: a review. Int. J. Sci. Eng. Res. **1**, 87–94 (2013)
9. Muthuraman, V.: A study on text recognition using image processing with datamining techniques. Int. J. Comput. Sci. Eng. Open Access (2019)
10. Dave, N.: Segmentation methods for hand written character recognition. Int. J. Signal Process. Image Process. Pattern Recognit. **8**, 155–164 (2015)
11. Dixit, S., Suresh, H.N.: South Indian Tamil language handwritten document text line segmentation technique with aid of sliding window and skewing operations. J. Theor. Appl. Inf. Technol. **58**, 430–439 (2013)
12. Shah, L. et al.: Rotation estimation of Gujarati script document using hough transform (2014)
13. Kumar, M., Jindal, M.K., Sharma, R.K.: A novel technique for line segmentation in offline handwritten Gurmukhi script documents. Natl. Acad. Sci. Lett. **40**, 273–277 (2017)
14. Akter, N., Hossain, S., Islam, M.T., Sarwar, H.: An algorithm for segmenting modifiers from Bangla text. In: Proceedings of the 11th International Conference on Computer and Information Technology. ICCIT 2008, pp. 177–182 (2008)

15. Pramanik, R., Bag, S., Kumar, R.: A fuzzy and contour-based segmentation methodology for handwritten Hindi words in legal documents. In: Proceedings of the 4th IEEE International Conference on Recent Advances in Information Technology. RAIT 2018, pp. 1–6 (2018)
16. Pareek, J., Singhania, D., Kumari, R.R., Purohit, S.: Gujarati handwritten character recognition from text images. Proc. Comput. Sci. **171**, 514–523 (2020)
17. Prasad, J.R., Kulkarni, U.V., Prasad, R.S.: Offline handwritten character recognition of Gujrati script using pattern matching. In: 2009 3rd International Conference on Anti-Counterfeiting, Security, and Identification in Communication. ASID 2009 (2009)
18. Kasthuri, M., Sivaprasatham, V.: Self-learning based cognitive reading and character recognition in image processing techniques. SN Comput. Sci. **1**, 1–12 (2020)
19. Carbune, V., et al.: Fast multi-language LSTM-based online handwriting recognition. Int. J. Doc. Anal. Recognit. **23**, 89–102 (2020)
20. Wu, J.W., Yin, F., Zhang, Y.M., Zhang, X.Y., Liu, C.L.: Handwritten mathematical expression recognition via paired adversarial learning. Int. J. Comput. Vis. (2020)
21. Ali, S., et al.: An efficient and improved scheme for handwritten digit recognition based on convolutional neural network. SN Appl. Sci. **1**, 1–9 (2019)
22. Ali, H., Ullah, A., Iqbal, T., Khattak, S.: Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network. SN Appl. Sci. **2**, 1–12 (2020)
23. Gatto, B.B., dos Santos, E.M., Fukui, K., Júnior, W.S.S., dos Santos, K.V.: Fukunaga–Koontz convolutional network with applications on character classification. Neural Process. Lett. (2020)
24. Demilew, F.A., Sekeroglu, B.: Ancient Geez script recognition using deep learning. SN Appl. Sci. **1**, 1–7 (2019)
25. Liu, Z., Pan, X., Peng, Y.: Character recognition algorithm based on fusion probability model and deep learning. Comput. J. **00** (2020)
26. John, J., Pramod, K.V., Balakrishnan, K.: Offline handwritten Malayalam character recognition based on chain code histogram. In: 2011 International Conference on Emerging Trends in Electrical and Computer Technology. ICETECT 2011, pp. 736–741 (2011)
27. Rahman, M.M., Islam, M.S., Sassi, R., Aktaruzzaman, M.: Convolutional neural networks performance comparison for handwritten Bengali numerals recognition. SN Appl. Sci. **1**, 1–11 (2019)
28. Sharma, A., Thakkar, P., Adhyaru, D., Zaveri, T.: Features fusion based approach for handwritten Gujarati character recognition. Nirma Univ. J. Eng. Technol. (2017)
29. Singh, P., Budhiraja, S.: Offline handwritten Gurmukhi numeral recognition using wavelet transforms. Int. J. Mod. Educ. Comput. Sci. **4**, 34–39 (2012)
30. Desai, A.A.: Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space. CSI Trans. ICT **2**, 235–241 (2015)
31. Farkya, S., Surampudi, G., Kothari, A.: Hindi speech synthesis by concatenation of recognized hand written Devnagri script using support vector machines classifier. In: 2015 International Conference on Communication and Signal Processing. ICCSP 2015, pp. 893–898 (2015)
32. Gupta, A., Srivastava, M., Mahanta, C.: Offline handwritten character recognition using neural network. In: ICCAIE 2011—2011 IEEE Conference on Computer Applications and Industrial Electronics, pp. 102–107 (2011)
33. Maloo, M., Kale, K.V.: Support vector machine based Gujarati numeral recognition. Int. J. Comput. Sci. Eng. (IJCSE) **3**, 2595–2600 (2011). ISSN 0975-3397
34. Nagar, R., Mitra, S.K.: Feature extraction based on stroke orientation estimation technique for handwritten numeral. In: ICAPR 2015—2015 8th International Conference on Advances in Pattern Recognition (2015)
35. Desai, A.A.: Gujarati handwritten numeral optical character reorganization through neural network. Pattern Recognit. **43**, 2582–2589 (2010)
36. Goswami, M., Mitra, S.: Structural feature based classification of printed Gujarati characters. In: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS, vol. 8251, pp. 82–87 (2013)

37. Gohel, C.C., Goswami, M.M., Prajapati, V.K.: On-line handwritten Gujarati character recognition using low level stroke. In: Proceedings of the 2015 3rd International Conference on Image Information Processing. ICIIP 2015, pp. 130–134 (2016)
38. Thaker, H.R., Kumbharana, C.K.: Structural feature extraction to recognize some of the offline isolated handwritten Gujarati characters using decision tree classifier. Int. J. Comput. Appl. **99**, 46–50 (2014)
39. Madushanka, P.T.C., Bandara, R., Ranathunga, L.: Sinhala handwritten character recognition by using enhanced thinning and curvature histogram based method. In: 2017 IEEE 2nd International Conference on Signal Image Processing. ICSIP 2017, Jan 2017, pp. 46–50 (2017)
40. Macwan, S.J., Vyas, A.N.: Classification of offline Gujarati handwritten characters. In: 2015 International Conference on Advances in Computing, Communications and Informatics. ICACCI 2015, pp. 1535–1541 (2015)
41. Álvarez, D., Fernández, R., Sánchez, L.: Stroke-based intelligent character recognition using a deterministic finite automaton. Log. J. IGPL **23**, 463–471 (2014)
42. Álvarez, D., Fernández, R.A., Sánchez, L.: Fuzzy system for intelligent word recognition using a regular grammar. J. Appl. Log. **24**, 45–53 (2017)
43. Joshi, V., Panday, S.P.: Character component segmentation and categorization of machine printed text in Devanagari (Nepali) script in digital image processing. In: Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security. ICCCS 2018, pp. 191–198 (2018)
44. Nopsuwanchai, R., Povey, D.: Discriminative training for HMM-based offline handwritten 2. In: Maximum Mutual Information Training of Analysis (2003)
45. Das, R.L., Binod, I., Prasad, K., Sanyal, G.: HMM based offline handwritten writer independent english character recognition using global and local feature extraction. Int. J. Comput. Appl. **46**, 975–8887 (2012)
46. Garg, N.K., Kaur, L., Jndal, M.: Recognition of offline handwritten hindi text using middle zone of the words. In: 2015 IEEE/ACIS 14th International Conference on Computer and Information Science. ICIS 2015—Proceedings, pp. 325–328 (2015)
47. Obaidullah, S.M., Halder, C., Santosh, K.C., Das, N., Roy, K.: PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification. Multimed. Tools Appl. **77**, 1643–1678 (2018)
48. Yousaf, A., Khan, M.J., Imran, M., Khurshid, K.: Benchmark dataset for offline handwritten character recognition. Proceedings of the 2017 13th International Conference on Emerging Technologies. ICET2017, Jan 2018, pp. 1–5 (2018)
49. Nehra, M.S., Nain, N., Ahmed, M.: Benchmarking of text segmentation in devnagari handwritten document. In: 2016 IEEE 7th Power India International Conference. PIICON 2016, pp. 0–3 (2017)