Proceedings of the 16th INDIACom; INDIACom-2022; IEEE Conference ID: 54597
2022 9th International Conference on "Computing for Sustainable Global Development", 23rd - 25th March, 2022
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Image Set Quality Optimization for Handwritten Gujarati Character and Its Modifier Recognition

Priyank D. Doshi
Department of Computer Science and Information Technology
Atmiya University, Rajkot, India
doshipriyank76@gmail.com

Pratik A. Vanjara
Department of Computer Science and Information Technology
Shree M & N Virani Science College, Rajkot, India
pavanjara@gmail.com

Abstract — The problem of recognizing handwritten Gujarati characters has been tried by many researchers but still it requires enough work from vowel recognition up to its online application. The problem becomes even more complex if we use characters with vowels. Machine learning and Deep learning are extensively used to solve the image classification problem. It is observed by reviewing survey papers that Support Vector Machine, Bayes Probability Model, Deterministic Finite Automaton (DFA), Hidden Markov Model techniques are used as classifiers in this problem but machine learning and deep learning gives more promising result. It requires large data set to train and test the model. We collected hand-written Gujarati 'Barakshari' and text images from more than 1000 people having different ages. Deep learning requires a comparatively larger image set than machine learning. Both can have their pros and cons and so it is very much essential to optimize the data set if we are using the 'hybrid mode' to get benefits from both. Different augmentation techniques are also applied to the image set to raise the size, quality, and variety.

Keywords— Support Vector Machine, Machine Learning, Neural Network, Deep Learning, Deep hybrid Learning, Hand Written Character Recognition.

I. INTRODUCTION

Initially, this research in Gujarati character recognition was initiated by Hetal R. Thaker and Dr. C.K. Kumbharana[1]. They have worked for structural feature extraction to recognize some of the offline isolated Handwritten Gujarati Characters using Decision Tree Classifier. After that many more researchers have contributed to this area.

In the review paper presented by us, it is concluded that for different languages, with a variety of classifiers and segmentation, technologies are available [18]. The solution for this is becoming more promising nowadays because of the popularity of machine learning and deep learning. Many tasks like feature extraction and training models are performed by artificial neural networks.

We are applying the proposed model as shown in Fig. 1 in a sequence of six basic steps from image acquisition, pre-processing segmentation, feature extraction, classification, and post-processing. Feature extraction and classification process are taken as a hybrid of machine and deep learning.

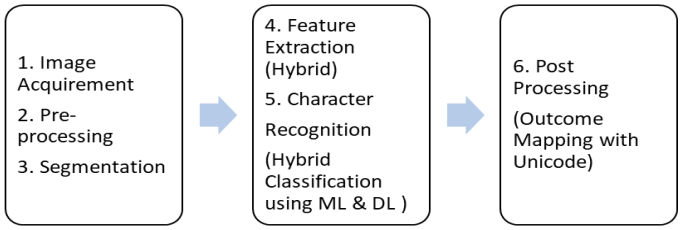

Fig. 1. A hybrid model for recognizing HCR in Gujarati

A. Advantages of Gujarati Character Recognition :

- The basic advantage of Gujarati Handwritten character recognition is to avoid re-typing in a specific font which is sometimes not as easy as we generally do.
- Benefits of this can be seen in many areas of daily life, like hand-written address recognition on the cover at the post office or currier office.
- This will be a solution for record-keeping in Gujarati for old documentation after conversion in digital text form.
- Many online platforms provide their services in local or regional languages to expand their business.
- For students up to 5th grade: Teaching these children in a local or regional language is mandatory.
- Therefore, expanding this scenario to people of every age shows that this type of research is highly encouraging and demanding.

B. Problem solution with Machine Learning & Deep Learning :

- This HCR recognition in the Gujarati language is a multiclass problem in a machine or deep learning. Each class represents a character or a combined character with a vowel. We need a sufficient amount of image set that can test the model. To get a good result it should be appropriate in quantity and cover all the classes evenly in proportion so that proper labeling can be done with each sample in supervised learning. Later we can map each outcome of the classifier with Unicode in post-processing.
- Any machine learning algorithm, if it is unsupervised, then dependency on the data set rises higher. The

 ISSN 0973-7529; ISBN 978-93-80544-45-8

algorithm differs by learning rate, gradient descent - stochastic, mini-batch, and batch with other researcher-wise intricacies. In all of these Image Sets, size and quality play an important role.

II. Efforts Reviewed in Dataset Creation with a Hybrid Model Concept

J. Pareek, D. Singhania, R. R. Kumari, and S. Purohit created Data set of the size 10000 images from 250 different people for the Gujarati Language[2]. They have used supervised machine learning. The training and test ratio of their data set used was 80% - 20% respectively. There was a total of 59 classes including vowels, modifiers, and consonants.

M. Sridharan, D. C. Rani Arulanandam, R. K. Chinnasamy, S. Thimmanna, and S. Dhandapani worked for Tamil letters recognition and they did it for 256 Tamil letters and six fonts total of 1536 letters are used to pre-train the system using deep convolutional neural network[3]. R. Chaudhari tried on the problem for joint characters or conjunct characters with low accuracy[4].

M. M. Goswami and S. K. Mitra summarized different OCR-based system database samples per class label for the Gujarati script[5]. As in Class variation, two of them are up to 10 and another two are up to 50. Others are ranging from 119 to 250 classes. Goswami used 239 classes and 16000 samples are collected from newspapers, books, etc.

H. Patel used 50 Conjunct characters for classification and showed in the paper[6]. Also, the Unicode table is given for Gujarati Characters. Used Zernike moment feature extractor.

D. S. Joshi and Y. R. Risodkar used a k-NN classifier but not enough specifications of the data set were found[7]. A. Shirke, N. Gaonkar, P. Pandit, and K. Parab used the Yolo labeling tool in the context of training the model[8]. Not enough specification of data set given.

Very few from above have mentioned details about data set and rarely one for machine learning. The process of image set creation is an important task and should be properly justified.

S. Liaqat, K. Dashtipour, K. Arshad, K. Assaleh, and N. Ramzan used a hybrid approach of deep learning and machine learning in the Posture detection problem[9]. They have applied support vector machine (SVM), logistic regression (KNN), decision tree, Naive Bayes, random forest, Linear discrete analysis, and Quadratic discrete analysis) and deep learning classifiers (i.e., 1D-convolutional neural network (D- CNN), 2D-convolutional neural network (2D-CNN), LSTM, and bidirectional LSTM) to identify posture detection. Ultimately, they can predict the health of different ages of people.

For offline recognition of Segmented Gurumukhi characters Anupam Garg, Manish Kumar Jindal, and Amarpreet Singh used principal component analysis and modified division point-based features and fed them in the classification process k-NN and SVM having three kernels linear, polynomial, and RBF – SVM. They have used 8960 samples of offline handwritten Gurumukhi characters[10]. They achieved 92.3 % accuracy. In this paper, they have worked upon characters but not modifiers and vowels. Since complexity rises when vowels are added into the dataset, it creates further interest in research in this area similarly in Gujarati character and modifiers recognition.

As we know image classification can solve many problems including character and vowels recognition, a survey on support vector machines has been presented by Mayank Arya Chandra and S. S. Bedi[11]. They have included techniques about support vector machines after and before 2000 A.C. In their survey they have positively focused on the importance of SVM as well as neural networks.

In the context of the hybrid concept of machine learning-based model Abid Sarwar, Mehboob Ali, Jatinder Manhas, and Vinod Sharma presented their work for the diagnosis of diabetes type – II[12]. Their data was collected from a group of 400 people. They have used Artificial Neural Networks, K-Nearest Neighbours, Naive Bayes, and support vector machines to make it a hybrid. The result of each one was ensembled and the final result was then provided.

In the problem to recognize handwritten Marathi numeral Deepak T. Mane, Rushikesh Tapdiya and Swati V. Shinde presented a model based on an ensemble neural network[13]. Using CNN, they have created 5 base pipelines and then its output goes to the Meta classifier network to get desired output. The whole work was based on machine learning so they used 63000 samples for training, 7000 samples used for validation, and 11500 samples for testing. They concluded and suggested using dilated customized CNN to increase accuracy and reduce computational cost.

We have observed Gujarati Language work done by Obaidullah, S. M., Halder, C., Santosh, K. C., Das, N. & Roy in Page Level Handwritten Document Image Data Set "PHDIndic_11" of 11 official Indic scripts[14].

The data set developed by Yousaf, A., Khan, M. J., Imran, M. &Khurshid, K. for English Language alphabets and numerals is publicly available and provides isolated characters and digits free of cost[15].

A novel database consisting of 26000 images of Hindi handwritten characters and modifiers for offline recognition by segmentation and augmentation process was developed by Nehra, M. S., Nain, N. & Ahmed, M[16].

Such data set created for the Gujarati Language is required to create and make it public for effective implementation HCR system.

As we have seen above, the Hybrid concept in feature extraction is successful, and also in classifier modeling we can apply hybrid concept using deep and machine learning to get a more robust algorithm for Gujarati characters with vowels recognition.

III. Observations and Findings

It may result in over-fitting when less data is provided. In deep learning, an overfitting problem arises when the result is a good fit on our model using the training data smaller in size, but it fails when new data is tested. So, it implies that the model is very specific to the training data and inappropriate for other data.

Secondly, Deep learning requires more computational power and resources which is not required in classical machine learning algorithms. GPU has a hardware requirement whereas in machine Learning we can train with fewer data set as compared to Deep Learning with a CPU.

Deep learning will take more time to train. Feature extraction task is carried out by Deep learning by adding extra layers itself. Using additional network layers, deep learning itself does feature extraction from unstructured data, whereas in machine learning we need to understand the dataset. We can say it creates its artificial network to learn and make smart judgments. Machine Learning makes judgments based on what it has learned. By engineering features, we can achieve final performance and accuracy in ML whereas, in deep learning, final classification is driven by fully connected neural network layers.

ML is used for categorization and scoring purposes whereas DL outputs may be in any shape, text, sound, or freeform components. DL can have more tuning capacity than ML. DL can be used in more complex learning machine learning problems.

The word hybrid indicates tractability that can be adopted in this recognition process like deep learning is more flexible in hyperparameter tuning and accuracy whereas machine learning can be trained on CPU and with less data size.

There are many such features from DL and ML which can be tuned to get the best results for the problem of handwritten character recognition in the Gujarati Language. In the context of Image Set Size DL requires a larger image set than ML. Such differences can be eliminated or tuned in by Hybrid Model for this problem.

To take benefits of both approaches we can infuse them and apply Deep Hybrid Learning. By using both DL and ML as a hybrid approach, good results can be achieved and drawbacks of both can be eliminated.

In the post-processing of our problem, we need to map each outcome to a single Unicode Gujarati character. There are combined characters including vowels in the segmentation process as shown in Fig. 1 and we need to map them combined characters given the first column of Table I.

To solve the overfitting problem, we need more training data. We collected handwritten Gujarati images shown in Fig. 2 & Fig. 3 from more than 1000 people of various ages. Fig. 2 contains the first part of the combined characters of Gujarati called 'Barakshari' that is the combination of characters and vowels. Fig. 3 contains handwritten paragraphs in Gujarati.

Image data augmentation is related to overfitting problems that can be overcome by improving the size and quality of training datasets surveyed by C. Shorten and T. M. Khoshgoftaar [17]. Augmentation is used to enrich the data set in terms of size and quality like flipping, rotation, zoom in and zoom out, etc. So, it will be a crucial task to set an appropriate image set to solve this problem.

As shown in TABLE I. we have three columns. The first includes a combined character with vowels, the second column is a Gujarati character and the third column includes vowels in the Gujarati Language. After segmentation, we can get vowels from any column as shown in Table I. Here, for a single character, we can get a combined character shown in the first column, a combination of the second and third column in Table I.

In Fig. 4, an example of a handwritten image is given. After the segmentation process, we can get images of the size 28X28 pixels which can be combined characters along with vowel or a single character.

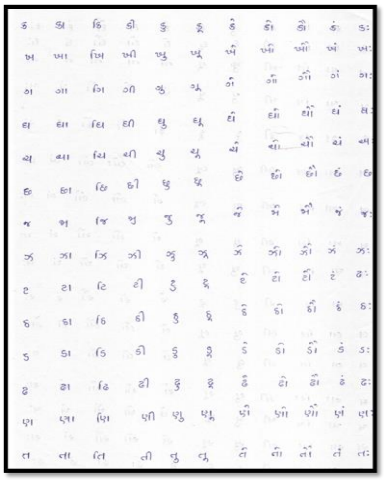

Fig. 2. Barakshari'

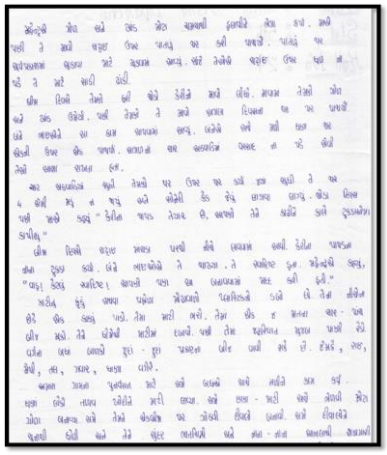

Fig. 3. Gujarati text

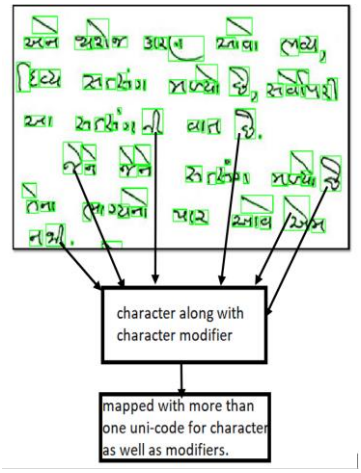

Fig. 4. Segmented Characters along with modifiers

Simply calculating the total number of possible unique segmented images from handwritten Gujarati Image can be 432 as we have 36 characters and 12 vowels. By multiplying these we get 432, so for sampling, we require a maximum of 432 classes.

Ignoring other special characters which are very infrequent in the Gujarati Language we can decide 432 as maximum unique segmented images. In each class, we include a good number of segmented images and in our case, it is more than 1000. Later the data set can be optimized for quality and size by different augmentation processes and can be used to make it a richer set so that it can be compatible with a hybrid model.

In supervised machine learning, we need to create a separate class for such images given in the first and third columns separately and label these images. As a result, we need to map each outcome of classification to one or more Unicode Gujarati characters in post-processing.

TABLE I. CHARACTERS AND ITS MODIFIERS

Character with modifier	Combination with one character only	
	Character	*Vowel (Modifier)*
કં	ક	અં
કા	ક	આ
કિ	ક	ઇ
કી	ક	ઈ
કુ	ક	ઉ
કૂ	ક	ઊ
કે	ક	એ
કૈ	ક	ઐ
કો	ક	ઓ
કૌ	ક	ઔ
કઃ	ક	અઃ

IV. CONCLUSIONS

Improvement of any machine learning, deep learning, or hybrid algorithm includes improvement in data set, appropriate selection of model, training, and evaluation methods. In our case, we have collected images from more than 1000 people. Thus, we got 43200 segmented images for 432 unique classes. By augmentation process, we can have image size more than two-fold or three-fold. The size of the data set plays an important role as we have fixed classes and a label for each class. In this case image set is the foundation of machine learning or deep learning algorithm. The proper size of training and test set, learning rate, gradient descent – stochastic, batch-wise, or mini-batch, and all other intricacies of Artificial Neural Network makes it more accurate and unique.

V. FUTURE SCOPE

The use of Machine Learning and Deep Learning altogether to solve the problem of handwritten characters with vowels recognition seems necessary and sufficient. The future scope will be the applicability of algorithms by putting them online through smartphones and websites. Android and iPhone are widely used in Urban and Rural areas of Gujarat. Also, we can use Hand Scanner at offices of public or private sectors including government offices like the post, railway, city survey offices, etc. where records and documents are received and stored in regional handwritten Gujarati language. Another stack holder can be schools where primary education will be in regional or mother language according to NEP 2020. Secondly, once the problem of hand-written Gujarati characters with vowels recognition is solved, the future scope will be detection and conversion of Gujrati phonetic language to text.

ACKNOWLEDGMENT

I cannot express enough thanks for their continued support and encouragement: Dr. Stavan C. Patel, my Head of Department; Dr. Bankim L Radadiya, Navasari Agricultural University, Department of Statistics, Surat, Gujarat, India. I offer my sincere appreciation for the learning opportunities provided by the management of Atmiya University.

My completion of this project could not have been accomplished without the support of students of Atmiya University, Atmiya School, and their parents. thank you for writing Gujarati Language Barakshari and Paragraphs.

REFERENCES

[1] H. R.Thaker and C. K. Kumbharana, "Structural Feature Extraction to recognize some of the Offline isolated Handwritten Gujarati Characters using Decision Tree Classifier," Int. J. Comput. Appl., vol. 99, no. 15, pp. 46–50, 2014, doi: 10.5120/17452-8381.

[2] J. Pareek, D. Singhania, R. R. Kumari, and S. Purohit, "Gujarati Handwritten Character Recognition from Text Images," Procedia Comput. Sci., vol. 171, no. 2019, pp. 514–523, 2020, doi: 10.1016/j.procs.2020.04.055.

[3] M. Sridharan, D. C. Rani Arulanandam, R. K. Chinnasamy, S. Thimmanna, and S. Dhandapani, "Recognition of font and Tamil letter in images using deep learning," Appl. Comput. Sci., vol. 17, no. 2, pp. 90–99, 2021, doi: 10.23743/acs-2021-15.

[4] R. Chaudhari, "Analysis of Handwritten Joint Characters in Gujarati Language," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 12, pp. 758–762, 2020, doi: 10.22214/ijraset.2020.32614.

[5] M. M. Goswami and S. K. Mitra, Printed Gujarati character classification using high-level strokes, vol. 704. Springer Singapore, 2018.

[6] D. Sihag, "[IJCST-V8I5P9]: Honey Patel Gujarati Ocr : Compound Character Recognition Using Zernike."

[7] D. S. Joshi and Y. R. Risodkar, "Deep Learning-Based Gujarati Handwritten Character Recognition," 2018 Int. Conf. Adv. Commun. Comput. Technol. ICACCT 2018, pp. 563–566, 2018, doi: 10.1109/ICACCT.2018.8529410.

[8] A. Shirke, N. Gaonkar, P. Pandit, and K. Parab, "Handwritten Gujarati Script Recognition," 2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021, pp. 1174–1179, 2021, doi: 10.1109/ICACCS51430.2021.9441811.

[9] S. Liaqat, K. Dashtipour, K. Arshad, K. Assaleh, and N. Ramzan, "A Hybrid Posture Detection Framework: Integrating Machine Learning and Deep Neural Networks," IEEE Sens. J., vol. 21, no. 7, pp. 9515–9522, 2021, doi: 10.1109/JSEN.2021.3055898.

[10] A. Garg, M. K. Jindal, and A. Singh, "Offline handwritten Gurmukhi character recognition: k-NN vs. SVM classifier," Int. J. Inf. Technol., vol. 13, no. 6, pp. 2389–2396, 2021, doi: 10.1007/s41870-019-00398-4.

[11] [M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," Int. J. Inf. Technol., vol. 13, no. 5, pp. 1867–1877, 2021, doi: 10.1007/s41870-017-0080-1.

[12] [A. Sarwar, M. Ali, J. Manhas, and V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning-based ensemble model," Int. J. Inf. Technol., vol. 12, no. 2, pp. 419–428, 2020, doi: 10.1007/s41870-018-0270-5.

[13] D. T. Mane, R. Tapdiya, and S. V. Shinde, “Handwritten Marathi numeral recognition using stacked ensemble neural network,” Int. J. Inf. Technol., vol. 13, no. 5, pp. 1993–1999, 2021, doi: 10.1007/s41870-021-00723-w.

[14] S. M. Obaidullah, C. Halder, K. C. Santosh, N. Das, and K. Roy, “PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification,” Multimed. Tools Appl., vol. 77, no. 2, pp. 1643–1678, 2018, doi: 10.1007/s11042-017-4373-y.

[15] A. Yousaf, M. J. Khan, M. Imran, and K. Khurshid, “Benchmark dataset for offline handwritten character recognition,” Proc. - 2017 13th Int. Conf. Emerg. Technol. ICET2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICET.2017.8281752.

[16] M. S. Nehra, N. Nain, and M. Ahmed, “Benchmarking of text segmentation in devnagari handwritten document,” 2016 IEEE 7th Power India Int. Conf. PIICON 2016, pp. 1–4, 2017, doi: 10.1109/POWERI.2016.8077422.

[17] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[18] P. D. Doshi, P. A. Vanjara, â€œA Comprehensive Survey on Handwritten Gujarati Character and Its Modifier Recognition Methods.â€• In: Joshi A., Mahmud M., Ragel R.G., Thakur N.V. (eds) Information and Communication Technology for Competitive Strategies (ICTCS 2020). Lecture Notes in Networks and Systems, vol 191. Springer, Singapore. https://doi.org/10.1007/978-981-16-0739-4_79