International Journal on Emerging Technologies **11**(2): 793-796(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Synthetic Data for Hindi Character Recognition

***Madhuri Yadav*[1] *and Ravindra Kumar Purwar*[2]**
[1]Research Scholar, USIC & T, GGSIPU Delhi, India.
[2]Associate Professor, USIC & T, GGSIPU Delhi, India.

(Corresponding author: Madhuri Yadav)

(Received 04 December 2019, Revised 30 January 2020, Accepted 05 February 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The success of deep learning techniques has provided solution to many pattern recognition problems. It has shown abrupt advancement in technology related to artificial intelligence and machine learning. Deep learning involves self learning using convolved filters and emphasize on automatic feature extraction. This self learning requires huge amount of training data to learn from different examples so that recognition system can handle variations in character at testing time. The major challenge in Hindi Character recognition is scarcity of handwritten database. The training dataset should contain variety of characters to build robust recognition system. The field of Hindi handwritten character recognition has scarcity of grayscale databases, thus techniques for generating synthetic dataset for Hindi isolated characters has been proposed in this paper. This synthetic dataset is generated using original handwritten isolated characters of Hindi language. The synthetic data is trained and tested on convolutional neural network architecture for recognition and experimental results show the efficacy of this dataset, it can be used independently or in combination with original handwritten characters.

Keywords: Data augmentation, Hindi characters, Handwritten isolated characters, Deep learning, Synthetic data.

I. INTRODUCTION

In last few decades, Human beings have become technology savvy and want computing systems which read, write and understand documents in their language. They want computers to read documents as they read computer media. This necessity of automation gave rise to the field of pattern recognition and machine learning. The growing trend of digitization has led to increase of research in the field of Optical Character Recognition (OCR). Automatic character recognition helps in mail sorting by address recognition in post offices and courier offices, biometric identification by automatic signature verification, digitization of forms and studies in academic institutions, restoration of historical documents.

OCR of handwritten documents is very challenging due to variations in character's shape. A single character can be written in different ways due to writer's writing style. Although there are numerous challenges in handwritten character recognition, research community is striving hard to automate character recognition. One of the recent works, investigated different Convolutional Neural Network (CNN) architectures and proposed a CNN architecture for Hindi handwritten characters [1]. Due to scarcity of benchmark dataset in this language, researchers propose their own dataset and test its efficiency using different algorithms. One such effort is proposed by [2] where database for isolated handwritten Hindi characters is created and then, it is trained and tested on deep learning based CNN architecture. An approach based on combination of HoG (histogram of oriented gradients) and Hu-moments is proposed for Hindi character recognition [3]. These features are evaluated on two classifiers: Support Vector Machines (SVM) and Multi layer Perceptron (MLP). SVM was proved to more efficient than MLP in this work. Other works used statistical and structural features such as regular expressions, gradient masks, end points, loops, intersection points, polynomial coefficients, curvelet transform as features. These works are the tremendous efforts of researchers to make optical character recognition a commercial success, which is yet not possible. The recognition systems of other languages such as Chinese, Arabic, English etc are far more successful than Hindi language [4]. The main reason behind this is the transition of recognition techniques from manual learning to deep learning. The recent works on recognition techniques of these languages use deep learning based CNN framework as manual learning techniques have reached a stagnant point. Another reason for their success is availability of benchmark databases for these languages. Some of the famous databases for Arabic language are Chinese language [5, 6].

Along with feature extraction and classification techniques, databases also play major role in building robust character recognition systems. The database must contain different variety of characters and they should be huge enough so that system can learn while training. Hindi language lacks in availability of standard or huge databases. Most of the works based on Hindi language create their own dataset and they are not publicly available. Thus, there are no standard databases to evaluate recognition accuracies of different techniques for Hindi language.

On the other hand, deep learning has provided solution to almost all image classification problems and thus, it should be used for Hindi character recognition as well. Deep learning techniques have emphasized the need of large dataset as they automate the process of feature extraction and learn through examples. Thus, to use deep learning techniques, there is a need of huge datasets which contain different type of characters.

Keeping in view the need of the hour, the main focus of this work is to generate synthetic data for Hindi language. It uses geometric rotations, horizontal and vertical flipping of characters and Gaussian noise for synthetic data generation. This work trains and tests the synthetic data on CNN architecture. This synthetic data can be augmented with original handwritten characters and help in building a huge dataset with varsity of characters.

Rest of the paper is organized as follows, Section II briefly describes the databases used in literature works of Hindi language, Section III explains the data augmentation techniques used in the proposed work, Section IV details the convolutional neural network architecture used for training and testing synthetic data, section V demonstrates the experimental results and the recognition accuracy achieved used different transformations and Section VI concludes the research work with future directions.

II. DATABASES FOR HINDI LANGUAGE

This section briefly explains about the databases used in the literature works of Hindi language. One of the well known database is ISI Kolkata database [7]. They collected data from mails, job applications and artificially created forms with pre-printed rectangular boxes. This database contains 30,000 characters and 22,256 numerals from 1049 writers. It has been used in [8, 9]. Another database created in [10] contains 20,305 isolated characters collected from 750 writers. This database was generated by scanning handwritten isolated characters at 300 dpi. Hindi character recognition technique based on structural and statistical features [11] used this dataset for result analysis. An offline database is created using polygonal approximations of online characters using electronic equipment, it is called HPL database [12]. The grayscale database consisting of 92 thousand images of 46 different character classes was generated by help of people with different age groups [2]. A technique based on curvelet transformed was proposed and evaluated on k-nearest neighbour classifier [14]. It collected 200 samples of each character class from 100 different writers and used it as database. Another dataset of 8224 handwritten character images of 32 character classes with a frequency of 257 samples per character class was proposed [15]. There are various other in-house databases but none of them have included transformations in their datasets. Thus, this work is an effort to generate a robust dataset which includes different transformations of original characters and can handle noisy characters.

III. DATA AUGMENTATION METHODS

It is believed that the reason behind success of convolutional neural networks is their deep layers, optimization algorithms and huge training data. The availability of huge datasets prevents network from overfitting and cover different variations of characters for better generalization. In this work, the main focus is on generation of synthetic data using geometric transformations and Gaussian noise. These transformations are applied on original Hindi handwritten characters collected from 108 writers. The details of collection of this database are provided in [16]. Originally, this database consists of 4428 characters and 41 classes.

A. Geometric transformations

The geometric transformation consists of affine transformations such as rotation, skew, vertical and horizontal flipping. The rotation is applied on character images with random range of 0° to 180°. 3 different character images are created from each character image by applying random rotations. Each character class has 108 images. Thus, a total of 324 images are created for each character class using rotation transformation. Fig. 1 represents the rotated character images of different character classes. The character images may not always be centered; they may be present anywhere in the frame. Thus, there is a need to train the network for off-center images. Hence, separate random shifting of height and width of characters is also used as a transformation. These transformations help in developing different data conditions for training data and build a recognition system which could handle different distortions. Fig. 2 and 3 represents the height and width shifts of few character images. Each character image corresponds to one width shifted and one height shifted image. Thus, each image has 2 randomly shifted images and the images in each character class increases from 108 to 216.

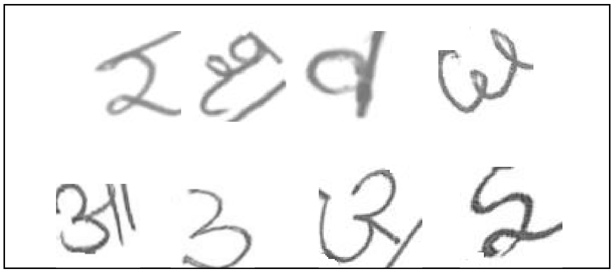

Fig. 1. Few rotated samples of different character classes.

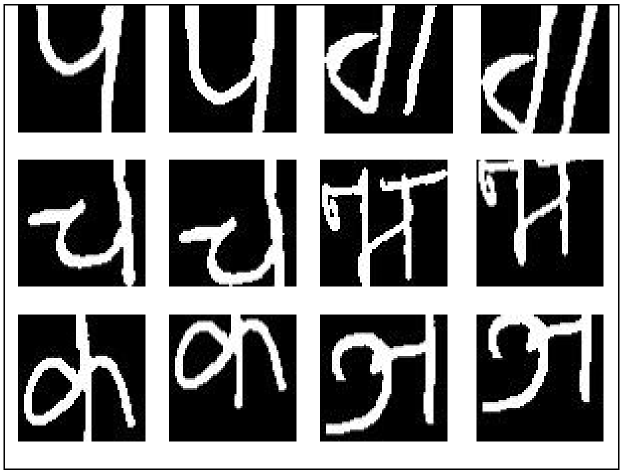

Fig. 2. Examples of random height shifting of different character classes.

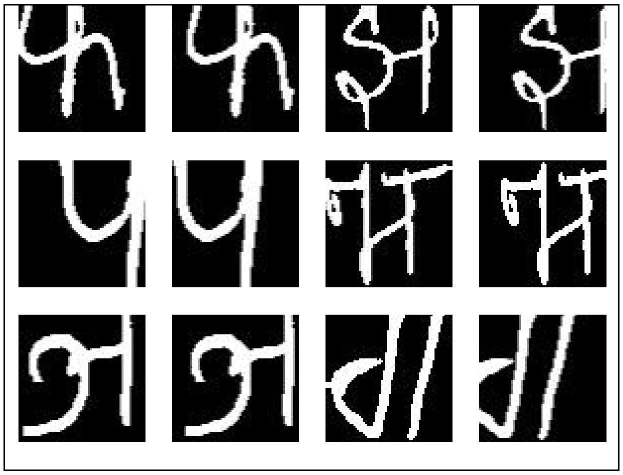

Fig. 3. Examples of random width shifting of different character classes.

Fig. 4 and 5 represents the vertical and horizontal flipping of character images.

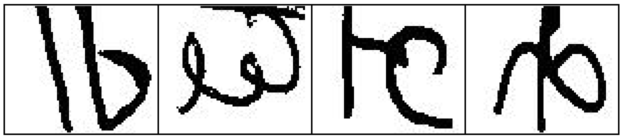

Fig. 4. Vertical flipping of different character classes.

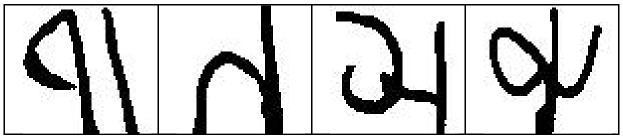

Fig. 5. Horizontal flipping of different character classes.

B. Gaussian Noise

There can be noise in character images due to data acquisition devices, poor illumination and dust particles. Gaussian noise is used in this work for synthetic data generation to train network for noisy images.

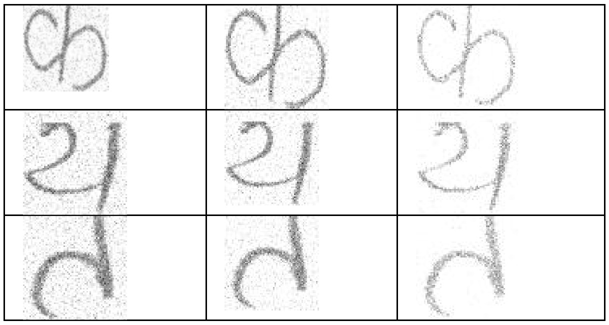

Fig. 6. Character samples showing noise of different variances 0.01, 0.05 and 0.2 in different columns respectively.

Character images with Gaussian noise of 0 mean and 0.01, 0.05 and 0.2 variance are created. Thus, each image corresponds to 3 noisy images with different noise variations. Fig. 6 represents character with different noise variance.

IV. CNN ARCHITECTURE

Convolutional neural network consists of basic layers which are convolutional layer, max pooling layer, dense layer and classification layer. This work uses CNN architecture with three convolution layers and two dropout layers. Max-pooling layers and dropout layers are used to reduce overfitting [1]. The architecture along with the input and output dimensions of each layer is as shown in Fig. 7.

The original dataset along with the synthetic data is trained on convolutional neural network. The recognition rates are discussed in next section.

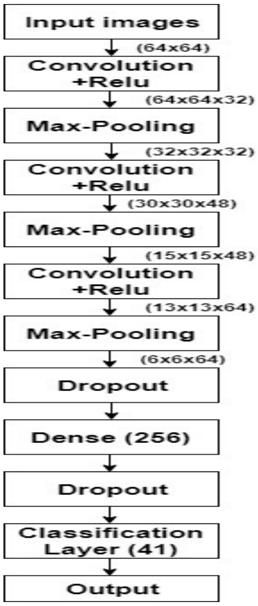

Fig. 7. CNN architecture used in this work.

V. RESULTS

The experiments are conducted using Keras library on intel core i5 processor.

Table 1 represents the total contribution of different synthetic data generation techniques. The CNN architecture is used to train various combinations of original dataset and synthetic data. It is trained for 50 epochs using adadelta optimization. The experimental results are tabulated in Table 2.

The original characters are trained in combination with synthetic data. After including these transformations in the dataset, the system becomes robust enough to handle these deformations on real system testing.

Table 1: Total contribution of different synthetic data generation techniques.

Deformation	Contribution in each class	Contribution in Database
Rotation	3 x 108=324	432 x 41=13284
Width shifting	1 x 108=108	108 x 41= 4428
Height shifting	1 x 108=108	108 x 41= 4428
Horizontal flipping	1 x 108=108	108 x 41= 4428
Vertical flipping	1 x 108=108	108 x 41= 4428
Gaussian Noise	3 x 108=324	324 x 41=13284
Total synthetic images		= 44280

Table 2: Recognition accuracies obtained using different combinations of synthetic data.

Dataset used	Recognition accuracy (in %)
Original characters +rotated data	90.86
Original characters+height translation	96.01
Original characters+width translation	96.39
Original characters +horizontal flipping	97.20
Original characters + vertical flipping	97.08
Original characters + guassian noise(0.01)	97.54
Original characters + guassian noise(0.2)	95.09
Original characters + guassian noise(0.05)	96.32
Rotation+flipping+translation+noise	92.68

VI. CONCLUSION

This work proposed synthetic training data for handwritten Hindi isolated characters. The synthetic data was created using original handwritten characters collected using different writers. The techniques used for generating synthetic data are rotation, height and width translation, horizontal, vertical flipping and Gaussian noise. The generated data is trained and tested using CNN architecture. The synthetic data can be used independently or in combination with original dataset. This work can be further extended by generating word database for Hindi characters.

VII. FUTURE SCOPE

The generated synthetic data is a vital solution for grayscale handwritten Hindi character database. In future, this database can be extended using other transformation techniques such as shearing. The different types of transformations in the dataset help in creating robust recognition system.

Conflict of Interest. No.

ACKNOWLEDGEMENT

I would like to thank anonymous reviewers for providing their valuable suggestions for improving quality of this paper. I am also thankful to editor for considering my paper for publication.

REFERENCES

[1]. Yadav, M., Kr Purwar, R., & Jain, A. (2018). Design of CNN architecture for Hindi Characters. *Advances in Distributed Computing and Artificial Intelligence Journal, 7*(3), 47-61.

[2]. Pant, A., Gyawali, P., & Acharya, S. (2015). Deep learning based large scale handwritten Devanagari character recognition. In *9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (pp. 1-6). Kathmandu: IEEE.

[3]. Yadav, M., & Purwar, R. (2018). Hindi handwritten character recognition using oriented gradients and Hu-geometric moments. *Journal of Electronic Imaging*, *27*(5), 16-30.

[4]. Yadav, M., Purwar, R., & Mittal, M. (2018). Handwritten Hindi character recognition: a review. *IET Image Processing*, *12*(11), 1919-1933.

[5]. Elzobi, M., Al-Hamadi, A., Al Aghbari, Z., & Dings, L. (2012). IESK-ArDB: a database for handwritten Arabic and an optimized topological segmentation approach. *International Journal on Document Analysis and Recognition (IJDAR)*, *16*(3), 295-308.

[6]. Liu, C., Yin, F., Wang, D., & Wang, Q. (2011). CASIA Online and Offline Chinese Handwriting Databases. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition* (pp. 37–41). Washington, DC: IEEE.

[7]. Bhattacharya, U., & Chaudhuri, B. (2005). Databases for Research on Recognition of Handwritten Characters of Indian Scripts. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition* (pp. 789–793). NW Washington: IEEE.

[8]. Sharma, N., Pal, U., Kimura, F., & Pal, S. (2006). Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier. In *5th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP* (pp. 805-816). Madurai: Springer.

[9]. Bhattacharya, U., & Chaudhuri, B. (2009). Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, *31*(3), 444-457.

[10]. Dongre, V., & Mankar, V. (2012). Development of Comprehensive Devnagari Numeral and Character Database for Offline Handwritten Character Recognition. *Applied Computational Intelligence And Soft Computing*, 1-5.

[11]. Khanduja, D., Nain, N., & Panwar, S. (2016). A Hybrid Feature Extraction Algorithm for Devanagari Script. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *15*(1), 1-10.

[12]. Verma, B. (1995). Handwritten Hindi character recognition using multilayer perceptron and radial basis function neural networks. In *International Conference on Neural Networks* (pp. 2111-2115). Perth: IEEE.

[13]. Sarkhel, R., Das, N., Das, A., Kundu, M., & Nasipuri, M. (2017). A multi-scale deep quad tree based feature extraction method for the recognition of isolated handwritten characters of popular indic scripts. *Pattern Recognition, 71, 78–93.*

[14]. Verma, G. K., Prasad, S., & Kumar, P. (2011, March). Handwritten Hindi character recognition using curvelet transform. In *International Conference on Information Systems for Indian Languages* (pp. 224-227). Springer, Berlin, Heidelberg.

[15]. Rojatkar, D., Chinchkhede, K., & Sarate, G. (2013). Handwritten Devnagari consonants recognition using MLPNN with five fold cross validation. In *International Conference on Circuits, Power and Computing Technologies (ICCPCT)* (pp. 1222-1226). Nagercoil: IEEE.

[16]. Yadav, M. & Purwar, R. (2017). Handwritten hindi character recognition using multiple classifiers, In *International Conference on Cloud Computing, Data science and Engineering*, 12–13.

How to cite this article: Yadav, M. and Purwar, R. K. (2020). Synthetic Data for Hindi Character Recognition. *International Journal on Emerging Technologies*, *11*(2): 793–796.