Multi-lingual handwritten character recognition using Deep Learning

Harikesh Pandey[1,*], Arun Prakash Agrawal[1]

[1] *Department of Computer Science, Sharda University, Greater Noida, India, 201306*

Abstract

The most difficult activity is handwritten character recognition (HCR), which is difficult be-cause it requires repeated human labor and is subject to human error. It is the most difficult task to complete in any language due to the comparable types of characters and various types of writing styles. However, we may address this problem with picture classification and deep learning algorithms. There are numerous models for character recognition in a single language, but there is no model for character recognition in multiple languages. Model performance will decline as the number of classes rises. For HCR, a variety of methods are available.

Convolutional Neural Network (CNN), a cutting-edge model for image categorization, is frequently utilized for this issue. In this, a brand-new architecture is suggested. By altering the hyper-parameters and selecting the proper activation function the proposed design was made more precise. There are three distinct openly available datasets, including English, Hindi, Bengali characters, and numbers were used to evaluate the proposed system. For uni-lingual models, we have attained accuracy of 95.54%, 93.60%, 91.58%, and 95.32%, respectively. Our accuracy rate for the suggested multilingual model is 92.47%. It can be seen from a comparative analysis with the current approaches that the performance of the suggested multilingual model has significantly improved.

Keywords

Convolutional Neural Network (CNN), Deep Learning (DL), handwritten character recognition (HCR), multilingual, symbols [1]

Symposium on Computing & Intelligent Systems (SCI), *May 10, 2024, New Delhi, INDIA*

* Corresponding author.

† These authors contributed equally.

harikesh.pandey@gmail.com (H. Pandey); arun.agrawal@sharda.ac.in (A. Agrawal)

1. Introduction

There are several reasons why it is vital to recognize handwritten characters. It is incredibly challenging to look through thousands of paper bundles to find a file. However, in many places, papers with handwriting are manually typed and input into the system making digitization of the files into the systems crucial. So, the easiest solution to save time and effort and provide masterpiece with a smaller mistake percentage is to develop a model that can recognize the handwritten letters. Handwritten character recognition (HCR) has several uses, including reading cheque amounts, postal addresses, handwritten correspondence, and submitted forms. The goal is to do the assignment accurately and efficiently. So utilizing deep neural learning to identify the pattern in this situation.

By using an ensemble approach and a deep learning strategy, we suggest a multilingual model that is able to distinguish the mathematical symbols and four distinct linguistic letters. It has been found that models based on convolutional neural networks (CNN) are effective for a variety of picture categorization tasks. There are numerous non-linear hidden layers in CNN, each with a substantial number of parameters and connections. Here improve the performance by adjusting the hyper-parameter, number of hidden layers, etc. Here we offer a CNN-based model that can recognize characters from many languages, including Hindi, English, Bangla, numerals in Bangla and Hindi, and all mathematical symbols. The dataset was compiled using online sources. For each dataset, we created a unique model, and we compared the outcomes to the methods already in use as described in the literature. We suggest an ensemble model to address the issues with the baseline models. The comprehensive trials conducted have demonstrated that the suggested model is accurate and quick even as the number of classes grows.

2. Literature Review

For Malayalam characters, Nair et al.[1] suggested a character recognition-based methodology. They gathered data from 112 distinct individuals for six handwritten Malayalam characters. They produced over 2 lakh photos by using image enhancement. In order to avoid overfitting and reduce model training time. Using CNN, Shiba Prasad Sen et al.[2] suggested for Bangla characters, they suggested a various feature map changes, max-pooling procedures, and additional activation functions 200 unique photos were gathered for the one-character class. The image is then processed and made into a 28 x 28 grayscale image (70% of the dataset is used for training and 30% for testing). Couple of convolution layers, pooling layers used and a fully connected classification layer with input and output layers are all included in their model [2]. A 28 x 28 image is present in the input node. The output of the first convolutional layer, which has 32 filters of the size 5 x 5, is a 24 x 24 matrix in their upcoming work, they use a huge database with more accuracy. A HCR method using Devanagari digits was proposed by Reena Dhakad et al., [3]. The dataset is implemented in MATLAB 7.0 along with the digit recognition system. For Chinese characters, Li Chen et al. suggested a model based on the HCR system [4], which made use of a CNN architecture. They used the MNIST and CASIA datasets in this study. Three components make up their model: sample creation, CNN models, and voting. For sample

generation, they used random distortion. Due to the combinations and more complicated characters, they used two separate CNN models.

A model based on a dataset of handwritten Devanagari characters that includes Hindi numerals was proposed by Shailesh Acharya et al., [5]. 92,000 distinct photos from 46 different classifications make up this collection. Training data (85%) and testing data (15%) are separated into two sets. 32 x 32 pixels make up each image. There are 784 neurons and 4 filters in the first convolutional layer. Each unit's input weight for the 5 x 5 size kernel is 25 and each unit has trainable bias. The 2 x 2 size pooling layer is then connected to it, which lowers the resolution. Numerous learning algorithms, such as clustering techniques and neural learning, were applied for handwritten character identification in Hindi and English [10–13]. The efforts are described in [22–26] for character recognition in Tamil, Arabic, and other languages as well.

Using data augmentation, Rashid Chowdhury et al., [18] stated using Bangla handwriting, Bangla lekha isolated is the dataset that was used in this model. It has 10 Bangla numeral digits, 24 compound characters, and 50 basic characters. One class of photographs contains 2,000 pictures. Following the conversion of these photos into a CSV file, the pixel values are additionally standardized. Furthermore, 10% of the data is used for validation [18]. Without adding further data, they maintain 91.81% accuracy [18]. With 70 epochs after data augmentation, they achieved 95% accuracy. A study based on images utilizing the CNN-ECOC (Error Correcting Output Code) classifier was presented by Mayur Bhargab Bora et al. [19].The outcome will got 88% training and 93% testing accuracy when the code word is compared to the current NIST dataset. A study on Gujarati HCR was presented by Jyoti Pareek et al. [20]. They gathered roughly 10,000 photos and 59 classes of the handwritten data. Then, it is separated into 80% training and 20% testing for Medical Care [20]. Deep CNN was used in a study on handwritten Bengali character identification by Suprabhat Maity et al. [21]. 11 vowels and 39 consonants from the Bengali character set are used, and 12k images are used for training and 3k for testing [21]. Roy et al., (2022) [27] This study used a Convolutional Neural Network (CNN) based method to recognize handwritten multilingual multiscript in both single and multiscript scenarios, taking into account scripts written in Bangla, Devanagari, and English, obtained results up to 93.29% in English. Agilandeeswari et al., (2023) [28] propose to enhance handwritten character recognition accuracy, a unique deep LSTM technique is presented in this paper. It is composed of the following stages: CNN feature extraction, an LSTM sequence of layers, and a Connectionist Temporal Classification (CTC) decoder.

3. Proposed CGM model

This section contains a complete explanation of the model as well as a presentation of CNN's work. The CNN architecture is used for this character recognition problem because, unlike other MLPs, it works in a layered architecture with more reliable and computationally modest than other MLPs. Three concepts, local connections, layering, and spatial invariance, are the key sources of inspiration for Deep CNN. To prepare the dataset for this model's input, all of the photos are transformed to the same dimensions. All of the input photos in this situation are 32 by 32, making it simple to feed them into the input

layer. The dataset images used in the CNN architecture are initially processed convolutional layer and determines the weights (feature map) for a specific image inside the specified field by computing with filters dot product image. The pattern is then learned from the images of a single character using an activation function that is applied in conjunction with it

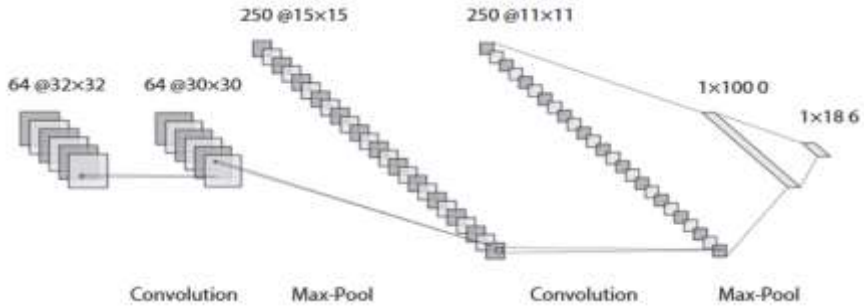

Figure 1. Architecture of the proposed approach

Different sources, including English, Hindi, and Bangla and numbers character recognition for different languages is the foundation of this concept. Each data set is assessed independently and compared to the existing models, demonstrating that our model is superior to the latter. Finally, we aggregate all datasets to assess the model's performance shown in Figure 1.

4. Experimentation

The performance of the proposed model was assessed using a single language dataset (Hindi, English and Bangla) for the first four experiments, and all dataset classes were combined for the final experiment, which allowed us to assess multilingual characters.

4.1. Data collection and preprocessing

English number and English (capital and tiny) characters symbols [6], dataset comprises of 45 x 45 pixel jpg files. There are 375,974 photos in this dataset. Figure 2 displays the sample dataset. The dataset2 of handwritten Bangla characters contains 10 Bangla numbers, 24 compound characters, and 50 basic characters [7]. There are 84 classes in total in this collection. Only 50 fundamental characters and 10 Bangla numerals have been considered. In Figure 3, the sample dataset is displayed. Collection of handwritten Devanagari characters from the Kaggle database, 3 pictures were taken dataset has 45 classes, including Hindi consonants and numbers [8]. This collection has 92,000 photos in total. Figure 4 displays the example dataset.

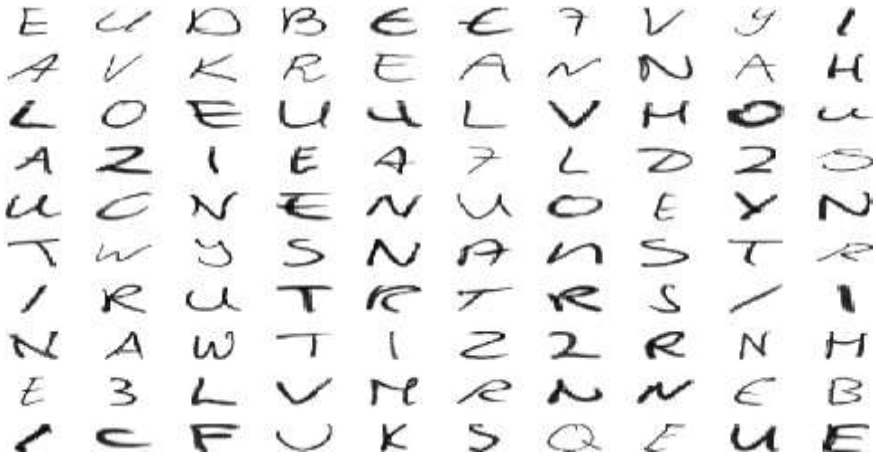

Figure 2. Sample English characters

Figure 3. Sample Bangla characters

Figure 4. Sample Devanagari characters

4.2. Dataset Distribution

The CNN model's architecture is displayed in this section. Both the combined and individual models utilize the same architecture. The integrated model for character

recognition has 186 classes in all. Figure1 depicts the model's architecture. A first convolution layer is applied to images of size 32 by 32, with kernel size 3 by 3 and feature size 64. Max pooling layer, which has a 2 2 size, is the following layer, and it receives tanh activation function. The output of this layer is fed into a second next layer with a feature size of 250 and a kernel size of 5 5, after which a second max pooling layer with a 2 2 grid and tanh activation function is applied. The result was often passed through different process and layers as passed work to the heavy layer and soft-max. This model was developed using the same suggested system architecture for English HCR, the 36-class dataset (10 integers and 26 alphabets) is split into three sections: 70% (151,749) of the data were used for training, 15% (32,507) were used for validation, and 15% (32,553) were used for testing [6]. Each class has a varied number of photos, and Figure.5 displays the dataset distribution. The training data set photos are all scaled to 32 32 because all of the images are in ".jpg" format and are 45 45 in size.

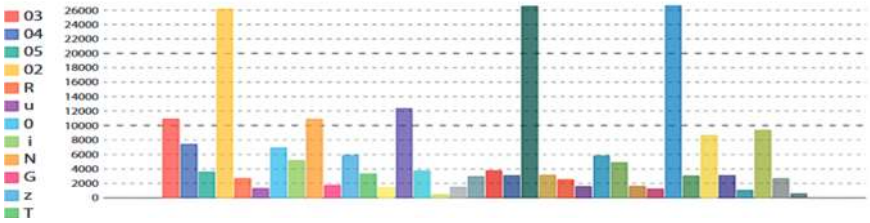

Figure 5. Dataset distribution for English dataset

This model was developed for Hindi HCR using the same suggested system architecture. The dataset, which is broken up into three portions, has 46 classes, including 36 basic letters and 10 Hindi numerals. 70% of the data (78,200 photos) were used for training, 15% for validation (13,800 images), and 15% for testing (13,800 images) [8]. Every class has the same amount of photos (2,000), and Figure 6 displays the dataset distribution.

Figure 6. Dataset distribution for Hindi dataset

4.3. Combined Model

The dataset under consideration has 186 classes (the class labeled "0" is the same in Bangla, Hindi and English, hence concatenated into a class) is split into three sections: 70% of the data (including 407,819) were used for training, 15% were used for validation, and

15% were used for 87,527 photos data to be tested. We get 93.46% validation accuracy and 92.47% test accuracy.

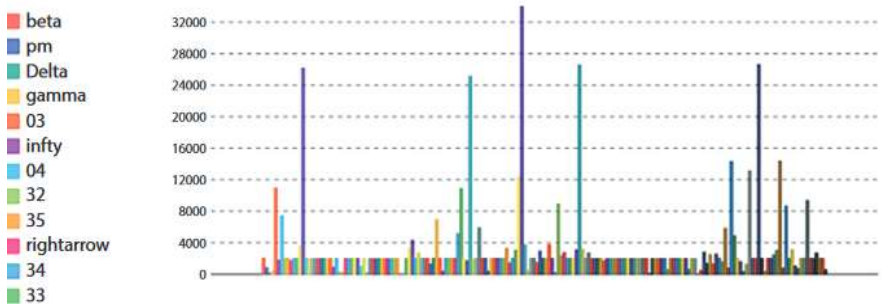

Figure 7. Dataset distribution

5. Results and Discussion

Each uni-lingual model is assessed separately in this section. For each model, all models are also evaluated against other models to show how well they perform. Finally, a multilingual model is compared to each single language model. It demonstrates that the suggested multi-lingual model strategy is the best to utilize, outperforms dissimilar models, and eliminates the demand for specific style.

5.1. Execution of Uni-Lingual Model on separate Dataset

Table 1. English dataset

Pattern	Precision	Records
Support vector (SVM)	92.75	TD3
CNN	95.54	HMS

Table 2. Hindi dataset.

Pattern	Precision	Records
Support vector (SVM)	91.75	TD3
CNN	93.60	HMS

Table 3. Bangla dataset

Pattern	Precision	Records
Multilayer Perceptron(MLP)	84.75	BHCR
CNN	91.58	BHCR

5.2. Execution of Multi-Lingual Model on integrated Dataset

While this model can recognize characters from several languages as English Hindi and Bangla. This comparison demonstrates that our model outperforms all mono-language models in terms of accuracy and performance. Our accuracy rate for the suggested multilingual model is 92.47%.

6. Conclusion

Many different methodologies and approaches for the recognition of handwritten characters have been proposed. The following problems are present in all of the suggested solutions: if we us writes in numerous language the system is incapable of grasp the characters from the further language additionally, there is no global model that is language-neutral; existing unimodel techniques concentrate on different languages.

In this study, we offer a CNN-based approach that can recognize characters from many languages, including Hindi, English, Bangla, and numbers. The dataset was compiled using online sources. On different 3 sets—the Hindi character, the Bangla character and the English record set, we tested our methodology. For every dataset, we created a unique model, and we compared the outcomes to the methods previously described in the literature. We suggest an ensemble model to address the issues with the baseline models. The comprehensive experiments conducted have demonstrated that the suggested model is quick and precise even as increase the number of levels, the proposed models demonstrated superior performance.

References

[1] Nair, P.P., James, A., Saravanan, C., Malayalam Handwritten Character Recogni-tion using Convolutional Neural Network. International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017.

[2] Sen, S.P., Shaoo, D., Paul, S., Sarkar, R., Roy, K., Online Handwritten Bangla Character Recognition Using CNN: A Deep Learning Approach. Part of the Ad-vances in Intelligent Systems and Computing book series AISC, vol. 695, 2018.

[3] Dhakad, R. and Soni, D., Devanagari Digit Recognition by using Artificial Neural Network. 2017 International Conference on Energy, Communication, Data Analyt-ics and Soft Computing ICECDS, 2017.

[4] Chen, L., Wang, S., Fan, W., Sun, J., Naoi, S., Beyond human recognition: A CNN-based framework for Handwritten Character Recognition. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015.
[5] Acharya, S., Pant, A.K., Gyawali, P.K., Deep Learning Based Large-Scale Hand-written Devanagari Character Recognition. 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2015.
[6] https://www.kaggle.com/xainano/handwrittenmathsymbols#data.rar
[7] https://data.mendeley.com/datasets/hf6sf8zrkc/2
[8] https://www.kaggle.com/ashokpant/devanagari-character-dataset-large
[9] Nasien, D., Haron, H., Yuhaniz, S.S., Support Vector Machine (SVM) For English Handwritten Character Recognition. 2010 Second International Conference on Computer Engineering and Applications, 2010.
[10] Yuan, A., Bai, G., Jiao, L., Liu, Y. Offline handwritten English character recogni-tion based on convolutional neural network, in: 2012 10th IAPR International Workshop on Document Analysis Systems, pp. 125-129, Gold Coast, QLD, Aus-tralia, 2012.
[11] Gaur, A. and Yadav, S., Handwritten Hindi character recognition using k-means clustering and SVM. 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, 2015.
[12] Singh, N., An Efficient Approach for Handwritten Devanagari Character Recogni-tion based on Artificial Neural Network. 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), 2018.
[13] Chaudhary, D. and Sharma, K., Hindi Handwritten Character Recognition using Deep Convolution Neural Network. 2019 6th International Conference on Compu-ting for Sustainable Global Development, 961–965, 2019.
[14] Rahman, Md. M., M.A.H., Akhand, Islam, S., Shil, P.C., Bangla Handwritten Char-acter Recognition using Convolutional Neural Network. I.J. Image, Graphics and Signal Processing, 2015.
[15] Basu, S., Das, N., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.K., A hier- archical approach to recognition of handwritten Bangla characters. Pattern Recognit., 42, 1467–1484, 2009.
[16] Bhowmik, T.K., Ghanty, P., Roy, A., Parui, S.K., SVM based hierarchical architec-tures for handwritten Bangla character recognition. Int. J. Doc. Anal. Recogn., 12, 2, 97–108, 2009, 2009.
[17] Drsouza, L. and Mascarenhas, M., Offline Handwritten Mathematical Expression Recognition using Convolutional Neural Network. 2018 International Conference on Information, Communication, Engineering and Technology (ICICET), 2018.
[18] 18. Chowdhury, R.R., Hossain, M.S., Ul Islam, R., Andersson, K., Hossainrsouza, S., Bangla Handwritten Character Recognition using Convolutional Neural Network with Data Augmentation. Joint 2019 8th International Conference on Informatics, Electronics & Vision (ICIEV) & 3rd International Conference on Imaging, Vision & Pattern Recognition (IVPR), 2019.
[19] 19. Bora, M.B., Daimary, D., Amitab, K., Kandar, D., Bangla Handwritten Character Recognition from images using CNN-ECOC. International Conference on Compu-tational Intelligence and Data Science ICCIDS, 2019.

[20] Pareek, J., Singhania, D., Kumari, R.R., Purohit, S., Gujarati Handwritten Character Recognition from Text Images. Third International Conference on Computing and Network Communications (CoCoNet'19), 2019.

[21] Maity, S., Dey, A., Chowdhury, A., Banerjee, A., Handwritten Bengali Character Recognition Using Deep Convolution Neural Network. MIND 2020: Machine Learning, Image Processing, Network Security and Data Sciences, pp. 84–92, 2020.

[22] Gondere, M.S., Schmidt-Thieme, L., Boltena, A.S., Jomaa, H.S., Handwritten Am-haric Character Recognition Using a Convolutional Neural Network. ECDA2019 Conference Oral Presentation, 2019.

[23] Ptucha, R., Such, F.P., Pillai, S., Brockler, F., Singh, V., Hutkowski, P.,Intelligent character recognition using fully convolutional neural networks. Pattern Recognit., 88, 604–613, 2019.

[24] Boufenar, C., Kerboua, A., Batouche, M., Investigation on deep learning for off-line handwritten Arabic character recognition. Cognit. Syst. Res., 50, 180–195, 2018.

[25] Ram, S., Gupta, S., Agarwal, B., Devanagri character recognition model using deep convolution neural network. J. Stat. Manage. Syst., 21, 593–599, 2018.

[26] Kavitha, B.R. and Srimathi, C., Benchmarking on offline Handwritten Tamil Char-acter Recognition using convolutional neural networks. J. King Saud Univ. - Comp. Info. Sci., 2019, (In Press)

[27] Roy, R.K., Mukherjee, H., Roy, K. et al. (2022)" CNN based recognition of hand-written multilingual city names". Multimed Tools Appl 81, 11501–11517 (2022). https://doi.org/10.1007/s11042-022-12193-827.

[28] Agilandeeswari, L., Swarup, S.K., Thrishala, T.V. and Kola, S., 2023, March. Handwritten Character Recognition Using Deep LSTM Approach. In Proceedings of the 14th International Conference on Soft Computing and Pattern Recogni-tion (pp. 120-128). Cham: Springer