Chapter 39
Classification of Bangla Compound Characters Using a HOG-CNN Hybrid Model

S. M. A. Sharif, Nabeel Mohammed, Sifat Momen and Nafees Mansoor

Abstract Automatic handwriting recognition is challenging task due to its sheer variety of acceptable stylistic differences. This is especially true for scripts with large character sets. Bangla, the sixth most widely spoken language in the world has a complex, large and rich set of compound characters. In this study, a hybrid deep learning model is proposed which combines the use of the manually designed feature Histogram of Oriented Gradients (HOG), with the adaptively learned features of a Convolutional Neural Networks (CNN). The proposed hybrid model was trained on the CMATERDB 3.1.3.3, a Bangla compound character data set which divides Bangla compound characters into 177 broad classes and 199 specific classes. The results demonstrate that CNN-only models achieve over 91% and 92% test accuracy respectively. Furthermore, it is shown that the proposed model, which incorporates HOG features with a CNN, achieves over 92.50% test accuracy on each division. While there is still room for improvement, these results are significantly better than currently published state of art on this data set.

Keywords Bangla handwriting recognition · Bangla compound characters Convolutional Neural Networks (CNN) · Deep learning · Histogram of Oriented Gradients (HOG) · CMATERDB 3.1.3.3

S. M. A. Sharif (✉) · N. Mohammed · S. Momen · N. Mansoor
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh
e-mail: sma.sharif.cse@ulab.edu.bd

N. Mohammed
e-mail: nabeel.mohammed@ulab.edu.bd

S. Momen
e-mail: sifat.momen@ulab.edu.bd

N. Mansoor
e-mail: nafees.mansoor@ulab.edu.bd

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_39

Introduction

Bangla language has a rich set of compound characters which are formed by combining multiple individual characters. While they do not appear as frequently as individual vowels and consonants, these compound characters have an important role in day to day verbal communication as well as in writing scripts [5]. An exhaustive study [4] identified over 334 compound characters with 171 pattern classes. Some of these pattern classes even contain multiple shapes. Such large number of pattern classes, some with subtle shape changes, make Bangla compound character classification a challenging task.

Handwriting recognition is a part of the image classification domain. Two dominant approaches to image classification can be observed from recent studies. The first approach uses predefined and/or hand-crafted features, e.g. SIFT, HOG etc., extracted either locally or globally (or both) which are then used to train a classifier e.g. SVM, K-Means, GA-based etc. The other approach is popularly known as the deep learning approach, where convolutional neural networks (CNN) are trained with images and their corresponding labels. These networks are remarkable because they learn useful features during the training phase.

Both approaches have been employed successfully to the task of handwritten character classification. This paper reports on the application of a Hybrid model, which combines CNNs and the popular Histogram of Oriented Gradients (HOG) image feature, to the task of classifying Bangla compound characters. It is shown that the proposed model achieves over 92.5% test accuracy in classifying Bangla compound characters, which appreciably better than the best reported results so far.

The rest of the paper is organised as follows: Section "Background" discusses the CMATERDB 3.1.3.3 data set, CNN, and HOG. Section "Proposed Method" describes the proposed hybrid model. Section "Experimental Setup" details the experiments performed for this study. Section "Results" presents the results and section "Conclusion" concludes the paper.

Background

More than 230 million people speak Bangla all over the world. It is the sixth most widely spoken language and is the second most widely spoken language in the Indian subcontinent. The Bangla writing script has a set of vowels, consonants as well as compound characters. Compound characters are composed of two or more consonants and at times also a vowel. The vocalisation of these characters is typically done by the simultaneous pronunciation of the individual characters. The shapes of these compound characters are necessarily complex and in some cases, the individual original characters are not recognisable in the final shape. Some examples compound characters are shown in Fig. 39.1.

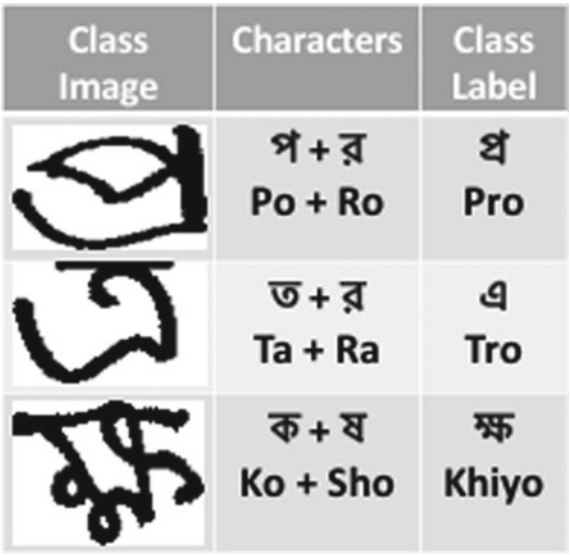

Fig. 39.1 Examples of Bangla compound characters from CMATERDB 3.1.3.3 model structure

The work in [4] documents the CMATERDB 3.1.3.3 handwritten Bangla compound character data set. The study performed a thorough automated analysis of 2.4 million words collected from three Bangla newspapers to identify 171 different compound character classes. These 171 classes do no include individual consonants with vowel allographs. Some compound characters have multiple visually distinct writing patterns. If these individual patterns are separated, then the total number of compound character classes rise to 199. Currently, the data set contains 44,152 training images and 11,126 testing images with annotations for 171 and 199 classes. The number of samples per character class is not equal and vary between 125 and 474 samples. Das et al. [3] applied a GA-and SVM-based approach to classifying this data set and achieved an impressive test accuracy of 78.93%. Das et al. [4] used a Quad-tree-based approach and reported an improved test accuracy of 79.35%. In both cases, the features extracted from the images were predetermined and not learnt during the training phase, as convolutional neural networks are designed to do.

A breakthrough study [6] showed that convolutional neural networks (CNN) [7] can be successfully applied to image classifications tasks with a very large number of classes. These networks have multiple layers of filters, and pooling/subsampling layers. The filters are typically two or three dimensional, depending on whether the image is gray scale or colour, and are used to extract image features by convolving them with the image. Unlike traditional systems, these filters are not predefined by the researcher/engineer but are learnt by minute adjustments during the training phase. Nonlinearities (i.e. tanh, RelU, sigmoid), are usually applied to the output of the convolutional layers to allow these networks to learn complex nonlinear functions. The output of the convolutional layers are typically two/three dimensional,

which are then flattened and passed through one or more fully connected layers for image classification.

There has been some recent CNN-based work on Bangla handwriting recognition . For instance, [10] employed a five-layer network similar to the popular LeNet architecture, which has four convolutional and pooling layers. The last layer is labelled as the F5 layer which extracts a feature vector of 300 dimensions. The work aimed to achieve a generalised character recognition feature extractor by training this network on a data set of 50 character classes. The features extracted by the trained model was then used to classify characters from other data sets by classifying the features using an SVM. This method yielded an accuracy rate of 98.375% on the test partition of the ISI Bangla numeral data set. Even more recently, a smaller CNN was used by Akhand and Mahtab Ahmad [9] to train on and classify the ISI Bangla numeral data set. This study reported accuracy rates of 98.98% on the test partition of the data set. The training set of 19, 392 images was augmented by rotating the images by fixed angles. Different networks were trained for different augmentation angles ($5°, 10°, 20°$ and $30°$). The best testing accuracy rate of 98.98% was reported when a $10°$ rotation angle was used. However, this method is equivalent to fine-tuning the training process to learn features which are applicable on the validation set, instead of generalised learning.

Most of these studies rely entirely on "learnt features" extracted by CNNs, and do not take into account the role of manually designed features. This paper presents the use of a hybrid approach to designing a network, which aims to bridge this gap. The proposed approach combines, in a simple way, the "learnt features" of a CNN with the very effective, manually designed, Histogram of Oriented Gradients (HOG) feature.

Histogram of Oriented Gradients (HOG) [2], is a feature which emanates from a different philosophy. Before the recent popularity of deep learning, features such as HOG, SIFT [8], etc.... were considered state of the art. These features were designed through the hard work and ingenuity of the researchers and were typically used in conjunction with linear classifiers, e.g. SVM. HOG features are particularly interesting because of their simplicity and effectiveness at distinguishing shapes [1]. It utilises image gradient information extracted from images after subdividing the image into small regions which are referred to as "cells". The gradients at sampled pixels within a cell are used to compute a histogram. The combined histogram of all the cells becomes the image descriptor. For gradient calculation, applying the simple masks: $[-1, 0, 1]$ and $[-1, 0, 1]^T$ on a non-smoothed image works best. The image is subdivided into cells of configurable size and an orientation histogram of the gradient values is calculated for every cell. For contrast normalisation, particularly, when cells sizes are small, multiple cell histograms can be normalised together.

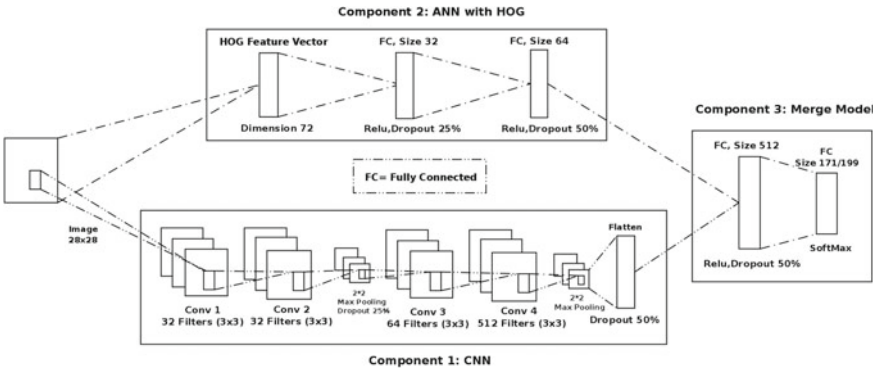

Fig. 39.2 Proposed hybrid model

Proposed Method

The aim is to integrate HOG features in the learning and classification process of a deep learning model. Figure 39.2 shows the scheme of the proposed model which has three components and two input layers.

The first component is comprised of a single CNN, which accepts a 28×28 grayscale image as input and successively applies convolution and pooling operations. The network has two convolution layers, each with 32 3×3 filters. The two convolutional layers are followed by a max-pooling layer, which is in turn followed by two more convolutional layer with 64 and 512 3×3 filters, respectively. The output of this last convolutional layer is a stack of two-dimensional filter outputs. These values of the filter outputs are unrolled, to form a vector input to further fully connected layers.

A fully connected layer which expects a HOG image descriptor as input forms the second part of the model. HOG features are extracted from each image with a cell size of 8 pixels, with a 8 bin histogram of edges created at each cell. The histograms are then concatenated to form the final HOG image descriptor. This results in a feature vector of 72 dimensions, which is what this part of the proposed structure expects as input. This component has two hidden layers of 32 and 64 dimensions, respectively. The output of the last layer as a feature vector of 32 dimensions and is concatenated with the output of the first component.

The concatenated feature vector is then used as input to a third fully connected network which has a hidden layer of 512 dimensions. The output layer has 171 or 199 dimensions depending on which classification scheme is used in a particular experiment.

All three components are created as a part of a single network and trained end-to-end together. As shown in Fig. 39.2, the Dropout [12] rates used are quite high; this was done to ensure that the network does not over train as it has a large number of parameters.

Experimental Setup

Data Set Preparation

This study uses the aforementioned CMATERDB 3.1.3.3 Bangla compound character collection. The training set of the collection is augmented [11] by randomly rotating between $-50°$ and $50°$. The original images have black writing on a white background; these are processed so that the background is black and the character is in white. All images are resized to be 28×28. The aspect ratio of the written characters is not preserved.

Models Used

In total, four different models were used to facilitate comparison using either a CNN-only approach or the proposed hybrid approach.

The two CNN-only models are labelled CNN-171 and CNN-199, respectively, to indicate whether they are classifying 171 classes or 199 classes. The networks have a similar structure to the first component of the hybrid model, except the output of the flattened layer is directly used in a fully connected network whose structure is similar to the third component.

The Hybrid models are labelled Hybrid-Hog-171 and Hybrid-HOG-199, respectively, to indicate whether they are classifying 171 classes or 199 classes. The only place where they differ is in the dimensionality of the last softmax layer (of the third component).

Model Training

Each model was trained on the augmented training set for 100 epochs. At the end of each epoch, the models performance on the test data set was measured and the best performing model, over the 100 epochs, was kept. The training batch size was set to 100. The Adadelta [13] optimiser was used to minimise the categorical cross entropy loss values. All training was done on an a machine running Ubuntu 14.04 while taking advantage of an NVIDIA GTX 980 TI GPU to speed up the experiments.

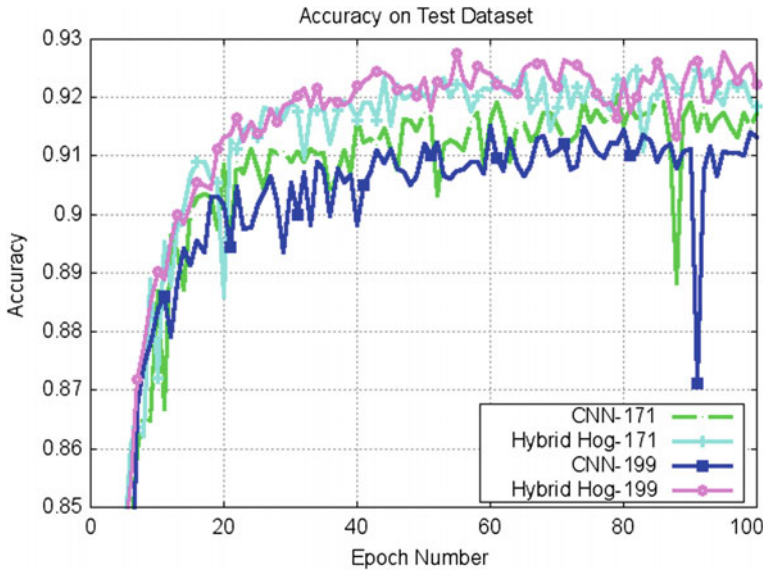

Fig. 39.3 Accuracy of the four models on the test data set during training

Table 39.1 Best validation accuracy and training epochs required

Experiment name	Accuracy (%)
CNN-171	92.05
CNN-199	91.53
Hybrid-Hog-171	*92.57*
Hybrid-Hog-199	*92.77*

Results

Figure 39.3 shows how the accuracy of the test set varied during training. Although only the best performing models were saved, it is evident from the comparison that the hybrid models consistently outperform the CNN-only models.

Table 39.1 shows the best performing test accuracy rates of the four models used in the experiments, and the epoch at which the rates were achieved. The hybrid models achieve improvements of 0.52 and 1.24% on the 171 and 199 classifications when compared to the CNN-only models.

Table 39.2 compares the performance of the proposed method with those which have previously been published. The proposed approach achieves a significant improvement over previous work, with improvements of 13.22% and 14.1% over the 171 and 199 classes respectively.

The proposed models actually achieve a higher accuracy on the 199 classes compared to the 171 classes. While this might seem counter-intuitive initially; the

Table 39.2 Comparison with published results

Work	Year	Feature selection	Classification	Number of classes	Test accuracy (%)
Nibaran Das et al. [3]	2012	GA	SVM	171	78.93
Nibaran Das et al. [4]	2014	CH,QTLR	SVM	171	79.35
Nibaran Das et al. [4]	2014	CH,QTLR	SVM	199	78.67
Hybrid-Hog-171 (Proposed)	2016	Hog-Adaptive	CNN-ANN	171	*92.57*
Hybrid-Hog-199 (Proposed)	2016	Hog-Adaptive	CNN-ANN	199	*92.77*

behaviour is easily explained when taking into account the nature of the class separations. The 171 classification problem includes cases, where the different patterns for the same shape is assigned to the same class. For the 199 classification problem, this added difficulty is not present.

Conclusion

This paper presents a Hybrid deep learning model which combines a convolutional neural network with HOG features. The proposed model was trained end-to-end using an augmented version of the training image set provided by the CMATERDB 3.1.3.3 collection. After training, the model achieves test accuracy rates of 92.57% and 92.77% on the 171 and 199 classes of the data set, respectively. These results are an improvement of over 13% and 14%, respectively, in accuracy rates compared to previous studies, which is significant. In future, multiple features are planned to be incorporated in such Hybrid models to study their effect.

Acknowledgements This work was supported by the ICT division of Ministry of ICT, Bangladesh [Grant number 56.00.0000.028.33.066.16-731].

References

1. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. vol. 2, p. 8 (2011)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
3. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A novel ga-svm based multistage approach for recognition of handwritten bangla compound characters. In: Proceedings of the International Conference on Information Systems Design and Intelligent

Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. pp. 145–152. Springer (2012)
4. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A benchmark image database of isolated bangla handwritten compound characters. International Journal on Document Analysis and Recognition (IJDAR) 17(4), 413–431 (2014)
5. Das, N., Basu, S., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.: Handwritten bangla compound character recognition: Potential challenges and probable solution. In: IICAI. pp. 1901–1913 (2009)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
7. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10), 1995 (1995)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
9. M. A. H. Akhand, Mahtab Ahmad, M.M.H.R.: Convolutional neural network training with artificial pattern for bangla handwritten numeral recognition. ICIEB 1(1), 1–6 (2016)
10. Maitra, D.S., Bhattacharya, U., Parui, S.K.: Cnn based common approach to handwritten character recognition of multiple scripts. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 1021–1025. IEEE (2015)
11. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813 (2014)
12. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1), 1929–1958 (2014)
13. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)