Convolutional Neural Networks for the Recognition of Malayalam Characters

R. Anil, K. Manjusha, S. Sachin Kumar, and K. P. Soman

Centre for Excellence in Computational Engineering and Networking,
Amrita Vishwa Vidyapeetham,
Coimbatore - 641112, India
{anil.soubhagia,manjushagecpkd sachinnme}@gmail.com,
kp_soman@amrita.edu

Abstract. Optical Character Recognition (OCR) has an important role in information retrieval which converts scanned documents into machine editable and searchable text formats. This work is focussing on the recognition part of OCR. LeNet-5, a Convolutional Neural Network (CNN) trained with gradient based learning and backpropagation algorithm is used for classification of Malayalam character images. Result obtained for multi-class classifier shows that CNN performance is dropping down when the number of classes exceeds range of 40. Accuracy is improved by grouping misclassified characters together. Without grouping, CNN is giving an average accuracy of 75% and after grouping the performance is improved upto 92%. Inner level classification is done using multi-class SVM which is giving an average accuracy in the range of 99-100%.

Keywords: Optical Character Recognition (OCR), Deep learning, Multi-stage classification, Convolutional Neural Network (CNN), Gradient based learning, Backpropagation, Multiclass SVM.

1 Introduction

In pattern recognition, Optical Character Recognition is an active and successful area for last few decades. There are many algorithms for pattern recognition. Intelligent Word Recognition (IWR), Optical Character Recognition (OCR), Optical Mark Recognition (OMR), Intelligent Character Recognition (ICR) is some common pattern recognition technologies, but OCR is most prelavent technology [1] [2]. The basic idea of Optical Character Regognition is the convertion of scanned bitmaps of printed or hand written text into data files which can be edited by machine. Using an OCR, a book or article can directly give to a computer to transform it to an editable text format. The ability to store text efficiently and productivity improvement by reducing human involvement are two major advantages of OCR system. The areas where this system can be used are banks, health care, postal departments, education, publication industry, finance, government agencies etc [2] [3].

S.C. Satapathy et al. (eds.), *Proc. of the 3rd Int. Conf. on Front. of Intell. Comput. (FICTA) 2014 – Vol. 2*, Advances in Intelligent Systems and Computing 328, DOI: 10.1007/978-3-319-12012-6_54

Any OCR system contains three main steps: 1. Preprocessing, 2. Feature Extraction and 3. Classification and Recognition. Preprocessing phase enhance and clean up the image by image binarization, noise removal, skew detection and correction, text segmentation etc. Feature extraction is a method used to measure some information from the data which is most relevant to the given classification task. The basic idea of classification is to map the segmented text portion in document image to equivalent text representation [1].

There are different softwares available for Optical Character Recognition. One of the open source OCR engine available is "Tesseract OCR Engine". Wide variety of image formats can read by this software and transform them to text in more than 60 languages. "OCRopus", is a Google sponsored project under IUPR research group. This OCR system use plugins for modular design. Another OCR system used for printed Malayalam documents is "Nayana", which is a product of Centre for Development of Advanced Computing(C-DAC) [4].

Good quality OCR systems are available for different languages like Latin, Arabic, and Chinese etc. Complete and efficient OCR systems for Indian language scripts are not widely available. Because they are more complex with large character set, irregular positioning of characters in a running text, structural similarity among character classes and existence of both old and new version of language scripts [3] [4]. In that Malayalam is one of the famous Dravidian languages. There are total 578 characters in Malayalam script and it contains 52 letters including 16 vowels and 36 consonants. The classification of Malayalam characters is a challenging task because this script is unicase, contains lots of similar structured characters and does not have any inherent symmetry [5].

Over the last few years, neural networks are the most researched area in pattern classification. Multi-layer neural networks trained with gradient decent have the ability to learn complex, high-dimensional, non-linear mappings from very large number of data which makes this a good solution for pattern recognition tasks. Multi layer neural networks with full connection are used as classifier in convolutional neural networks. CNN shows excellent recognition rates for digit recognition which is proposed by Lecun *et. al.*[6] and for Malayalam characters, compared to other classification methods CNN shows better results [2] [7].

In this paper Malayalam character recognition using multi stage classification architecture is presented. I.e primarily the recognition is done using CNN, called outer classification. Then in second stage the misclassified characters from outer classification is further classified using multiclass SVM, called inner classification. Second section discuss about CNN and learning method used in the network. Section 3 depicts a brief description about Multi-class SVM. Classification results are shown in Section 4.

2 Convolutional Neural Networks

In the area of machine learning, deep learning is a set of algorithms that is used for modelling high-level data abstractions by using architectures created with several

non-linear transformations. Neural Networks got importance because of their learning ability. In pattern recognition, designed neural network under goes two phases. The first phase is training phase, in which the network will first initialize the weights or trainable parameters with random values. Then the network will update the weights according to the error arouse with each sample provided by the training set. This type of learning method is called stochastic method of learning. In Batch learning weights gets updated only after considering all the samples present in the training set. Gradient descent backpropagation is one of the commonly used weight updation algorithm [7].

2.1 Gradient Based Learning

Gradient-based learning is one of the popularized methods used in automatic machine learning. Here learning machine computes a function

$$Y^p = F(Z^p, W). \tag{1}$$

Where Z^p is p th input pattern, W is collection of adjustable parameters in the system and Y^p is the output. Training set is created using this input vector Z^p and desired class label vector. By finding error function or loss function the amount of parameter adjustment, W can be identified. There are two error values e_{test} and e_{train} for test set and train set respectively. So the final aim is to minimize the value of $e_{test} - e_{train}$. This value indicates how the classifier performs on input patterns. Parameters $W = W(t)$, where t is the current training step, related to the gradient of the scalar-valued error function, can be modified using the equation

$$W(t+1) = W(t) - \varepsilon \frac{\partial E(W(t))}{\partial W(t)}. \tag{2}$$

Parameter can be modified after giving all input patters or after the presentation of one pattern [7] [8] [9].

2.2 Multilayer Perceptron

In the area of machine learning, perceptron is an algorithm used for supervised classification which will classify the input into several classes. Single perceptron is a computational model of the biological neuron. In single perceptron connections from other neurons are represented by the input vector l. Then all input vectors are multiplied with corresponding weights w and added. This scalar product is yielded with a threshold or bias, θ:

$$u = \sum_{k=1}^{n} l_k w_k - \theta = \langle l, w \rangle - \theta. \tag{3}$$

Then the perceptron's output is calculated by giving this scalar output to the transfer function or squashing function, i.e $y = f(u)$. For a multilayer perceptron Fermi function or inverse tangent is used as transfer function. Multilayer are capable of classsifying data which is not linear seperable. A multilayer perceptron is obtained by combining two or more layers together, where one layer is fully interconnected with the previous layer. It is also called feed-forward network because data is propagated forward through the network [7] [8] [9].

2.3 Backpropagation

Backpropagation algorithm is used for learning in neural networks. Gradient calculation is the main idea of backpropagation which is done by propagation from output to the input. The error function can be taken as

$$E(W) = \sum_{p} E^p(W) = \sum_{p} \left\| T^p - y^p \right\|_2^2. \tag{4}$$

Where y^p is the output vector from the last layer. To reduce the error rate of a pattern, gradient of E^p can be used to adjust the weights accordingly. Then the learning rule for neurons in the output layer can be calculated as

$$w_{ij}(t+1) = w_{ij}(t) + \varepsilon y_i \delta_j. \tag{5}$$

Where $\delta_j = (T_j - y_j) f'(u_j) y_i$, y_j - output of neuron j in the ouput layer, y_i - previous layer output of neuron i. Likewise hidden layers can also be learned. The weight adjustments carried out in hidden layer are responsible for all errors in the hidden layer. Here the error is given back to the hidden layer, so the name backpropagation [8] [9] [10].

2.4 Convolutional Neural Networks

A CNN has one or more convolutional layers which is having full connection. CNN also uses tied weights and pooling layers. LeNet-5 is a typical convolutional network for character recognition and which is used here. Le-Net 5 architecture contains 7 layers, without counting the input. Each unit in a particular layer accepts input from a set of units in the small neighborhood in the previous layer. In this architecture, convolutional layers and sub-sampling layers are organized alternatively. Starting with a convolutional layer, there are three convolutional layers, two sub-sampling layers and an output layer. The Output layer is composed of Euclidean Radial Basis

Function units (RBF), one for each class. That is each output RBF unit computes the Euclidean distance between its input vector and its parameter vector [7] [8] [9] [10].

3 Multiclass SVM

Support Vector Machine (SVM) is usually used for binary classification. One approach for multiclass SVM is building and combining several binary classifiers. Another way is considering all data in one optimization problem. The complexity to solve multiclass SVM problem varies with number of classes [11]. The earliest method for classification using multiclass SVM is one-against-all method. This will create k SVM models, where k is the number of classes. Another popular method is one-against-one method. Here $k(k-1)/2$ classifiers are created where each of them is trained on data from two classes. Directed Acyclic Graph Support Vector Machines (DAGSVM) is another approach for this problem. Training phase is same as one-against-one method. A rooted binary directed acyclic graph with $k(k-1)/2$ internal nodes and k leaves is used in testing phase [12] [13]. These methods can be implemented by using simple and efficient open source software called Libsvm. One-class-SVM, nu-SVM classification, C-SVM classification, nu-SVM regression and epsilon-SVM regression problems can be efficiently solved by this software.

4 Experimental Results and Discussion

This section describes about our attempt to apply the CNN based character classification approach to Malayalam language script. The experiments are conducted in MATLAB programming environment. LeNet-5 network is used for implementing CNN. 119 different character classes (including consonants, vowels, vowel modifiers and compound characters) of Malayalam Language script is used for our experiments. The dataset is created by scanning different Malayalam story, novel books and then segmenting each character images using levelsets [14].

At first for evaluating the performance of CNN on Malayalam language script, we have used an incremental training approach. We started with 5 classes for training first and calculated the accuracy. Then number of classes is slowly increased. When the number of classes reached range of 40 the performance of CNN started to decrease. CNN's performance was decreasing when we increased number of character classes beyond that. The misclassification rate of different set of trials (each set contains different 30 character classes) is shown in Table 1.

After this the network is learned using the 119 characters of Malayalam character dataset. The network shows an average training misclassification as 24% and for testing the misclassification rate is 25%. Therefore for whole Malayalam character dataset average classification accuracy is 76%. The Sample graph showing misclassification rate of set contain 119 characters is shown in Figure 1. It shows root mean square error in the above portion and misclassification rate for train and test set in the below portion. Root mean square error is the measure of difference between

predicted values and the actual observed values. Here the algorithm is trained for 5 iterations. Based on the confusion matrix obtained from this result, the characters which are misclassified mostly are grouped into classes. So the total number of classes is reduced to 85 from 119. The network is again learned and tested for these 85 classes. The result shows a performance improvement in the network. The misclassification rate is come down to 10.5% for testing and 11.9% for testing. A graph showing root mean square error and misclassification rate for a sample training and test set is shown in figure 2. Here the algorithm is trained for 10 iterations, but from iteration 5 itself the network shows a good accuracy.

Accuracy is improved when the classes are grouped together and trained. The grouped classes are further classified using multiclass SVM. The inner classification using multiclass SVM is giving accuracy in the range of 99-100%.

Table 1. Misclassification rate of different set of classes

Classes(each set 30 classes)	Error rate (train)	Error rate (test)
Set9	4.16667%	3.33333%
Set13	0.133333%	0%
Set14	0.6%	0.166667%
Set15	5.56667%	4.33333%
Set16	0.266667%	0%
Set17	9.83333%	6.55556%
Set18	10.9333%	6.22222%
Set19	8.4%	6.88889%

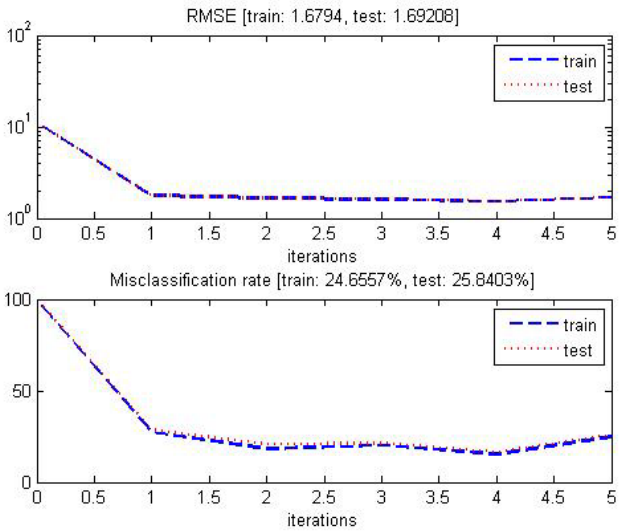

Fig. 1. Misclassification rate for set contain the 119 characters

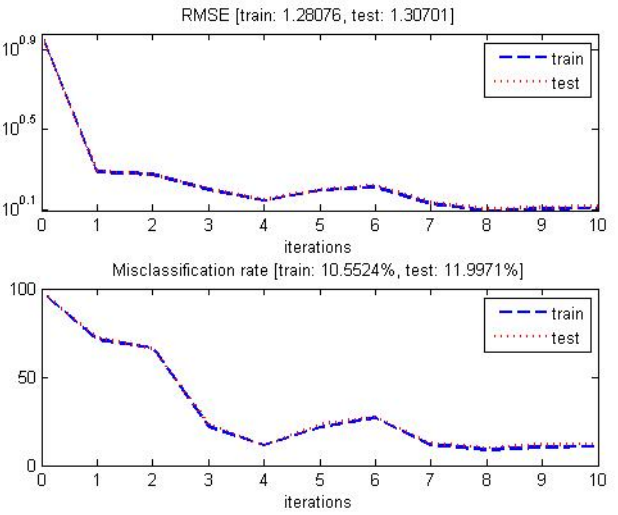

Fig. 2. Misclassification rate for set contain the grouped characters

5 Conclusion

Optical Character Recognition for Malayalam characters is an active research area which always needs an improvement in accuracy. This work is based on recognition of Malayalam characters using convolutional neural networks (CNN) and multiclass SVM. Compare to other deep learning architectures, CNN has better performance in both image and speech applications. To improve the performance of CNN, most misclassified characters are grouped and learned the network. The outer classification performance of proposed method is 92%. Multiclass SVM is used for inner classification and the accuracy is in the range of 99-100%.

References

1. Neeba, N.V., Namboodiri, A., Jawahar, C.V., Narayanan, P.J.: Recognition of Malayalam Documents. In: Advances in Pattern Recognition, Guide to OCR for Indic Scripts. Springer, London (2009)
2. Neeba, N.V., Jawahar, C.V.: Empirical evaluation of character classification schemes. In: Seventh International Conference on ICAPR 2009. Advances in Pattern Recognition, pp. 310–313. IEEE (2009)
3. Anil, R., Pradeep, A., Midhun, E.M., Manjusha, K.: Malayalam Character Recognition using Singular Value Decomposition. International Journal of Computer Applications (0975 – 8887) 92(12) (April 2014)
4. Divakaran, S.: Spectral Analysis of Projection Histogram for En- hancing Close matching character Recognition in Malayalam. International Journal of Computer Science and Information Technology (IJCSIT) 4(2) (April 2012)

5. Chaudhuri, B.B.: On OCR of a Printed Indian Script. In: Advances in Pattern Recognition (ed) Digital Document Processing. Springer, London (2007)
6. Lecun, Y.E.: Learning algorithms for classification: A comparison on handwritten digit recognition. Neural Networks: The Statistical Mechanics Perspective, 261–276 (1995)
7. Bouchain, D.: Character Recognition Using Convolutional Neural Networks. Institute for Neural Information Processing 2007 (2006)
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
9. Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: CNN: Single-label to Multi-label, arXiv preprint arXiv: 1406.5726 (2014)
10. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column Deep Neural Networks for Image Classification. In: Computer Vision and Pattern Recognition, pp. 3642–3649 (2012)
11. Soman, K.P., Loganathan, R., Ajay, V.: Machine Learning with SVM and other Kernel methods. PHI Learning Pvt. Ltd (2009)
12. Ramanathan, R., Arun, S., Nair, V., Vidhya Sagar, N.: A support vector machines approach for efficient facial expression recognition. In: International Conference on Advances in Recent Technologies in Communication and Computing, ARTCom 2009, pp. 850–854. IEEE (2009)
13. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2), 415–425 (2002)
14. Cherian, M., Radhika, G., Shajeesh, K.U., Soman, K.P., Sabarimalai Manikandan, M.: A Levelset Based Binarization and Segmentation for Scanned Malayalam Document Image Analysis. In: IEEE International Conference on computational Intelligence and Computing Research (2011)