International Journal of Control and Automation
Vol. 13, No.2, (2020), pp. 1215 - 1221

CNN Based Feature Extractionand Classificationfor Degraded Historical Documents

Devendran K[1], Keerthika P[2], Manjula Devi R[3], Santhosh sivan P[4], RebhashiA[5], and Ragul M[6]

[1]*Assistant Professor, Department of Computer Science and Engineering, Kongu Engineering College, Erode, India.*
[2]*Associate Professor, Department of Computer Science and Engineering, Kongu Engineering College, Erode,India.*
[3]*Associate Professor, Department of Computer Science and Engineering,Kongu Engineering College, Erode,India.*
[4,5,6]*Student, Department of Computer Science and Engineering, KonguEngineeringCollege, Erode, India.*
[1]skdeva@kongu.ac.in, [2]pkeerthika@kongu.ac.in, [3] manjuladevi@kongu.ac.in, [4]santhoshsivan550@gmail.com, [5]rebha1999@gmail.com, [6]ragulfirst@gmail.com.

Abstract

Preservation of important historical records/papers is a difficult task because of large volume and periodic degradation of texts. So character identification of historical records/papers is inevitable. Image Binarization is the pre-processing step for character recollection in historical records/papers. The Binarization process converts a greyscale image into a binary image. Thus image binarization is very important for character recognition. But it is a challenging task due to complex background and noises in the images. As binarization and text line segmentation also play a vital role in character recognition. In this paper we present an additional step, they are Feature extraction and Classification. The text line segmentation and binarization are implemented using Global Threshold values derived from Otsu's algorithm and Local threshold values form Niblack and Savaula's algorithms. The Feature Extraction and Classification is implemented by Convolution Neural Network. The result obtained from the methodology is less sensitive to noise and has high contrast.

Keywords:*Convolution neural network, Character recognition, Feature extraction, Binarization.*

1. Introduction

Mostly character detection in historical records/papers starts with binarization. Implementation of binarization involves the separation of the forepart amid the backdrop. The difficult task in character identification of the historical record/papers is the disjunction of colored and faded text from the backdrop noise and bleed over. The consequence faced is the contrast of the colored text and faded text is often similar to the contrast of the backdrop and the bleed through. Many methodologies are developed for better binarization. The subsist methodologies based on handwritten character identification require the features to be extracted and the display level is low. So a Convolution Neural Network (CNN) based handwritten character identification is applied. It is a deep learning technique with deep neural networks classes and also a fast system toward character recognition. The Convolution Neural network is used for classification and feature extraction which is considered as a major two stages. Classification also features extraction that comes next to the binarization classification.

ISSN: 2005-4297 IJCA

Binarization is executed by applying threshold values. They are two types of threshold values, they are Global and regional. The Global method uses a single threshold for all images. Otsu's algorithm is the most successful global thresholding method. The regional method uses the threshold for each pixel. Niblack algorithm and savaula's algorithm are the two most successful algorithm on using the local thresholding method. After Binarization is performed for an image then Text Line Segmentation is implemented. Text line segmentation deals with images of binary form. So the very first step is implemented as binarization, which converts the grayscale image into a binary image. Text Line Segmentation is an important step to obtain good recognition rates. As we have said earlier that classification and feature extraction is implemented by the Convolution Neural network, which is implemented by the yield from the binarization, and segmentation is fed to the Convolution Neural Network model.

2. Proposed Work

In this study, we proposed four levels such as Binarization, Feature extraction, Text line segmentation, and Classification for character recognition of historical records/papers. In Binarization the algorithm such as Otsu's, Niblack, and Savaula are combined to binarize images of historic record/paper. It is also a form of segmenting an image into constituent objects. In segmentation, the Bounding Box method is used to separate the text. Feature Extraction and Classification is implemented using Convolution Neural Network, a deep learning technique.

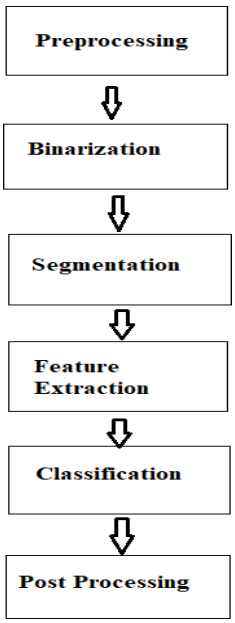

Figure 1. Proposed Architecture

3. Methodology

3.1. Binarization

In Binarization, a greyscale image of a handwritten document is converted into a binary image. To implement Binarization Otsu's, Niblack and Savaula's algorithm is used.

3.1.1. Otsu's Algorithm

In Otsu's Algorithm, a global threshold value is used. That is a single threshold value is used for a whole image. Otsu's algorithm iterates through the values of all the possible thresholds. And it calculates a measure of spread for each side of the threshold of the pixel levels. It detects the pixel whether it belongs to either foreground or background.

3.1.2. NiblackAlgorith,

Niblack binarization algorithm uses a local threshold value for binarizing the image. In this algorithm, the threshold value is calculated for each pixel in the image. The local grey level mean and the standard deviation is used to calculate the thresholding value for each window.

$$T= m + k\sigma \quad (1)$$

Where, m is local grey level mean, σ is standard deviation and k = 0.2.Here the window size is predetermined by the user.

3.1.3.Savaula's Algorithm

Savaula's binarization algorithm also uses a local threshold value for binarizing the image. The local grey level mean and the standard deviation is used to calculate the thresholding value for each window. Savaula's algorithm gives better results compared to the Niblack algorithm. It performs well when the background has a light texture, uneven illumination, and big variations. The threshold at each pixel is calculated using equation 2

$$T = m(1 - k\ (1 - \sigma/R) \quad (2)$$

Where, R is a dynamic range of standard deviation, k is a user defined parameter and m is a local grey level mean. The author suggested k=0.2 and R=125.

3.2. Text Line Segmentation

The requirement for text line segmentation is a binarized image. So binarization is a necessary step to implement before Segmentation. Text Line segmentation is performed using Bounding Box method.

3.2.1. Bounding Box Method

In Bounding box method the portions with the similar characteristics gets segmented by overlaying a rectangular box. To segment the text the bounding Box is created around the text by sliding window technique. It is a computationally expensive task. In Bounding box technique to detect the textaslidingwindowispassedthroughtheimageinthatwindow.Todetect text portion with different sizes, different window size is triedout.

3.3. Feature Extraction

In Feature extraction process an initial set of raw data is reduced to more manageable groups of processing which is called dimensionality reduction. Feature Extraction is implemented using convolution Neural Network.

3.3.1. Working of Convolution Neural Network (CNN)

Convolution Neural Network consists of several layers, where the individual layer consists of neurons that hold some data. Every neuron in the input layer consists of a Feature of the handwritten input data sequence in an image. By applying diverse filters we can extract important features using convolution layers. The pooling layer can be applied to decrease the dimensionality reduction of an input image i.e Spatial size. In Fully connected layer every node is interconnected and consists of feature, that is used for Classification. The output layer is the last layer in the Convolution Neural Network layer. It consists of several nodes depend on the number of classes. When an input image is given CNN applies different filters to create a feature map and applies a Rectified Linear Unit function to increase non-linearity in the image. Then the pooling layer is applied to each feature map and then it evens the pooled image into one log vector. And then it inputs the vector into a fully connected neural network. CNN trains through forwarding propagation and backpropagation for many epochs. This process is repeated until we have a well-defined neural network with trained weights and feature detectors.

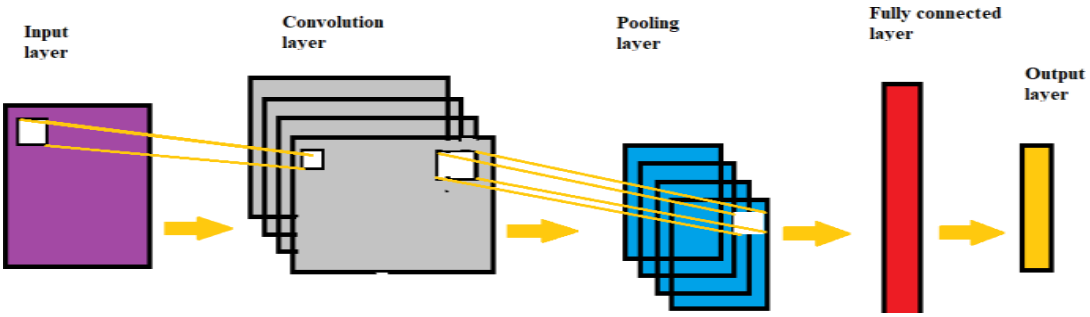

Figure 2. Layers in Convolution neural Network

3.3.2.Feature Extraction and Sliding Window

To implement Feature Extraction using Sliding Window technique we want a one- to-one mapping between frames used to extract handcrafted features and frames provided to the neural network. We can implement this by scanning two sliding windows of different sizes with the same shift on the images, that is the original size one and the normalized one. We use a width of 9px for the hand written features and a size of 39px for the pixels. The hand written features comprises of 26 statistical and geometrical features.

4. Classification

After Feature extraction text classification has to be implemented. Basedonatrainingsetofdatacontaininganewobservations,Classification process identifies a set of categories on which a new observation belongs. The correct class for each record is known and the output nodes can be assigned correct values 1 for the node corresponding to the correct class, and 0 for the others. It is possible to compare the network's calculated values for the output nodes to these correct values, and calculate an error term for each node. These error terms are then used to adjust the weights in the hidden layers, so during the next iteration the output values will be closer to the correctvalues.

5. Result and Discussion

The data set used in this paper is HDLAC 2011. After performing the Feature Engineering, Selection, and implementing a model and getting some output in forms of a

probability or a class, the next step is to find out how effective is the proposed model based on some evaluation metrics using test datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms. Classification performance metrics used are Accuracy, Precision, recall which can be used for sorting algorithmsprimarilyusedbysearchengines.

Table 1. Comparison of different algorithms on the HDLAC dataset

METHOD	PRECISION	RECALL	F MEASURE
W.Niblack	0.75	0.89	78.2
Savaula	0.89	0.90	91
FABEMID	0.88	0.87	89.66
ANN	0.83	0.81	80.22
CNN	0.89	0.92	91.92
Proposed	0.91	0.89	93.28

The Table 1 illustrates the performance of various methods on the basis of predictive measures on concern with precision, recall, FMeasure and the proposed model resulted with an accuracy 92%.

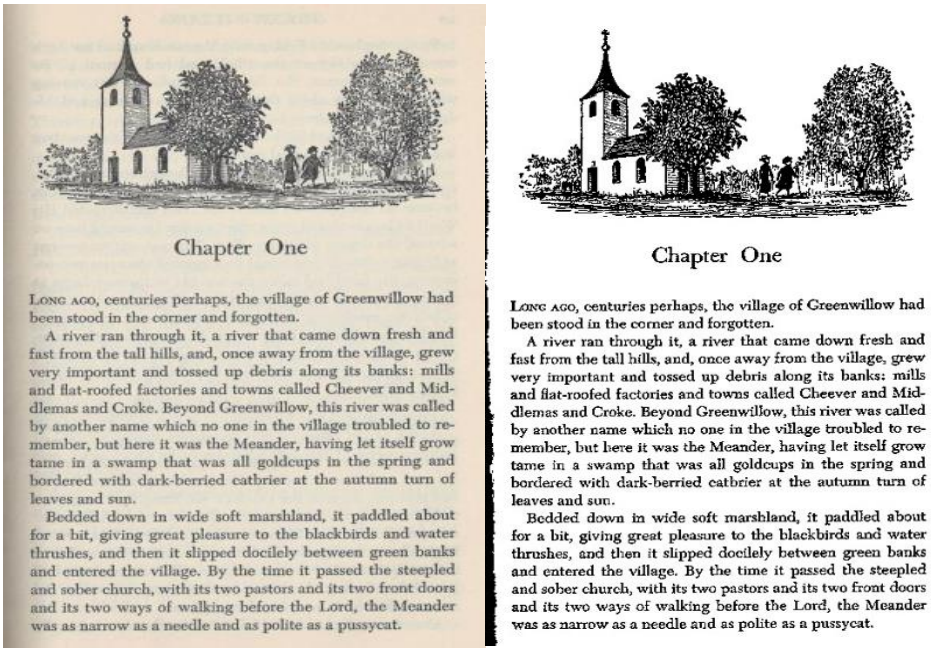

Figure 3. Original image Vs Binarization

The Figure 3 depicts the resulted output after the binarization process and figure 4 depicts the segmented image and its respective feature extraction.

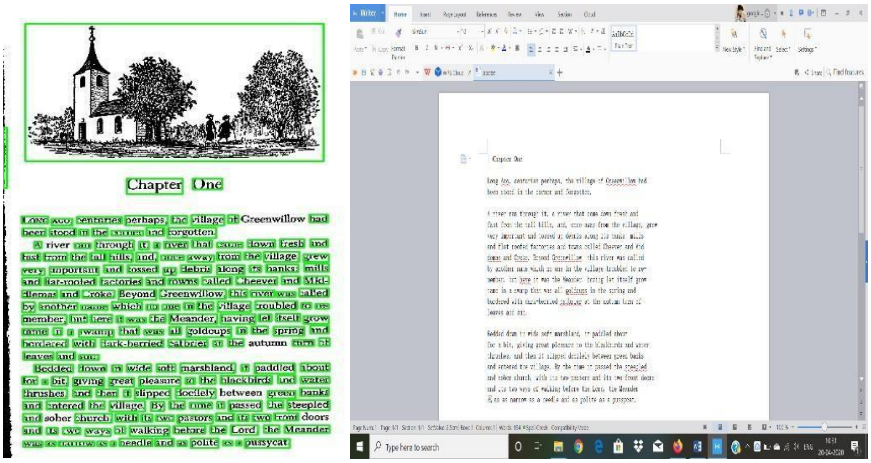

Figure 4. Segmented Image and Feature Extracted Image

6. Conclusion

The proposed method is robust to solve faded characters, smear, strain, non-uniform illumination, low contrast and large signal dependent noise. Experiments on HDLAC 2011 data sets show that the proposed approach is appropriate for binarization of the document image. Another contribution in the work is to use Feature extraction and Classification to extract and identify the appropriate characters in the Documentimage.

Acknowledgement

This work was carried out as part of “Main project” during eighth semester BE Computer Science and Engineering in partial fulfillment ofthe requirement for the award of Bachelor of Engineering Degree in Computer Science and Engineering under the AnnaUniversity

References

[1] J. Pastor-Pellicer, S. Espana-Boquera, F. Zamora-Mart ´ınez, M. Z. Afzal, andM. J.Castro-Bleda, “Insights on the use of convolutional neural networks fordocumentimagebinarization,” inInternational Work- Conference on ArtificialNeural Networks. Springer, 2015, pp.115– 126.
[2] A.S. Kavitha a, P. Shivakumarab,*, G.H. Kumar a, Tong Lu “Text segmentation in degraded historical document images”, b Faculty of Computer Science and Information Technology, University Of Malaya,B-2-18, Malaysia
[3] Y. S. Huang and C. Y. Suen, ―A method of combining multiple experts for the recognition of unconstrained handwritten numeralsIEEE Trans. Pattern Anal. Machine Intell., vol. 17, pp. 90–94.
[4] Oivind Due Trier, Anil K Jain, TorfinnTaxt. ―Feature Extraction Methods for Character Recognition-A survey‖,1995
[5] M.H. Mohamed Dylaa, F. Morain-Nicolierb“ Text line segmentation andbinarization of handwritten historical documents using the fast and adaptive bidimensional empirical mode decomposition”
[6] BJ.F.K.S.M.A. Bhuiyan, R.R. Adhami, A novel approach of fast and adaptive bidimensional empirical mode decomposition, IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, 2008, pp. 1313–1316.
[7] J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognit. 33 (2000)236–255.
[8] Salma ShofiaRosyda and Tito WaluyoPurboyo, “A Review of Various Handwriting Recognition Methods” International Journal of AppliedEngineeringResearch,Vol.13,2018, pp 1155-1164.
[9] Darmatasia and Mohamad Ivan Fanany, “Handwriting Recognition on Form Document Using Convolutional Neural Network and Support Vector Machines”, CoICT.2017.

[10] Asha K, Krishnappa H K "Handwriting Recognition using Deep Learning based Convolutional Neural Network", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November2019.
[11] Mrs. Safna K M "English Handwritten Character Recognition using Convolutional Neural Network (CNN)", IJSRD - International Journal for Scientific Research &Development| Vol. 6, Issue 02, 2018 | ISSN (online):2321-0613.
[12] J.Pradeep, E.Srinivasan, S.Himavathi, "Performance Analysis of Hybrid Feature Extraction Technique for Recognizing English Handwritten Characters", IEEE, 2012.
[13] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional Neural Network Committees for HandwrittenCharacterClassification",pages 1135– 1139.IEEE,Sept.2011.
[14] Chris Tensmeyer and Tony Martinez "Document Image Binarization with Fully Convolutional Neural Networks", 2017 14th IAPR International Conference on Document Analysis and Recognition
[15] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique fordegraded document images,"IEEE transactions onimageprocessing,vol.22,no.4,pp.1408–1417,2013.
[16] Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", In Advances in neural information processing systems, pages 1097–1105, 2012.
[17] Z. Liu, H. Wang, S. Peng, "Texture classification through directional empirical mode decomposition," Proc. of the 17th International Conference on Pattern Recognition – ICRP20