Broken Character Recognition using Connected Components and Convolutional Neural Network

Roshan D Suvaris, S Sathyanarayana

Abstract*:Recognizing broken characters in scanned and ancient scanned text document is not easy because the characters may be broken and unclear. Many researches have been carried to recognize these broken characters. In this research paper we have described a new broken characters recognition method for English text documents only. The proposed method uses a hybrid approach which uses connected component concepts and convolutional neural network to identify the broken characters. The input to the approach is scanned or ancient text document which contains unclear text that is difficult to recognize and hence our new proposed methodology will recognize these characters with greater accuracy and it will give the recognized characters to the user. The projected technique has attained a precision up to 92% in recognition.

Keywords : Connected Components, Convolutional Neural Network, Image Processing.

I. INTRODUCTION

Optical characters recognition is and have been electronically converting, handwritten, printed or other type of text from images, scanned documents etc.

Optical character recognition is a method which is used to recognize the characters in text documents or images. Recognizing the characters is a tedious undertaking task, even though there are lot of algorithms have already been developed, still more work is required to perceive the characters in various situations and scenarios. Few algorithms work better compared to other algorithms in certain conditions. It can recognize the characters in the archive if the document and the text is clear. Majority antiquated reports or documents, low resolution and other type paper documents are all digitized in the present world. However these digitized text documents are not crystal clear and may contain broken character or other type of characters due to long storage or due improper use of scanner and improper photocopies, which makes data unreadable. The optical character recognition algorithms are used to recognize these digitized copy text.

It is difficult to recognize the characters due to various textual styles, noises found in the background and due to background complexity and so on.

Distinctive character recognition algorithms has been already developed which are used to recognize characters from printed text documents, postal documents, recognize the character in stamps, number plates, and handwritten characters of different languages such as English, kannada, hindi, tamil, Arabic bangla and so forth and furthermore they have utilized various approaches to distinguish the characters for example Transfer learning, CSM neural network, deep neural network, local binary pattern networks, dynamic baysian networks, CNN classifier and efficient word interference, k-means based feature learning etc which will perform good in various conditions.

In this algorithm we are using a hybrid approach in which connected components is used to create sub graph for characters and convolutional neural network is used to recognize the subgraph characters and we have used chars74k dataset to train the network.

II. LITERATURE REVIEW

To recognize the characters of different documents such as text document, scanned document, handwritten documents, bills, post, scene images, number plates in different countries thousands of algorithms are already developed by different researchers which will work different type documents in different conditions. In this section we describe some of the methods which is used for character recognition.

Charles et al [1] proposed a method in which character is recognized using neural network at a given location, then word recognizer will use the character recognizer to find the most relative word inside the specified rectangle on the input page. It uses the camera based input image and its accuracy is up to 95% for the word recognition

Michael droettboom [2] has proposed a method in which words are represented using graph and it uses connected component methodology to represent the character subgraphs. And these subgraphs are fed to k nearest neighbor method for recognition.

Laurence likforman et al [3] uses a method which recognizes the character in ancient printed books by coupling two hidden markov models (Vertical and Horizontal HMMs). The two coupled dynamic baysian networks are modelled to show the interaction between these two.

Rakesh Kumar Mandal and NR Manna [4] has given a new algorithm to recognize handwritten English character. In this method the character matrix is compressed to reduce non-significant elements then these elements are given to CSIM neural network to recognize.

Dileep Kumar Patel et al [5] in their paper they have proposed a method to recognize the handwritten characters using discrete wavelet transform and Euclidean distance metric. This method has to be more accurate and faster.

Revised Manuscript Received on January 30, 2020.

***** Correspondence Author

Roshan D Suvaris*, Research Scholar bharathiar university Coimbatore, works as Asst. Professor at AIMIT St Aloysius College Mangalore. Email: roshansuaris@gmail.com

Dr. S Sathyanarayana, First Grade Womens College Mysore, Karnataka India. Email: ssn_mys@yahoo.com

DOI: 10.35940/ijrte.E1011.0285S20

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Dickson Neoh Tze How et al [6] used a deep convolutional neural networks to recognize the Malaysian vehicle license number plate which is trained using computer font's a-z and 0-9. The neural network has been trained using back propagation with gradient descents. Its performance is better with reasonable accuracy. P Rajendra et al [7] also used supervised learning using CNN. It also uses smoothing filters and contour detection.

XinHao Liu et al [8] has used convolutional neural network to recognize the character in the scene text and then weight finite state transducer (WFST) based word recognition is used to recognize the text in natural scene images.

Savitha Chaudhary et al [9] has proposed a method which is used to recognize text from scene images. This method first uses MSER to find the text regions then convolutional neural network is used to predict the character.

Yejun tang et al [10] proposed a transfer learning method using CNN in this the first model is trained and the weights of this model is used to initiate another model which is used as a feature extractor and used as a classifier. The second model is fine tuned to recognize the characters in the document.

III. PROPOSED METHODOLOGY

The goal of this research paper is to recognize the broken characters from old scanned documents. The algorithm begins with segmenting the document or image horizontal histogram method to get the lines. Then vertical histogram segmentation is applied to extract the words from lines [11]. Once the words are extracted, the characters in the words may be broken so an undirected graph is created in which each node/vertex represents the character component. The vertices are connected to one another if they are present inside certain threshold of distance and an edge is drawn between the two vertices. This will roughly create graph for each word in the document or image.

The connected component is a subgraph in which any two vertices are connected by path and no vertices are connected with any other vertices in the super graph.

In this, character segmentationis not used because character segmentation may give two character instead of one character, for example character "m" may be recognized as "r" and "n". for example

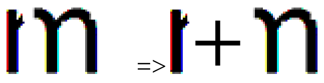


Fig 1: showing Segmentation fault

Next all possibilities are checked to create the subgraph for each character i.e. how components can be joined Depth firstsearch is performed from each vertex and subgraphs are created. Each vertex in the graph is numbered and from low to high vertices are visited to avoid already visited nodes [2].

After this subgraph are taken and it is converted to images which represent the individual broken character in the word and these images character are given as input to the already trained Convolutional neural network.

The CNN consists of four layers. It has been also known as CNN or convNet that is so far analyzing images, although image analysis has been most wide spread use of CNNs they can also be used for other data analysis and classification problems.Convolutional Neural network has four layers namely convolution, RectifiedLinear Unit layer or ReLu, Pooling layer and fully connected layer. Each layer perform different tasks. The convolutional layer defines a collection of filters or activation maps, each with the same dimension as the input. The filter operates on a small region of the input. The filter is moved on the every possible position on the image. The convolution layers has got four steps – line up the features and the image map and then multiply feature pixel by image pixel and the values whatever we get then divide by total number of pixels in the feature. Filter is moved over the image and the above four steps are repeated. All the filters are applied to the given image and the result is taken.

The second step in Convolutional Neural Network is ReLu layer which is a activation function. It will only activate a node if the input is above a certain quantity and it is calculated using the following function

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$$

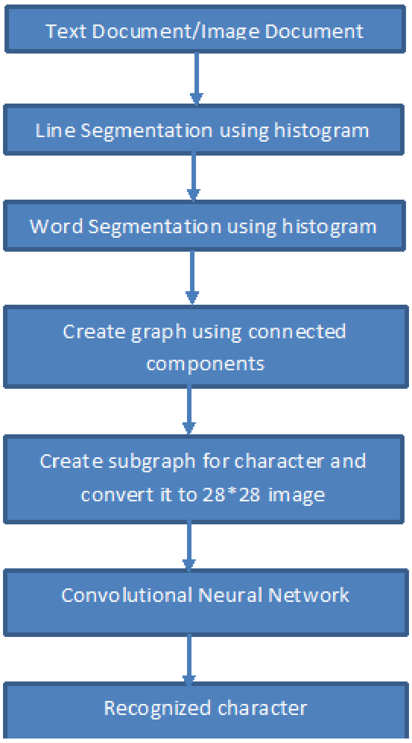

Text Document/Image Document
Line Segmentation using histogram
Word Segmentation using histogram
Create graph using connected components
Create subgraph for character and convert it to 28*28 image
Convolutional Neural Network
Recognized character

Fig 2: flow chart of proposed methodology.

After the ReLu layer pooling layer is used in which the image is shrinked. In this we take a window size of 2 or 3 and from that we take an only the maximum value from the input image and place it in the resulting image.

Lastly fully connected layer is used in which actual classification is done. We create a single list from the output got from the pooling layer which is used to recognize or classify the input images.

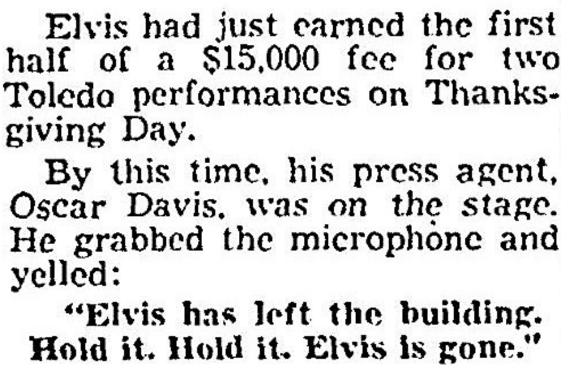
The Convolutional neural networks are already trained and tested using chars74k dataset.We use convolutional neural networks due to higher rate of recognition when compared to others.

IV. IMPLEMENTATION

The method is implemented using python. The python supports lot of libraries which is used in image processing like tensorflow, keras, pandas etc.

In this method we have used chars74k dataset

The text document is first scanned or any digital image is taken and it is given to the algorithm as an input. Sample broken character text image is shown in fig 3.



Elvis had just earned the first half of a $15,000 fee for two Toledo performances on Thanksgiving Day.

By this time, his press agent, Oscar Davis, was on the stage. He grabbed the microphone and yelled:

"Elvis has left the building. Hold it. Hold it. Elvis is gone."

Fig 3: input image

The image may contain some background noise so first a Gaussian filter is applied to the image. The Gaussian filter is a low pass filter which is used to remove the noise and to smooth the image. In the next phase histogram methodology is used to read the lines and words in that lines. The lines from the text image is taken using horizontal histogram methodology. The horizontal histogram is generated for the entire document which separates different lines from one another and then each lines are taken and words from those lines are extracted from those lines using vertical histogram.

In the next phase of the algorithm each word is taken and a graph is generated for each word. The vertex represents a connected component in the word. This creates many graphs which are equivalent to different strings or words in the document. Next every possible way in which components of the character can be connected are evaluated and subgraphs are generated for each character. The maximum number of connected components for a character may vary on amount of degradation in the scanned image. Once the characters are generated it is converted to image. The sample output for the first line for the above figure is shown in fig 4.

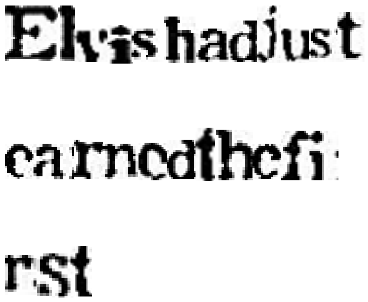
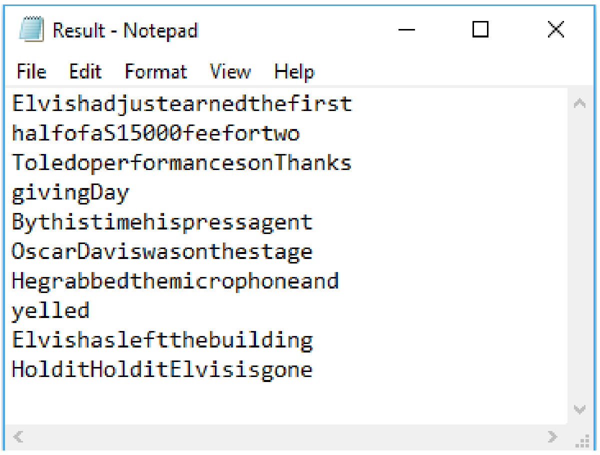


Fig 4: Characters images generated using CC

Now each character is read and it is passed to Convolutional Neural Network. The CNN is trained using chars74k data set. The chars74k dataset is actually contains English characters with different fonts and Kannada characters. The individual characters are read and it is recognized by this network. The output of this phase is shown in fig 5.



```
Elvishadjustearnedthefirst
halfofaS15000feefortwo
ToledoperformancesonThanks
givingDay
Bythistimehispressagent
OscarDaviswasonthestage
Hegrabbedthemicrophoneand
yelled
Elvishasleftthebuilding
HolditHolditElvisisgone
```

Fig 5: Result after applying CNN to individual character image

V. CONCLUSION

We have utilized connected components in this method because if we use the normal character segmentation methods then we may get two characters for one character. That is the reason we have used connected component method with CNN. The accuracy of convolutional neural network for printed characters and handwritten character for different languages with normal characters is high, so in this method we have utilized CNN. The Convolutional neural network can recognize any type and kind of characters. By using this method we have checked different documents and got the accuracy up to 92% for broken character recognition. We can further additionally improve this methodology by training the neural network for special characters for example comma, semi colon, colon etc. from the scanned documents.

REFERENCES

1. Charles Jacobs, James Rinker, Paul viola and Patrice Y. Simard "Text recognition of Low-resolution Document Images" International journal on Document analysis and Recognition ICDAR 2005.
2. Michael Droettboom "Correcting broken characters in the recognition of historical printed documents" In Proc. of Joint Conf. on Digital Libraries, pp.364-366, 2003.
3. Laurence Likforman-Sulem, Marc Sigelle "Recognition of broken characters from historical printed books using Dynamic Bayesian Networks" International Conference on Document analysis and Recognition ICDAR 2007.
4. Rakesh Kumar Mandal, N R Manna "Handwritten English character recognition using column-wise segmentation of image matrix" wseas transaction on computers Issue 5, Volume 11, May 2012.
5. Dileep Kumar Patel,Manoj Kumar Singh, Sushil Kumar Yadav Tanmoy Som, "Handwritten character recognition using multiresolution technique and Euclidean distance metric" Journal of Signal and information processing, 2012.
6. Dickson neoh Tze How, Khairul salleh Mohammed Sahari "character recognition of Malaysian vehicle license plate with deep convolutional neural networks" IEEE International Symposium on Robotics and Intelligent Sensors (IRIS2016) December 2016.
7. P Rajendra, Rahul Boadh and K sudheer Kumar "Design of a Recognition System automatic vehicle License plate through a convolutional neural network" International Journal of Computer Applications, Volume 177 No.3, 2017.

8. XinHao Liu, Kunio kashino, Xiaomeng Wu and Takahito Kawanishi "Scene Text Recognition with CNN classifier and WFST-based word labelling" International Conference on Pattern Recognition (ICPR) 2016.
9. Savitha Choudhary, Sanjay Chichadwani and Nikhil Kumar Singh "Text Detection and Recognition from Scene images using MSER and CNN" International conference on advances in electronics, Computer and Communications (ICAECC – 2018).
10. Yejun tang, Akio furuhata, Yanwei Wang, Qian Xu, and liangrui Peng "CNN based transfer learning for historical Chinese character recognition" IAPR workshop on Document analysis systems IEEE 2016.
11. Nallapareddy Priyanka,Ranju Mandal and Srikanta Pal, "Line and Word segmentaion approach for printed Documents" IJCA special issue on Recent trends in Image Processing and Pattern Recognition RTIPPR 2010.

AUTHORS PROFILE

Roshan D Suvaris, Pursued his master degree Master of Computer Application from NMAMIT nitte under VTU University. Currently he is a research scholar in Bharathiar University Coimbatore. Currently he is working in AIMIT St Aloysius College Mangalore. His interest of research is image processing, machine learning.

Dr S Sathyanarayana, Completed his BTech, Msc and Phd. His area of interest is Decision Support system and Cloud Computing, Relational Database Management Ssystems.