Keyword Spotting in Images Using Convolutional Neural Network

J Jagadeesan, Mayank Khandelwal, Rahul Kevadia, Prayank Shah, Sooryabhan Singh

Abstract*: Machine vision is the new and faster way to train a machine to better understand the videos and images. Machine vision can also be used to detect the text in an image using various machine learning approaches. Machine vision and deep learning have been booming with a greater speed in past few years thus thus there is always a better solution or a problem every now and then. This is the main difference between the existing system and the proposed system as there has been many advancement in the deep learning field it gives a better and faster approach to deal with the real world problems. With this we can gather text and make the machine better understand it by the use of image manipulation and machine learning's neural networks. Image manipulation includes thresholding, normalization, alignment and the neural network is used to extract the features and match it with the ones on which a machine is trained.

Index Terms*: Machine Vision, Machine Learning, Neural Network, Deep Learning.

I. INTRODUCTION

Extracting, Inspecting and Analysis of images using computational power to gather information about a certain object. With the use of Neural Networks and concept of Machine Learning we can train the computer to get the text from the given image.

This paper is a Machine learning algorithm to extract a keyword from a text present in an image. With the help of method such as thresholding the image is converted to binary image to better detect the edges and reduce the noise and then the Deep Convolution Neural Network is used to train the machine to gather the keywords. Neural networks are the algorithms, modeled to simulate the human brain, that are designed to recognize patterns. They are used for interpreting data through a kind of machine perspective, labeling and clustering data.

A neural network can recognize any number, images, sound, text or time series. Deep Learning is one of the most reliable machine learning method. It trains a system to predict outputs from the given set of inputs. There are 3 main layers in a deep learning model , i.e., Input Layer, Hidden Layer and Output Layer. Input layer is where the raw input is provided to annotate and train the system. Output layer is where the accuracy, validation accuracy and the actual outputs are compared. Hidden layer is the core of the Deep learning neural networks and a deep learning model contains more than one hidden layer. In this layer the processing of the data takes place and generally consists of the Convolution, Max Pooling, Dense and Flatten layers.

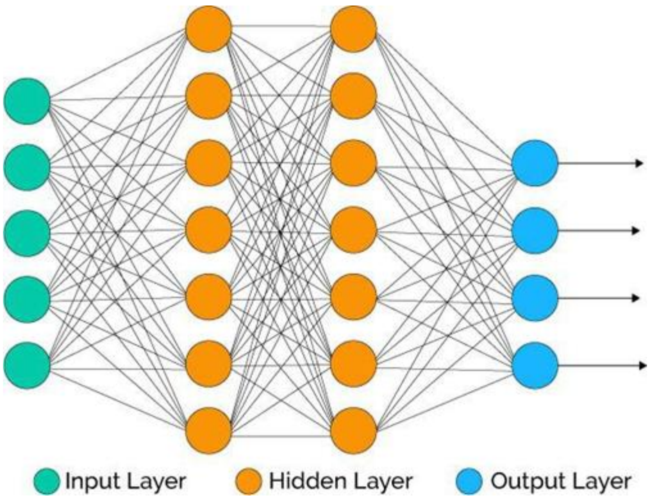

Fig. 1. Basic Neural Network

II. LITERATURE SURVEY

This section is to discuss about the different Methodologies Review/ Literature Review and Motivation Outcomes from the given refrences.

1) **"Efficient Learning-Free Keyword Spotting", George, Georgis, Nikolas, Basilis (Issue 2018) [1]**

This paper is proposed to spot the keywords using the OCR technique and by using the binarization technique to enhance the text quality for easy spotting. The method discussed here are old and may be slow and the accuracy is not better.

Revised Manuscript Received on April 07, 2019.

Dr. J Jagadeean Ph.D., Professor and Head, Computer Science and Engineering, SRM IST, Chennai, India.

Mayank Khandelwal B. Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Prayank Shah B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Rahul Kevadia B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Sooryabhan Singh B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

2) **”Text Detection Based On Convolutional Neural Networks With Spartial Pyramid Pooling”, Rui Zhu, Xiao-jiao, Qi-hai, Ning Li, Yu-bin (Issue 2016) [2]**

This Paper Proposed The Method To Learn Text Using Tradition Convolutional Neural Network. Method used here is slower than the current faster and more efficient Neural networks and produces less accuracy is detecting the text.

3) **”Binarization of Old Halftone Text Documents”, Chandranath, Pratnik, Bidyut (Issue 2015) [3]**

This Paper Proposed The Method To Binarize Or Threshold The Image. Method proposed here is to binarize an old document image to enhance the clarity if the text written on them which help in better detection of the text.

4) **” Feature Reduction For Neural Network Based Text Categorization”, Savio, Y. Lam, Dik Lun Lee (Issue 2016) [4]**

This paper proposed the method to reduce and gather the features. In this method the features extraction is done by reducing the noise and then categorizing them into different features category.

5) **”Combination Of Document Image Binarization Techniques”, Bolan Su, Shijiam Lu, Chew Lim Tan (Issue 2011) [5]**

This Paper Proposed The Various Methods To Binarize An Image. In this proposed paper they discussed about the various binarization techniques to improve the text quality of the document image.

III. SYSTEM OVERVIEW

This system consists of various layers of processing. Three main layers are Data annotation, Neural Network and Output Section. In the Data Annotation part the raw input is given in the format of images and then are labeled accordingly to create a process-able data set. And then this data-set is divided into train, validation and test sets on which the deep learning algorithm is applied to gather features and learn.

Neural network section is the main deep learning algorithmic part. In this part first the proper libraries are taken into the consideration and then a neural network is constructed (like VGG, ResNet, ResNxt, AlexNet etc.) and with the help of these neural networks the system is trained on the data-sets given. First it is trained on the train set and then using different classifier it is tested on the validation set so as to get the most suitable classifier for the processing. Then the trained model is tested on the test set to check the accuracy of the model on the data-set.

Then in the output section it is checked with the image if the given input gives the desired output or not.

IV. PROPOSED SYSTEM

A. Contribution

Proposed method is to take care of the difficulties encountered during the identification of the text, such as variations in the images size, image quality, lighting and different hand-writing styles. These variations should be handled in such a way that resource required by the system are low and cost effective. There are mainly three steps involved in this proposed system, 1) Preprocessing, 2) Feature extraction and 3) Deep learning, which is the neural network.

These steps are related to each other and each one of these steps are planned to check different set of hand-written documents from the document images that have been classified as very crucial to the performance of the system. The variation in hand writing is very common when the collection of images of documents of different writing styles is taken into the consideration.

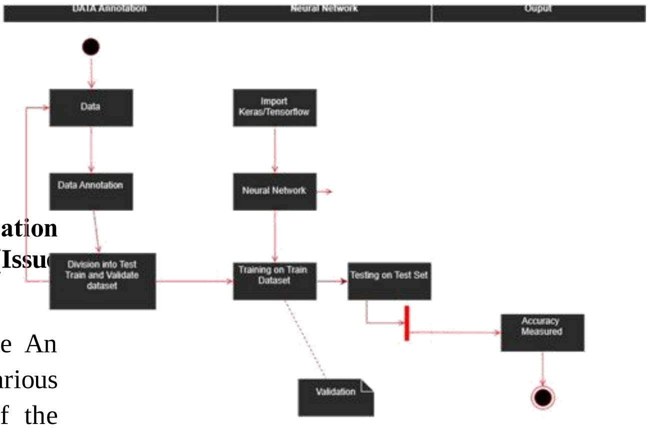

Fig. 2. System Architecture/Data Flow Diagram

One of the major problems faced by the descriptor is the rotation and translation of the image. To end this, the pre-processing method proposed here is to handle the rotation and vertical translation of the document image. The pre-processing method is dependent on the choice of the parameters and classifier whose performance is better on all the datasets present or provided.

In order to overcome these problems and avoiding any kind of error at an early step, different instances of the normalized image are generated for different preprocessing parameter. The proposed method will perform better than any of the older systems giving better accuracy, with cost-effective method.

B. Preprocessing

Thresholding is one of the simple way to perform image segmentation. Thresholding creates a binary form of image which is to make the gray scale image into either black or white.

Thresholding are of three types: 1) Simple, 2) Adaptive, 3) Otsu's Binarization

1) **Simple Thresholding**

If the pixel value is more than the threshold value, then that pixel is given a value which is either 0 or 1 which means either black or white. Figure 4 shows the binarization of the image in the Figure 3, using the OpenCV thresholding function.

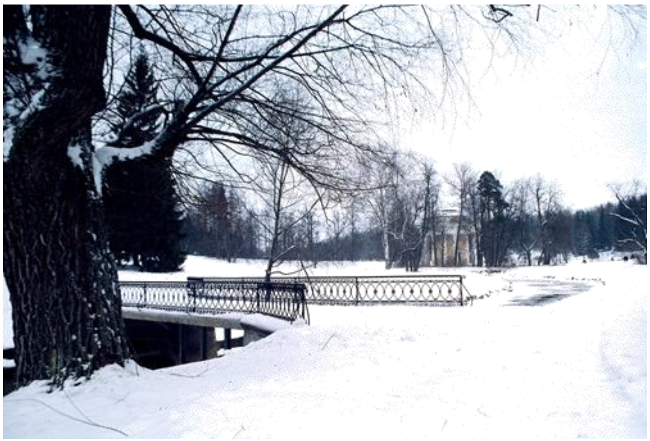

Fig. 3. Normal Image

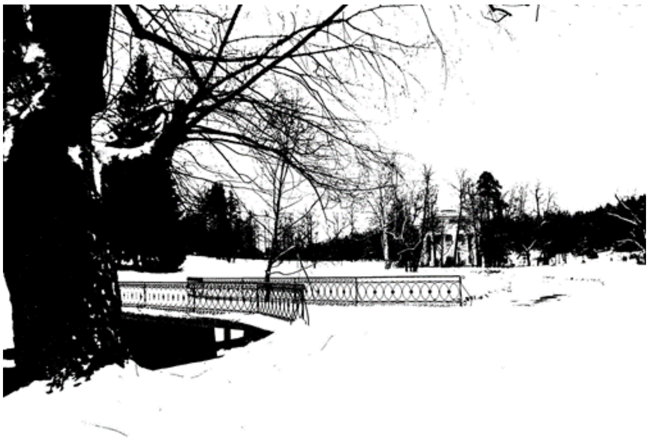

Fig. 4. Simple Thresholding

2) **Adaptive Thresholding**

This algorithm is used to calculate the threshold value of the image. THis is done to get different thresholds for different regions of the same image. It also works good with the images with different illumination as shown in Figure 5.

3) **Otsu's Binarization**

It automatically calculates a threshold value from image histogram for a bimodal images as it may not produce accurate results without binomial images. Output of the Otsu's binarization is as shown in Figure 6

For this, we use opencv's threshold function. Then the optimal threshold value is found and given as output as a second image as shown in Figure 6.

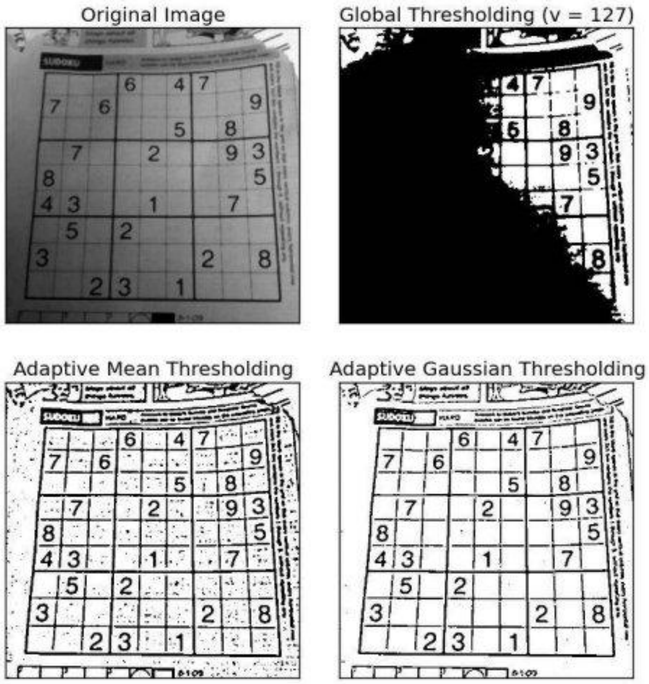
Original Image
Global Thresholding (v = 127)
Adaptive Mean Thresholding
Adaptive Gaussian Thresholding

Fig. 5. Adaptive Thresholding

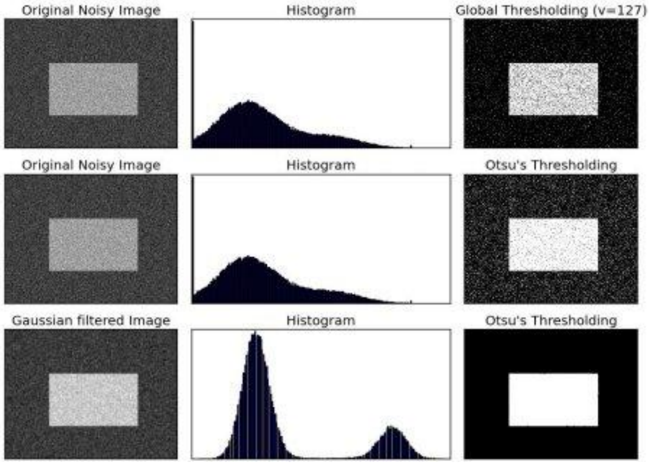

Fig. 6. OTSU's Binarization

C. Feature Extraction

Feature extraction in other words can be called as pattern matching. In deep learning we gather some features from text, images, audio or video and these features are first taught to the machine, that these are the features gathered from the given data and then these gathered features are used to match the existing pattern in the text, image, audio or video.In machine learning the feature extraction is done by the data scientist preparing the data but on the other hand in deep learning the feature extraction is done along with the classification of data which provides more and more understanding of the data by the system and in a quicker way.

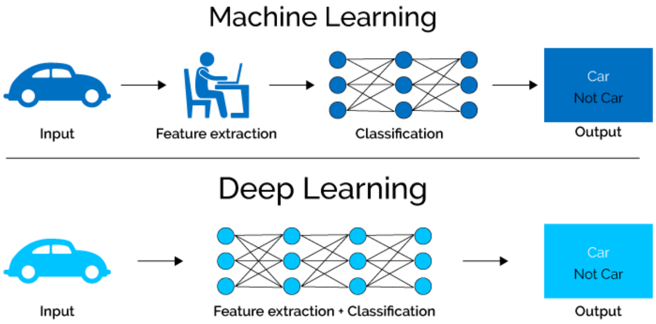

Fig. 7. Extraction in ML vs DL

D. Neural Network

Neural network is used to help a machine learn things how a human brain thinks and interpret things. With the use of neural network and deep learning technique, it became easy to interpret data in the form of images, audio and video, and with the use of Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, it became faster than the existing machine learning algorithms giving better accuracy. A Convolutional Neural Network have 4 basic layers: 1) Convolution, 2) Pooling, 3) Dense and 4) Fully Connected. A deep learning model with the CNN+RNN is way faster in processing. Dataset is also more important for the proper training of the model hence more the data better will be the performance of the model.

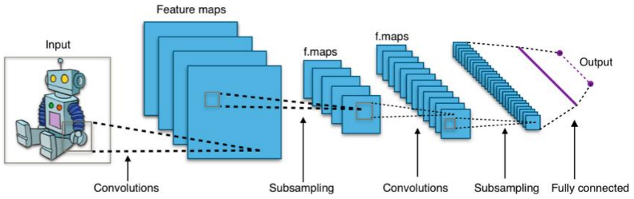

Fig. 8. Convolutional Neural Network

1) Convolution Layer

The convolution layer is the main building block of a convolutional neural network. The convolution layer comprises of a set of independent filters. Each filter is independently convolved with the image and we end up with 6 feature maps of shape 28*28*1.

2) Pooling Layer

A pooling layer is another important building block of a Convolutional neural network. The function of the pooling layer is to reduce the number of parameters and computation required in the neural network. Pooling layer independently operates on each feature map. Max pooling is one of the most common approach used.

3) Dense

A dense layer is a layer full of neurons in a neural network. All the neurons receive inputs from all the neurons present in the adjacent layer. The important components of the layer includes the weight matrix **W**, a bias vector **B**, and the activation function of the previous layer **A**.

4) Fully Connected Layer

The fully connected layer in the Convolutional Neural Network represent the feature vector for the input. This feature layer holds the information that is vital for the input. When the network gets trained, this feature vector is then further used for classification and regression. During training, the feature vector is used to find the loss. The convolution layers before the Fully Connected layer hold information regarding local features in the input image such as edges, blobs, shapes, etc. Each conv layer of keras hold several filters that represent one of the local features. The Fully Connected layer holds composite and aggregated information from all the conv layers of keras that matters the most.

V. EXPERIMENTAL OUTPUT

Machine vision and deep learning have been booming with a greater speed in past few years thus thus there is always a better solution or a problem every now and then. This is the main difference between the existing system and the proposed system as there has been many advancement in the deep learning field it gives a better and faster approach to deal with the real world problems. The exiting system is using the old OCR technique to scan the keywords and spot them, which takes a lot more time than the proposed system. Due to use of OCR with Convolutional Neural Networks the proposed system is way faster than the existing model and gives a better accuracy than the existing model. The proposed system gives a better, accurate and faster results than the existing systems due to use of Convolutional Neural Network (CNN)+Recurrent Neural Network (RNN) which makes the system faster than any existing models. The accuracy of the proposed system is 86% which is better than many of the existing system. Accuracy depends on the size of the dataset (as shown in Figure 10) and how much we train the system. The larger the amount of data, better the accuracy of the system. Dataset provided is a huge dataset of handwritten documents which provides a better understanding for the system.

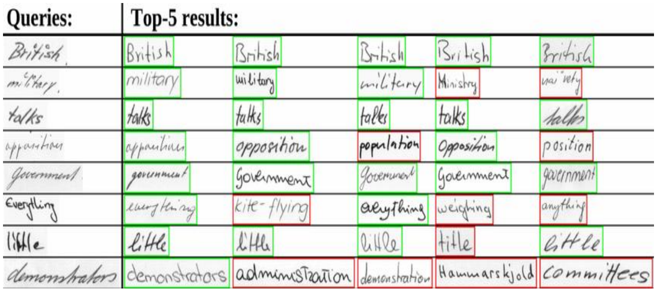

Fig. 9. Keywords Spotting

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

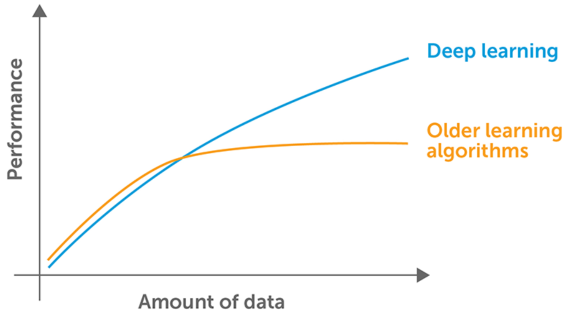

Fig. 10. Increase in accuracy depends on amount of data

VI. CONCLUSION

Thus hereby we conclude that the proposed system removes all the drawback of the existing system and enhance with better image preprocessing, and feature gathering. It provides the detailed solution to the existing system. Drawbacks such as usage of the older technologies and software are removed and use of newer and powerful hardware is done. Use of convolutional neural network gives it a better, faster way to gather information and give the output.

ACKNOWLEDGMENT

The authors would like to thank Dr. J Jagadeesan, Ph.D., Professor and Head, CSE Department, SRM Institute of Science and Technology, for his continuous support and also for his encouraging and fruitful advises which helped in accomplishing this task.

REFERENCES

1. ”Efficient Learning-Free Keyword Spotting”, George, Georgis, Nikolas, Basilis (Issue 2018)
2. ”Text Detection Based On Convolutional Neural Networks With Spartial Pyramid Pooling”, Rui Zhu, Xiao-jiao, Qi-hai, Ning Li, Yu-bin (Issue 2016)
3. ”Binarization Of Old Halftone Text Documents”, Chandranath, Pratnik, Bidyut, Michael (Issue 2015)
4. ”Feature Reduction For Neural Network Based Text Categorization” Savio, Y. Lam, Dik Lun Lee (Issue 2016)
5. ”Combination Of Document Image Binarization Techniques”, Bolan Su, Shijiam Lu, Chew Lim Tan (Issue 2011)
6. ”A Survey of Document Image Word Spotting Techniques”, A. P. Giotis, Sfikas, B. Gatos and C. Nikou (Issue 2017)
7. ”A Novel Word Spotting Method Based on Recurrent Neural Networks”,
8. ”Unsupervised Word Spot-ting in Historical Handwritten Document Images using Document oriented Local Features”, K. Zagoris, I. Pratikakis and B. Gatos (Issue 2017).
9. ”Deep feature embedding for accurate recognition and retrieval of handwritten text”, P. Krishnan, K. Futta and C.V. Jawahar (Issue 2016)
10. ”Zoning Aggregated Hyper-columns for Keyword Spotting”, G.Sfikas, G.Retsinas and B.Gatos (Issue 2016)
11. ”Robust scene text detection with convolution neural network induced MSER trees”, Weilin Huang, Yu Qiao, and Xiaoou Tang (Issue 2014)
12. ”Symmetry-based text line detection in natural scenes”,Zheng Zhang, Wei Shen, Cong Yao, and Xiang Bai (Issue 2015)
13. ”Deep features for text spotting”, Max Jaderberg, Andrea Vedaldi, and Andrew Zisser- man (Issue 2014)
14. .”Efficient scene text localization and recognition with local character refinement”, Lukas Neumann and Jiri Matas (Issue 2015)
15. ”A combined approach for the binarization of handwritten document images”, K. Ntirogiannis, B. Gatos, T. Pratikakis (Issue 2014)
16. ”A segmentation-free word spotting method for historical printed documents”, Thomas Konidaris (Issue 2015)
17. Wikipedia.com
18. Opencv.org

AUTHORS PROFILE

Dr. J Jagadeesan P.hD, Professor and Head, Computer Science and Engineering, SRM IST, Chennai, India.

Mayank Khandelwal B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Prayank Shah B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Rahul Kevadia B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.

Sooryabhan Singh B.Tech, Computer Science and Engineering, SRM IST, Chennai, India.