Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549463

END TO END SCENE TEXT UNDERSTANDING FOR COMPUTER VISION USING MACHINE LEARNING

Mr. T. Gnana Prakash[1], B. Tejaswini[2], Ch. Varshini[3], Ch. Harathi[4], G. Bhavani[5]
[1]Assistant Professor, Department of Computer Science Engineering, VNR Vignana Jyothi, Institute of Engineering and Technology
[2,3,4,5]B. Tech, Department of CSE, VNR Vignana Jyothi, Institute of Engineering and Technology.

Abstract:
The textual representation of landscapes and other natural settings gives us an insight into the location and can reveal some pertinent details. A challenging problem of considerable practical interest is reading characters in unconstrained scene images. This problem can be observed in medical fields. Scene text identification in a general condition is still a very open and difficult research subject, even though various text identification techniques have been established. Usually, simple images contain text that consists of characters of very similar size and font. In deep scene images, where style attributes will vary, this may not be the case. Thus, End to End Scene Text Understanding using computer vision can be applied to detect text from complex images with complex backgrounds and fonts. Using the text we extracted from the images, we can store it in a document and use it in the future.
Keywords: Image Processing, Text Recognition, Machine Learning, CNN.

1.0 INTRODUCTION

The use of multimedia technologies has grown significantly in recent years. Images play a significant role in multimedia technology and can contain a variety of information, including faces, people, scenes, text, etc. Text is shown to be the most prevalent type of material in photographs. one of the most crucial aspects to comprehending the contents of the image. Industries utilize text detection and recognition to read package labels, numbers, etc[1]. It is used to retrieve specific text contents from web pages as well as video captions. Both automatic number plate identification at toll booths and street board reading for unmanned vehicles are used for it. The assistance of the blind is a highly significant application of text detection and recognition[2].

Fig 1.1: Image with different fonts

The majority of information today is either on paper or in the form of images or videos. Images contain a lot of information. Text extraction with current technologies is only possible on plain backgrounds. Considering this, a technique to extract. Text drawn from broader contexts. Text extraction is helpful for a variety of purposes. Digital libraries, multimedia systems, information retrieval systems, and geographic information systems are a few of these uses.

Fig 1.2: Images with different backgrounds

The textual material in the photos has valuable data for indexing and automatic annotating. Text detection, localization, classification, and identification are all steps in the extraction of this valuable information. Our reasoning aids in locating and matching the using learned data, appropriately representing characters[3]. According to a study on text identification from real photographs, deep learning approaches are more precise and yield better results than image processing techniques. Several methods based on deep learning are used for text extraction processes in natural situations. Some examples include Feature Pyramid Networks (FPN),CNNs, and RNNs. Images of natural landscapes can be detected and their text extracted using image processing methods like The MSER (Maximally Stable Extremal Regions) detector, picture binarization, and morphological processes like erosion and dilation are used. After text recognition, the test is converted to speech by reading the image's alphabet utilizing some libraries and turning them into voices[4]. This conversion has a positive societal impact by assisting those who are blind in understanding the words contained in scene photographs

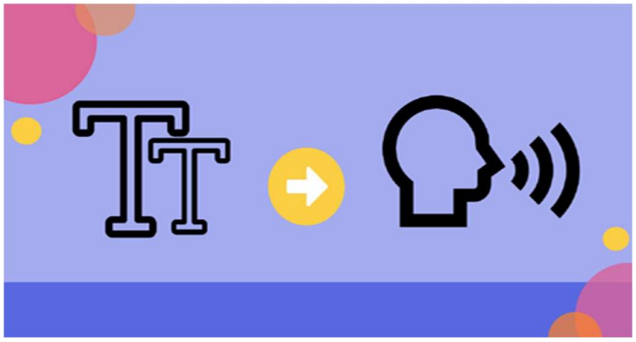

Fig 1.3: Text to speech conversion

After text recognition, the test is converted to speech by reading the image's alphabet utilizing some libraries and turning them into voices. This conversion has a positive societal impact by assisting those who are blind in understanding the words contained in scene photographs.

2.0 RELATED WORK

Ruijie Yan et al, [4] went through a process where they retrieved geometric structures as elements from local level to global level in the feature maps from a text image. Here they focused on specific features and encoded using MEA. They used CNN for feature extraction, MEA method for encoding and decoder for final text output. **Weilin Huang et al.,[5]** have proposed a model which has four main steps: component detection, components filtering, text-line constructing, and text-line filtering. The connected component method separates textual and non-textual information at the pixel level. Stroke width transform (SWT) is a connected component method used for component detection. **Pengwen Dai et al., [6]** adapted an attention network that is able to recognize text with different scale orientations. In this method, first the network they adapted will rotate the text according to their scales then dynamic scaling is done. **Asghar Ali Chandio et al., [7]** constructed a model that, without pre-segmenting the input into individual letters, modifies the order of the relevant properties from a whole word picture. The three main components of this model are the deep convolutional neural network (CNN) with shortcut connections used for feature extraction and encoding, the recurrent neural network (RNN) used for feature decoding of the convolutional features, and the connectionist temporal classification (CTC) used to map the predicted sequences into the target labels. **HamamMokayed et al.,[8]** chose a method which has two modules that feature extraction and defects component detection. In the feature extraction in order to separate character components from the detected text binarization approach is used.**MandanaFasounaki et al., [9]** by combining several image processing techniques obtained an approach where initially, MSER features are applied in order to identify ROI and then some geometric elimination and SWT elimination takes place and followed by the connection of characters where non-character regions are eliminated without OCR assistance. The results show that this model can effectively give promised results by combining various text detection methods. **Wenhao He et al., [10**] Convolutional feature extraction is the first step in the proposed method, which then combines multi-level feature fusion and multi-level learning in the network part to detect text from complex images. A separate module performs pre-processing to demand a word-level text. The most effective quadrilateral boundary regression method is direct regression. These tactics have achieved cutting-edge performance and have been proven effective in experiments. **Jeff Donahue et al.,[11]** gone through an algorithm, which utilizes neural network model and is fully differentiable and feed forward design where the generator contains the aligner which is used align the random input to aligned input, and the decoder which is used to convert the aligned input to the audio as output. **RandheerBagi et al., [12]** A portable scene text detector that can deal with scene photos' crowded surroundings has been proposed. It is a fully trainable deep neural network which leverages contextual cues from oriented area suggestions, global structural traits, and local component information to identify text occurrences.

3.0 METHODOLOGY

End-To-End scene text understanding in this project consists of various steps. They are collecting images, pre-processing, and sizing the images as required, building/training, and testing the model. Recognizing the region of interest, converting it to text, and then converting the recognized text to speech.

Our suggested approach can be used to accurately recognize text from intricately detailed photos with intricate backgrounds and typefaces. For this, the entire procedure is split into two steps:

- The process of localizing different areas of the scene and deleting non-text areas is known as text detection.
- Text recognition is the process of turning unreadable code from visual text.
- Our approach can reliably recognize multi-oriented text occurrences with precise bounding box localization by converting the challenging multi-oriented natural scene problem to a relatively simpler horizontal picture recognition problem.

The proposed methodology can be addressed in three main steps:

Localize text by bounding boxes:

Fig 3.1: Region of Interest

This step involves identifying the area of interest. An area of an image that you want to filter or manipulate in any other way is called a region of interest (ROI). By producing a binary mask, or binary image, with the same size as the picture you wish to create an ROI for, process with all other pixels set to 0, and the ROI-defining pixels set to 1.

In this stage, the full word is divided into more manageable, distinguishable parts. With localization data, it is possible to pinpoint certain elements in a picture. We train the convolutional neural network with images and datasets such as SVT and MSRA that contain fonts with diverse sizes, colors, and orientations to efficiently perform text detection and bounding-box regression at all locations and multiple scales in an image.

Crop the images and recognize text:

Thiswork,some threshold measures like K-means clustering are used to filter out extraneousbackground information and create character stanzas from picture pixels,followingthepredicted elements and geometrics.

Fig 3.2: Recognizing text from image

To handle text with irregular shapes more successfully, we also incorporate a spatial transformer into this network. The first two steps are an example of the text extraction stage.

The concept is easy to understand. We pass the entire image into the convolutional layer at once rather than sending individual regions into it one at a time, producing a feature map. The region proposals are then projected onto the feature map using the same external technique as before. As opposed to the raw image, we now have the areas in the feature map, which allows us to advance the regions in some fully linked layers and output the categorization and bounding box adjustments.

Converting the text into speech:

Fig 3.3: Converting text to speech

In the last phase, aligned text from the image is extracted and transformed into audio using a convolution network. The aligned content is then converted to an audio file using a decoder.

IMPLEMENTATION DETAILS

There are several processes involved in this project's interpretation of end-to-end scene texts. They are gathering photos, pre-processing and scaling them as necessary, constructing and training the model, then testing it. Identifying the area of interest, translating it into text, and then turning the translated text into speech.The dataset for this research is ICDAR 2015. The data folder is made up of photos, which include 461 image files, and the ground truth, which includes details about those 461 photographs. Images of various sizes and noise are included in the dataset. The photos have been cropped to (512,512) pixels in size. As a result, the boxes' ground truth has also been altered. The range of [-1, 1] has been applied to all of the photos.

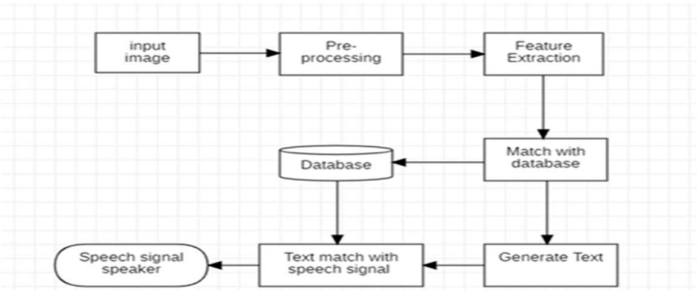

Fig 3.4: System Architecture

A matrix of dimensions is created using the ground truth coordinates as (grid height, grid width, 1, 5). The necessary dataset was produced, and then a model was developed to identify the text. The CNN model is made up of various layers that can determine how much weight is given to certain features. The image is converted by the convolution network into a single-dimensional vector with scores assigned to the class.

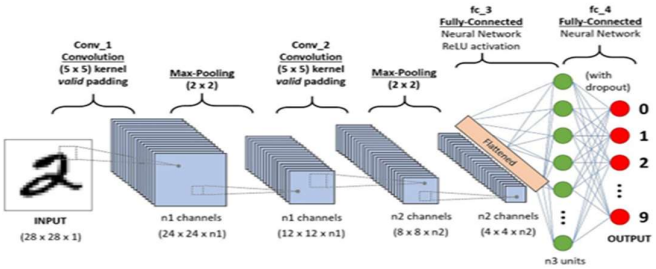

Fig 3.5: Layers in CNN

A CNN (Convolutional Neural Network) is a technique based on deep learning methods that can give input, assign different objects in the image with varying importance (learnable weights and biases), and be able to differentiate between them. The task of CNNs is to reduce the size of the pictures into a format that is easier to examine without compromising crucial components for producing an accurate forecast.

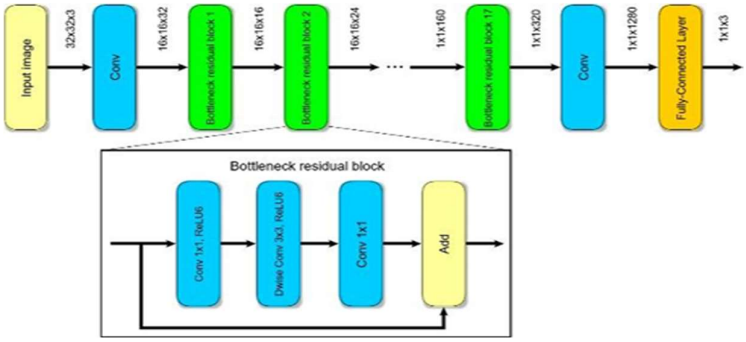

Fig 3.6: MobileNetV2 CNN

The model is created using a convolutional neural network called MobileNetV2. The first layer conv2D layer performs an elementwise multiplicative function on the two-dimensional input data as it slides over it. The outputs will henceforth be combined into a single output pixel. The overfitting issues are lessened by the second layer dropout. The last layer is used to normalize the inputs of the previous layers by re-centering and re-scaling, which speeds up and improves the stability of neural network training.

How well a neural network reflects the training data is determined by comparing the target and anticipated output values using a loss function.

The model is trained and prepared for testing after completing all the training processes.25% of the dataset was used as test data. A green box enclosing the text that was recognized in the natural scene image is output after the text's identification. gTTs are now used to transform the recognized text into speech.

4.0 RESULTS AND DISCUSSIONS

We have a dataset comprising 461 photos of natural scenes and the ground truth information that goes with them. To do this, we boosted the model's accuracy by raising the number of epochs to 30, while also lowering the loss. Several models were produced through training for various datasets. The accuracy percentage was raised by creating models for the numerous natural scene photos. Eventually, the accuracy acquired after 30 training epochs is around 86.7%.

```
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
87/87 [==============================] - ETA: 0s - loss: nan - accuracy: 0.8676
```

Fig 4.1: Accuracy

An area of an image that you wish to filter or manipulate in any manner is called a region of interest (ROI). A binary mask picture can be used to represent an ROI. Pixels in the mask picture that are part of the ROI are set to 1, while pixels outside the ROI are set to 0.

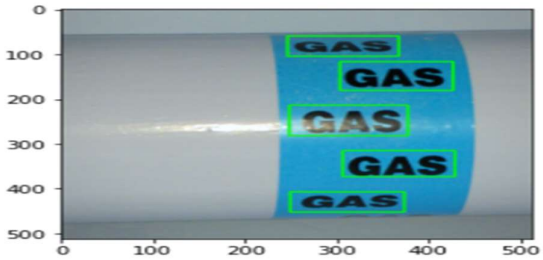

Fig 4.2: Region of Interest(1)

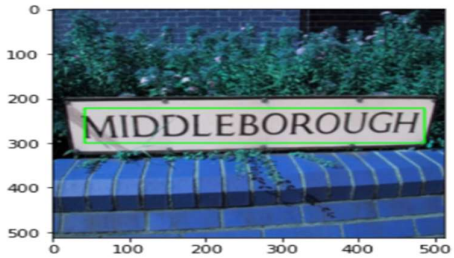

Fig 4.3: Region of Interest(2)

From the region of interest, the text is extracted and recognised.

```
In [26]: extractedInformation
Out[26]: 'MIDDLEBOROUGH f\n\n'
```

Fig 4.4: Extracted information

Fig 4.5: Recognised text speech

For better user interface we have added GUI for the whole model using django framework.

Fig 4.6: GUI interface 1

Fig 4.6: shows the first page to the user when the website is opened. It has option to upload image and to scan the image.

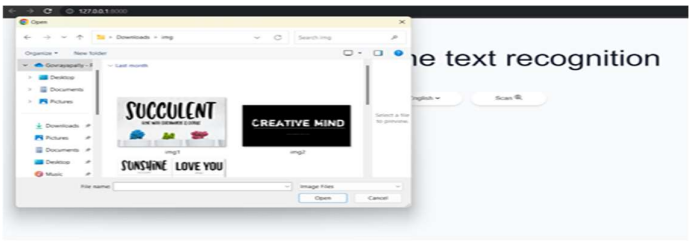

Fig 4.7: GUI interface 2

Fig 4.7: On clicking upload image it will open a file explorer inorder to select the image.

Fig 4.8: GUI interface 3

Fig 4.8: The selected image is displayed on the user interface. If the user clicks on scan button without selecting image, error message will be displayed.

Fig 4.9: GUI interface 4

Fig 4.9: The text is recognised and displayed on the user interfaceand a copy text feature is added for user future reference.

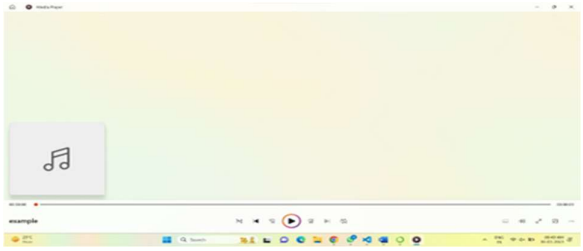

Fig 4.10: GUI interface 5

Fig 4.10: The text from the natural scene image is converted to speech as an audio file.

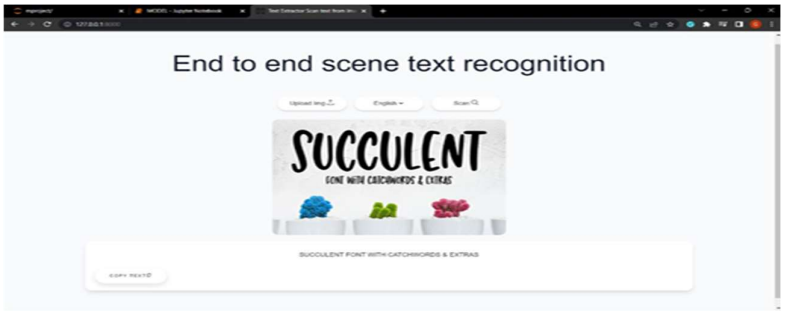

Fig 4.11: Overall GUI interface

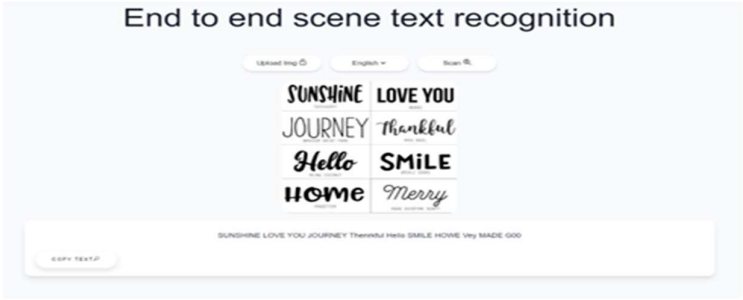

Fig 4.12 Output for different fonts

Fig 4.13 Output for a black background

CONCLUSION

This project aims to extract the words from the photos of nature scenes and convert them to text. This study demonstrates how convolutional neural networks can recognize various text typefaces in photos more precisely. This project is designed to exclusively recognize text written in English. The project first determines the Region of Interest, which is followed by text conversion of the determined regions. The generated text is also rendered as a voice to make it more user-friendly and accessible to those who are blind. Convolution neural networks are used in this project to identify the region of interest for any text that is present in the image. CNN receives photographs of real-world scenes to make predictions in real-time. CNN filters such as max pooling, convolution, and flattening are used in this project. This device is sturdy and yields accurate readings even with diverse typefaces. Our model has performed well in predicting outcomes.

REFERENCES:

1. Revathy A S, Anitha Abraham, Jyothis Joseph: A Survey on Text Recognition from Natural Scene Images. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.
2. Luo, Canjie, LianwenJin, and Zenghui Sun. "Moran: A multi-object rectified attention network for scene text recognition." Pattern Recognition 90 (2019): 109-118.
3. Zhou, Xinyu, et al. "East: an efficient and accurate scene text detector." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.
4. Yan, Ruijie, et al. "MEAN: multi-element attention network for scene text recognition." 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.
5. Huang, Weilin, et al. "Text localization in natural images using stroke feature transform and text covariance descriptors." Proceedings of the IEEE international conference on computer vision. 2013.

6. Dai, Pengwen, Hua Zhang, and Xiaochun Cao. "SLOAN: Scale-adaptive orientation attention network for scene text recognition." IEEE Transactions on Image Processing 30 (2020): 1687-1701.
7. Chandio, Asghar Ali, et al. "Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network." IEEE Access 10 (2022): 10062-10078.
8. Mokayed, Hamam, et al. "A New Defect Detection Method for Improving Text Detection and Recognition Performances in Natural Scene Images." 2020 Swedish Workshop on Data Science (SweDS). IEEE, 2020.
9. Özgen, Azmi Can, MandanaFasounaki, and Hazim Kemal Ekenel. "Text detection in natural and computergenerated images." 2018 26th signal processing and communications applications conference (SIU). IEEE, 2018.
10. He, Wenhao, et al. "multi-oriented and multi-lingual scene text detection with direct regression." IEEE Transactions on Image Processing 27.11 (2018): 5406- 5419.
11. Donahue, Jeff, et al. "End-to-end adversarial text-tospeech." arXiv preprint arXiv:2006.03575 (2020).
12. Bagi, Randheer, Tanima Dutta, and Hari Prabhat Gupta. "Cluttered textspotter: An end-to-end trainable lightweight scene text spotter for cluttered environment." IEEE Access 8 (2020): 111433-111447.