International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 5, July-August 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Text and Object Recognition using Deep Learning for Visually Impaired People

R. Soniya[1], B. Mounica[1], A. Joshpin Shyamala[1], Mr. D. Balakumaran[2]

[2]Assistant Professor, M.E,
[1,2]Department of Electronics and Communication Engineering,
[1,2]S. A. Engineering College, Chennai, Tamil Nadu, India

ABSTRACT

The main aim of this paper is to aid the visually impaired people with object detection and text detection using deep learning. Object detection is done using a convolution neural network and text recognition is done by optical character recognition. The detected output is converted into speech using text to the speech synthesizer. Object detection comprises of two methods. One is object localization and the other is image classification. Image classification refers to the prediction of classes of different objects within an image. Object localization infers the location of objects using bounding boxes.

KEYWORDS: *object detection, CNN, text detection, OCR, TTS*

How to cite this paper: R. Soniya | B. Mounica | A. Joshpin Shyamala | Mr. D. Balakumaran "Text and Object Recognition using Deep Learning for Visually Impaired People" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.624-628, URL: www.ijtsrd.com/papers/ijtsrd31508.pdf

IJTSRD31508

I. INTRODUCTION

Although much advancement has been made in the image processing field, extracting objects from the color image is not an easy task. This paper deals with taking out the objects and text alone from a frame. The object is recognized by passing the input video to a trained model which leads to the classification of objects between various classes. The model is trained by MS-COCO (Common Objects in Context) [1] and further tuned finely by using PASCAL VOC0712 [2] has some pre-trained weights. It makes use of single-shot detectors (SSD)and mobile nets result in fast and real-time object detection.

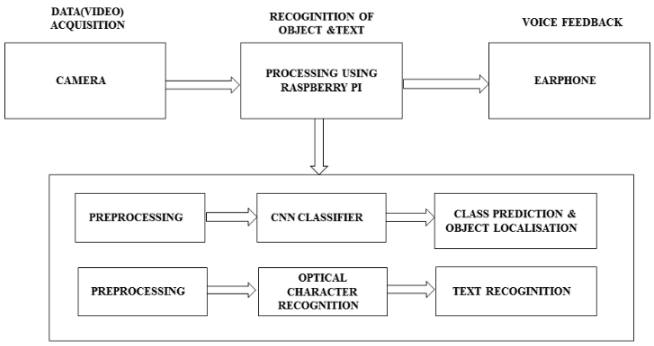

Figure 1:Overall diagram for detection and recognition

A Text detection model is based on a fully convolution neural network [2][3]. It was adopted to centralize the text regions.

Generally, the feature of the actual input image is first extracted. The one convolution is applied to output dense per-pixel predictions of text to be present. For each positive sample, the channel assumes a probability. As referred in figure:1 The preprocessed image input is forwarded to the neural network.

There it compares the input with the trained dataset after detecting the object and text. It generates output as to what object or text is that. This is followed by text to speech conversion process to help the visually impaired person.

II. CONVOLUTION NEURAL NETWORK

CNN [3] is primarily used for pattern recognition within the image, thus making it suitable for the image-focused task by reducing parameters to set up a model. It consists of three layers namely, convolution layer, pooling, fully connected layer. Figure 1: Shows simplified CNN architecture comprising of feature extraction and classification.

Convolution layer performs convolution operation between an array of input pixels and kernel of a particular size which slides all over the input frame

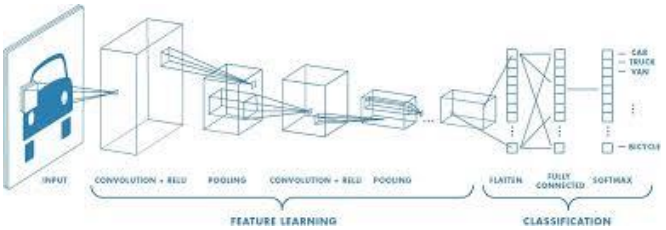

Figure 2: Simplified CNN Architecture

The resulting output is passed through the activation function (Rectified linear unit) in-order to remove nonlinear functionalities figure (2). The Pooling layer simplifies the overall dimension of the image by performing the down sampling operation. It operates over each input activation map and scales using the "MAX" function. Most of the CNN contains a max-pooling layer with a kernel of 2x2 size.

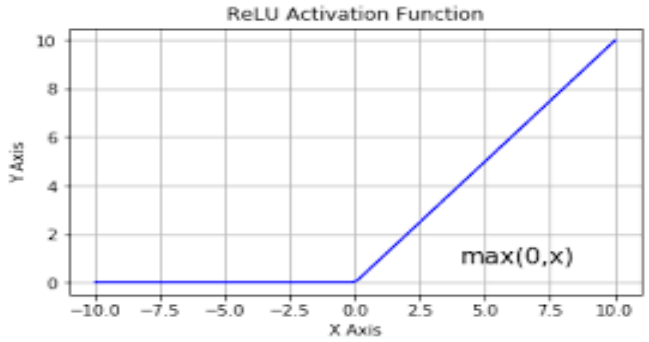

Figure 3: Rectified linear unit (ReLU) activation function

The output from the final pooling or convolution layer is flattened and fedas input to a fully connected layer. The final layer uses soft max activation function which is used to classify based on probabilities obtained.

III. CAFFE OBJECT DETECTION MODELS

Caffe [4] (convolution Architecture for fast feature embedding) is a deep learning framework that supports different architecture for image classification and segmentation. The following are the model used for object recognition.

A. Single-shot detectors:

It is the combination of Yolo's regression and a faster RCNN mechanism [5]. It simplifies the complexity of the neural network. The local feature extraction method is effective in SSD where multiscale feature extraction. The recognition speed of SSD is 59 fps (frames per second). It quenches the need for real-time implementations. The only demerit is that its fragile detection of tiny objects. At the time of prediction, the network generates a default box and adjusts it to match with the object shape [6],[7].

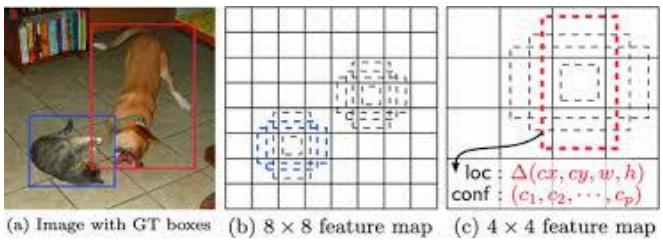

Figure 3:localization and confidence using multiple bounding boxes

B. Mobile Nets

While developing object detection networks we generally utilize existing network architecture, such as VGG or Res Net The limitation is that these network architectures can be very huge in the order of 200-500MB. So it doesn't support resource-constrained systems. Mobile nets differ from traditional CNN through the usage of *depth-wise separable convolution* (**figure 4)[7],[8].**

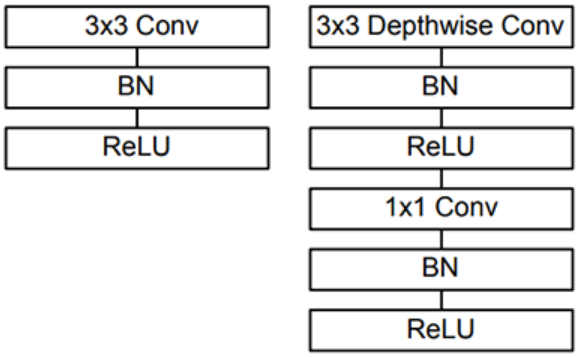

Figure 4: *(Left)* Standard convolutional layer with batch normalization and ReLU. *(Right)* Depth- wise separable convolution with depth-wise and pointwise layers followed by batch normalization and ReLU (figure and caption from Liu et al.).

Here, convolution is divided into two stages of 3x3 depth wise convolutions followed by a 1x1 point wise convolution.

IV. TEXT RECOGNITION

Text recognition utilizes the help of OCR to detect the text from an image. OCR is a popular model use to get the character from the text. OCR is a self learning translator of character in typed, printed or any documented and pictured forms. It has various methods to recognize characters like thinning thickening, feature extraction [9]. The major steps corresponding to text detection comprises data collection, pre processing, segmentation, features extracting and classification. In case of first step, pre trained data sets or own data sets can be used for the research. Pre processing includes slant detection, slant correction, gray scale conversion, normalization; noise removal etc. segmentation is done to splits the words from sentences for better detection. The feature extraction step extracts the features from the separated words and compares it with features of trained images to confirm with the matching classes. Finally classifiers are used with respect to convenience for the classification process.

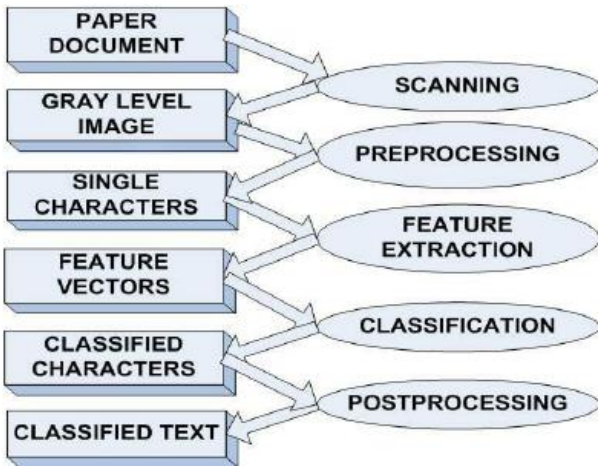

Figure 5:Basic steps of OCR[10]

A. Data collection:
It involves acquiring the text from a frame in the form of video of resolution using webcam [9]

B. Preprocessing:
Once the data is collected, various preprocessing techniques are used to enhance and to improve the quality of the image. It is an important stage prior to feature extraction because it determines the quality of the image in the successive stages [10].

C. Segmentation:
Segmentation is nothing but separating the various letters of a word. Only if there is a good segmentation, better recognition rate can be obtained [11].

D. Feature extraction:
It is a stage where the required information is obtained by using minimum number of resources. It also reduces the overall dimensions of the image.

E. Classification:
It is process of categorizing a character into its suitable category. There are different function to classify the image[12],[13].

V. TEXT TO SPEECH SYNTHESIZER

Text to speech system (TTS) converts text into voice using a synthesizer [14] which produces artificial speech of human.It converts an arbitrary plain text into a corresponding waveform.

The main components of this system are text processing and speech generation [15]. The text processing unit is responsible for producing a sequence of phonemic units according to the input text and also includes text normalization and tokenization. Then makes the text into prosodic units, which is termed as phonetic (also called as text-to-phoneme or grapheme-to-phoneme conversion). These are then realized by speech generation system. A general text to speech system is shown in figure 5[16]

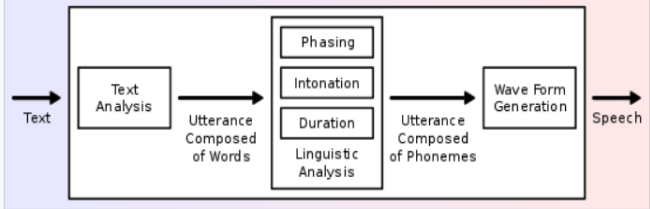

Figure 5: A overview of TTS system

The Text normalization involves transforming the whole text into standard format [17].Tokenization are used to split the whole text into words, keywords, phrases and symbols it also discards some punctuation like commas.

Speech synthesizer has to generate sounds to make those words in input text. There are 40 phonemes available for 26 letters in English alphabet it is because some letters can be read in multiple ways. But this may get harder for practical implementation commonly called as prosody, in linguistics. Within a word, a given phoneme may sounds differently because the phonemes that come before and after it.

Thus grapheme is the alternative approach where words are segmented into individual letters. Once the sequences of words are converted into list of phonemes, then it will be converted into speech. Following are the different speech synthesis techniques.

A. Concatenative synthesis
It is a synthesizer that make use of preloaded record of human voices which is referred to as units. The units contains lots of examples of different things, break the spoken sentences into words and the wordsinto phonemes[18]. Thus it can easily rearrange to create new words and sentence.The duration of units is defined and are in the range of 10 ms up to 10 seconds.

B. formant synthesis
It is created by using additive synthesis and an acoustic model instead of human speech input during runtime[19]. Factors like noise, frequency and voice are changed with time to produce signals of artificial speech .This is often termed as rules based synthesis. Most of the systems that use formant synthesis have robotic or artificially generated voice which are unique from human speech. It is trusted for its intelligence and high speed, eliminating acoustic flaws that usually infect concatenative systems. Visually impaired use high speed synthesized speech to navigate through computer screens Using screen reader. Formant synthesizers normally do not have a database of speech samples so, they are found to be smaller programs than concatenative systems.

C. Articulatory synthesis
It is computational technique used for synthesizing speech with respect to model of human vocal tract (voice process). It is a mechanism to mimic a human voice[19].

VI. EXPERIMENTAL SETUP

A. Object Detection
To execute object detection using CNN it makes use of the Caffe framework for deploying the object detection model. The datasets include 20 classes [7], [20].

The model contains apre-trained. Caffe model file, many. Proto txt files. Which has its parameterized structure within the proto text file and also weights are predefined.

Thus a single frame of the input video is resized to the dimension of 300x300(height & width) and forward passed into the model. So that it can detect 19 objects in an image. Multiple objects can be detected from a single frame.

B. trained dataset
The dataset contains the following classes including *airplanes, boats, bicycles, bottles, birds, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, potted plants, people, sofas, trains,* and *television monitors.*

The probability during each detection is also checked. If the confidence level is above the threshold level defined the particular class name which is referred to as label along with their accuracy will be displayed and also a bounding box is computed over the objects detected

And also the label detected will be created as a separate text file which helps in producing the audio output of the object detected.

A. labels and their accuracy during detection

TABLE I Different objects and text fonts that can be detected

OBJECTS
AEROPLANE
BICYCLE
BIRD
BOAT
BOTTLE
BUS
CAR
CAT
CHAIR
COW
DINING TABLE
DOG
HORSE
MOTORBIKE
PERSON
POTTEDPLANT
SHEEP
SOFA
TRAIN
TV MONITOR

(a)

TEXT FONTS
TIMES NEW ROMAN
NUMERALS
ARIAL BLACK
CALIBRI
ITALIC
GEORGIA
CASTELLAR
ROCKWELL
ALGERIAN
STENCIL

(b)

Since the labeling had to be done manually and also because of the size of RAM we couldn't produce enough data, or data of assured quality, to properly train the neural network model. We have reached the possible accuracy for a CNN with images as input data, we have only created a model for detecting 20 objects.

With the help of OCR it detects the different font styles with approximate predictions of the word or sentence. Table 6.1(a) shows the list of objects and Table 6.1(b) shows list of text fonts which can be detected by the proposed algorithm. Along with the detection it also produces voice output of the detected object and text through the earphone connected to the audio jack.

As far as the results are concerned the following are observed to be achieved from the project. In case of text, it is possible for the device to detect from either image inputs or live reading of the texts. It is capable of recognizing standard text formats of different fonts. Especially bold letters can be detected fast and accurately. It is noted that words in light background are recognized well with precision.

These results are based on the frame intensity and pixel variations. The outcome is subject to vary in accuracy depending on the camera specifications and focus done to the object. The accuracy of the predicted objects differ due to different condition such as contrast sensitivity, illumination, noise and also it produces maximum accuracy if the object is closer to the camera in the range of 70 cm with good lightning condition.

VII. CONCLUSION

This device is capable of acting as a smart device. It provides visually impaired people with the list of objects in the surrounding along with accuracy of prediction. Therefore the person will get to know about the hindrances in his path. This is more like a voice assistant that tells name of the things around us. Multiple objects in a single frame is detected which is an added advantage.

REFERENCES

[1] T.-Y. Lin, J. Hays, M. Maire, P. Perona, S. Belongie, D. Ramanan, L. Bourdev, L. Zitnick, R. Girshick and P. Dollar, "Microsoft COCO: Common Objects in Context," arXiv:1405.0312v3, pp. 1-15, February 2015.

[2] X. Zhou et al., "EAST: An efficient and accurate scene text detector", Proc-30th IEEE conf.Comput.vis. Pattern Recoginition, CVPR 2017, vol.2017-janua, pp.2642-2651, 2017.

[3] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network", Proc. IEEEComput. Soc. Conf. Comput.Vis. Pattern Recognit., pp.5676-5685, 2018.

[4] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303-338, 2010. ****

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convo-lutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)****

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, andA. C. Berg, "Ssd: Single shot multibox detector," in ECCV, 2016.(ssdku)***

[7] Liu, Wei, et al. "Ssd: Single shot multibox detector." European conference on computer vision. Springer, Cham, 2016.

[8] A. G. Howard, M. Zhu, B. Chen and D. Kalenichenko, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision," 2017.

[9] "mobilenet_ssd pretrained model" https://github.com/chuanqi305/MobileNet-SSD/blob/master/README.md

[10] A Survey on Optical Character Recognition System Noman Islam, Zeeshan Islam, Nazia Noor, Journal of Information & Communication Technology-JICT Vol. 10 Issue. 2, December 2016.

[11] Lund, W.B., Kennard, D.J., & Ringger, E.K. (2013). Combining Multiple Thresholding Binarization Values to Improve OCR Output presented in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA: SPIE.

[12] Shaikh, N. A., & Shaikh, Z. A, 2005, A generalized thinning algorithm for cursive and non-cursive language scripts presented in 9th International Multitopic Conference IEEE INMIC, Pakistan, 2005. Pakistan: IEEE

[13] Shaikh, N. A., Shaikh, Z. A., & Ali, G, 2008, Segmentation of Arabic text into characters for recognition presented in International Multi Topic Conference, IMTIC, Jamshoro, Pakistan, 2008.

[14] Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J, 2011, Convolutional neural network committees for handwritten character classification presented in International Conference on Document Analysis and Recognition, Beijing, China, 2011. USA: IEEE.

[15] Archana Balyan, S. S. Agrwal and Amita Dev, Speech Synthesis: Review, IJERT, ISSN 2278-0181 Vol. 2 (2013) p. 57 – 75.

[16] E. Cambria and B. White, "Jumping NLP curves: a review of natural language processing research," IEEE Computational Intelligence, vol. 9, no. 2, pp. 48–57, 2014.

[17] *van Santen, Jan P. H.; Sproat, Richard W.; Olive, Joseph P.; Hirschberg, Julia (1997). Progress in Speech Synthesis. Springer. ISBN 978-0-387-94701-3.*

[18] Liu, F., Weng, F., Jiang, X.: A Broad-Coverage Normalization System for Social Media Language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Number July (2012) 1035–1044.

[19] Mark Tatham and Katherine Morton, Developments in Speech Synthesis (John Wiley & Sons, Ltd. ISBN: 0-470-85538-X, 2005)

[20] "Deep learning for computer vision with Caffe and cuDNN" NIVIDIA Developer Blog. October 16,2014

[21] (ocr) "A Complete Workflow for Development of Bangla OCR" International Journal of Computer Applications (0975 – 8887) Volume 21– No.9, May 2011