International Journal of Engineering Sciences Paradigms and Researches: (Vol. 27, Issue 02) and
(Publishing Date: January 2016)

A Smart Assistant for Visually Impaired using Deep Learning Techniques

Satyajit Nayak[1], Saswati Sahoo[2] and Dr. S Krishna Mohan Rao[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: January 15, 2016

Abstract

Increasing pollution and changing life styles has severely affected human health specially our sense organs. More exposure to screen has increased vision related problems even at very early age of life. The developing technologies should be utilized to help the persons with no or very less vision to lead an independent life in society. Computer vision is one such field that can be utilized to develop some cost effective products that can be very useful for these scenarios. The detection and recognition of text from natural image can be very useful for visually imparted persons as well as in various other applications like developing a smart system to help driver in getting voice signal for every road sign, and even warning if we did not follow the one. The proposed work uses deep convolutional neural network to implement a text detection and recognition system that is much simpler and faster as compare to traditional hand crafted feature based methods.

Keywords: *Visually Impaired, deep neural network, Smart Assistant.*

I. Introduction

"VISION" is one of the most precious gift we have received from nature. But many among us could not receive this gift or lost this gift because of different reasons. Life of all these people becomes very difficult and they need to be dependent on others for each work. One of main reason for this is excessive exposure to screen. Many measures are already taken the government as well as non-government organizations to help such persons.

Artificial intelligence has lead to many smart devices that can help human in various fields of life. These technologies can also be utilize to help visually impaired persons. Deep convolutional neural network is one such technology that has made object detection face detection possible.

The computer vision is the key to develop various products that can help to provide artificial vision to various people. This artificial vision can be for face detection, object detection, text detection and recognition or the combination of all of these. The present work is a initial step for development of one such cost effective and easily portable or wearable device. The current project considers a scenario where a person suffering from a visual impairment needs a tool to carry around and receive a voice signal for the texts that are available around him. This will help in getting information from sign boards at various places.

Some products are available in the international market like one shown below but they are very costly (between 1500$ to 2000$):

Assisted Vision Smart Glasses: They are constructed using transparent OLED displays, two small cameras, a gyroscope, a compass, a GPS unit, and a headphone. They are based on the concept of making the thing which is close to the camera brighter as compare to other far things. Thus a person, who can distinguish light and dark, can use them for obstacle detection. The main problem with theses glasses is they are very costly and cannot identify text from images.

Another device developed by a Eyra, 'Horus', that can recognize face, text and objects. It is wearable device to be worn on head like a belt that is having a camera. The camera images are given to GPU where a pre-trained deep neural network is saved to process the image and describe the image for the person in form of voice. The size of the device is same as that of smart phone along with wearable belt as shown in Fig.1.

Apart from this one more device, available in market is "figure reader". This is also a wearable device but to be wore on figure like a ring. When user points his figure on text that text will be recognized by the device and speaker will give voice for that text. This device alert the user by giving small vibration if there is deviation of the pointing figure from the line.

Microsoft AI and research has developed an application named 'seeing AI' that helps person with no or less vision to use smart phone as virtual vision device. This app help them in detecting objects, color, text etc.

But when the exact location of text is not known or the distance between the user and text is much more, these scanner based devices will not be much affective.

In India many researchers are working in the same field to utilize artificial intelligence and deep learning neural network to help blind people. Dr. Amit Ray is working in Compassionate AI lab, to utilize AI for the benefit of blinds. In 2017, BrailleMe, the award-winning product developed by Surabhi Srivastava, IIT Bombay, made it possible for the visually impaired to access any digital information instantaneously in their own tactile script. The price of the product was just 300$.

Fig.1: The Horus: wearable, about the size of smartphone

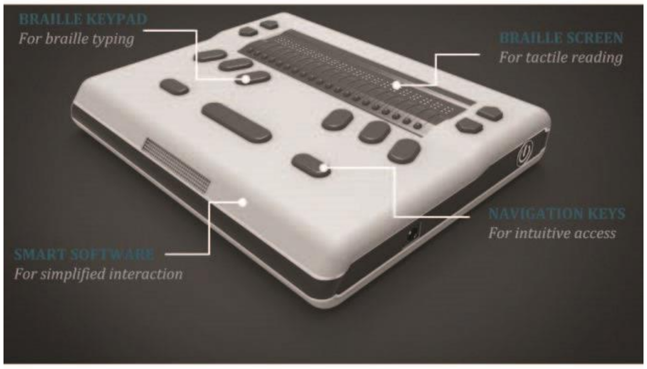


Fig.2: BrailleMe

II. Procedure

A. Text detection and recognition system

Feature detection from an image, without giving the exact information about the feature, is what a deep convolutional neural network can do. This can be utilized to detect the text from any natural scene image with varying backgrounds. Deep convolutional neural network need to be trained properly with sufficient data set to achieve the goal.

The proposed system enables a visually impaired person to understand text on sign boards, banners, hoardings. This system captures the image from its surroundings using a camera and the image will be internally processed and the speech output is given through the speaker or earpiece connected to it.

There are two main blocks: image processing block and voice processing block, In the image processing block the image captured using camera is converted to text. In voice processing block the output of the previous block i.e., text extracted from the captured image is converted to speech. For convenience of the user the voice may be altered to masculine or feminine voice.

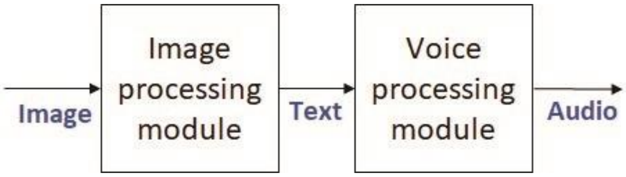


Fig.3: Block diagram

Image processing block ● Image is acquired using a camera.

- Text is separated from the image after processing in a pre-trained deep network.
- Here we get .txt file from a .png file or .jpg file

Voice processing block

- Now the text file is further converted into speech using a text to speech synthesizer.
- There are two ways to do this: one is text to phoneme conversion where text is compared with the words present in dictionary and giving output, other one is learning based speech output approach.

B. Implementation

OpenCV is installed in the Raspberry Pi to perform the image processing. OpenCV is Open Source Computer Vision which is a set of libraries including all the programs that support real-time computer vision. This OpenCV installed in the Raspberry Pi supports in executing the image processing captured with the camera.

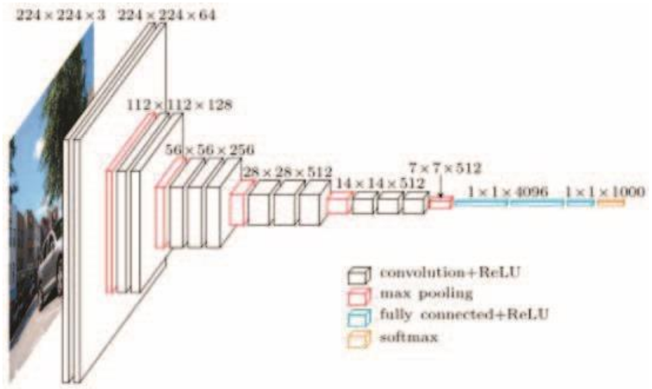
The camera used in the project is the Pi Camera which is interfaced with the Raspberry Pi using specific commands. The Speaker can be of a standard audio output device such as Headphones / Earphones which helps the user to listen to the voice output.

Before the start of deep learning era various hand crafted feature based methods were proposed to detect text from natural scenes. These methods are based on technologies like Maximally Stable Extremal Regions (MSER) and Stroke Width Transform (SWT). But in deep learning era end to end training is possible for a deep convolutional neural network to detect the text from natural scene. The only requirement is large data set to train the system and a proper system that can carry such large data set.

Fig.4 Convolutional Neural Network

C. Convolutional neural network

Artificial intelligence is bridging gap between humans and machine constantly. It helps the machine to see the world as a human. This can be done by deep learning where output is predicted for a given input. Here **C**onvolutional **N** eural **N**etwork (CNN) , a deep learning algorithm which takes input and allots confidence to different characteristics in the input image which helps in differentiating one from another is used.

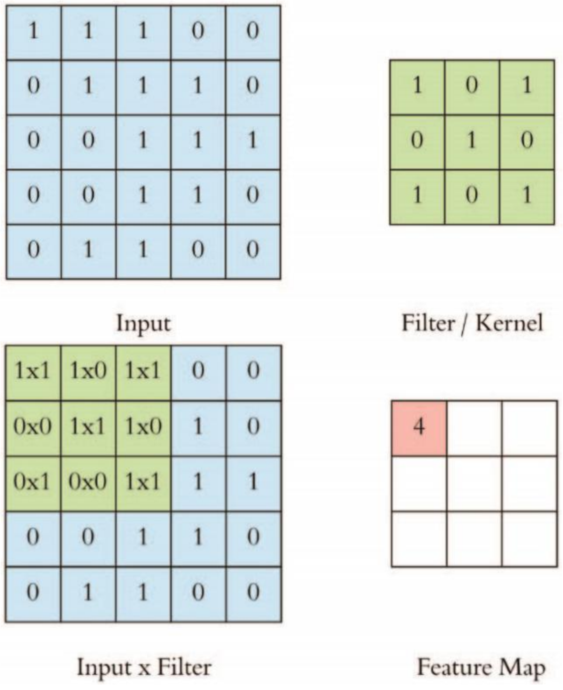


A series of convolution and pooling operations are performed followed by fully connected layers to get the output for a given input.

Convolution layer is the main building block CNN. Convolution means merging two sets of data. Here, the filter, also called as kernel is convolved with input image to get a featured map.

Consider the Fig.5, where blue area is the input and green area is the filter. The filter slides over the input, where, matrix multiplication is done element wise and the sum gives feature map. In the above mentioned example a 2D matrix is considered but in general the image is taken as a 3D matrix of pixels. Usually many filters which slide on input and result in different feature maps are combined together to get a single output from the convolution layer. Stride tells us by how much value the filter slides over the input. Generally, the stride value is 1.

In fig.5 feature map size is not same as the input. So, padding can be used in the image by bordering the input with zeros, so that the dimensions of both input image and feature map will match. By doing this the possibility of image to shrink is eliminated.



Input

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Filter / Kernel

1	0	1
0	1	0
1	0	1

Input x Filter

1x1	1x0	1x1	0	0
0x0	1x1	1x0	1	0
0x1	0x0	1x1	1	1
0	0	1	1	0
0	1	1	0	0

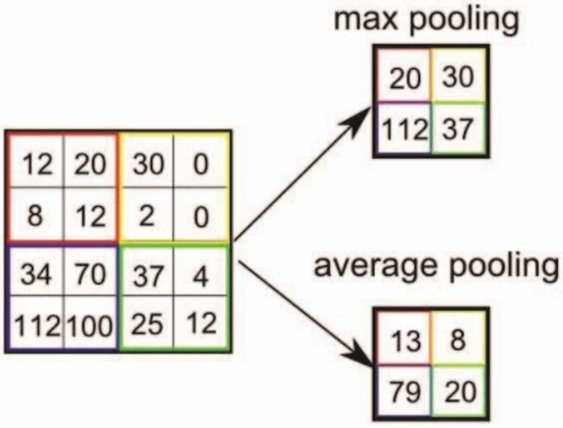
Feature Map

4		

Fig.5 An example showing input, filter and Filter sliding over the input

Pooling is used to reduce the dimensions. The height and width of the feature map is reduced but the depth is maintained the same. There are two types of pooling

- max pooling
- average pooling



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

max pooling

20	30
112	37

average pooling

13	8
79	20

Fig.6 Types of pooling

Max pooling which is commonly used type of pooling considers the maximum value from the pooling window. Average pooling takes all the values from a pooling window and computes the average value as shown in fig.6.

Fully connected layers are added after the convolution and pooling layers to complete the CNN architecture. The output from the convolution and pooling layers is 3D but the output from the fully connected layer is 1D.

So, the output from the last pooling layer is flattened to make it 1D.

Text detection is a tough task because

- The light may be so bright causing saturation of the image or the light may not be sufficient enough
- The surface of the text may be reflective which make it tough to capture the image due to reflection and refraction phenomena's.
- The resolution of the camera may below standard value
- When compared to a scanner the sensor noise is high of a camera
- The images may be blurred at times
- The text may be at an angle which makes it hard to detect the text

III. HARDWARE REQUIREMENT

The components required for the hardware implementation are as follows

- Raspberry Pi
- Power supply
- Camera
- Speaker
- Mouse/ Push button
- HDMI cable

Raspberry pi is a small computer whose size is of a credit card. A mouse and keyboard can be used to operate it when connected to a display. We have chosen raspberry pi as it supports python, the language in which the code is written. And also, the cost of the pi is low and it is portable. Here, we are using Raspberry pi 3 B+ model.

Specifications :

Raspberry PI 3B+:

- SOC - Broadcom BCM2837B0
- CPU - 1.4 GHz
- Memory - 1GB
- Networking - Ethernet, 2.4/5 GHz wireless
- Storage - MicroSD slot
- 40 pin GPIO
- Power Source - 5V
- Ports: HDMI, audio-video 3.5mm jack, 4xUSB, Camera Interface, Display Interface, Ethernet.

Pi camera v2 :

- Sony IMX219 Sensor.
- 8 MP camera capable of taking picture of 3280 x 2464 pixels.
- Capturing video at 1080p 30fps , 720p 60fps and 640 x 480p 90fps resolutions.
- Supports the latest version of Raspbian OS.
- Supports Raspberry Pi 1, Pi 2 and Pi 3 and Models A, B and B+.
- Applications of Pi Camera: CCTV security, auto motion detection, time lapse photography.

Power supply:

- PSU Current Capacity : 2.5 A
- Total Peripheral Current Draw from USB : 1.2 A ● Active Current Consumption from Bare-Board : 500 mA

Speaker:

- Earphones can also be used as audio output.
- Standard Speaker can be used with audio amplifier.
- Bluetooth speaker can also be used for wireless audio output.

Display Output:

- 1.3/1.4a HDMI cable
- Transfer speed of up to 10.2 Gbps

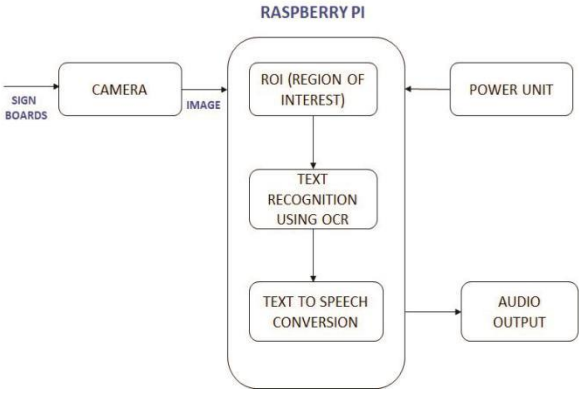


Fig.7 Hardware block diagram

IV. CONCLUSION

The result of text detection and recognition blocks are shown in Fig.8, Fig. 9 and Fig.10. it can be seen that the text in various images with different backgrounds is detected successfully by the system. The performance of the system can be further improve by using a suitable GPU for training the deep convolutional neural network.

The text detected is converted to voice by application available for text to voice conversions.

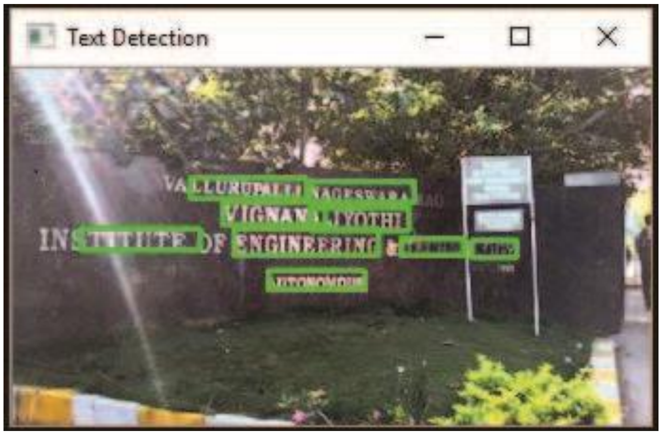

Text Detection
VALLURUPALLI NAGESWARA RAO
VIGNANA JYOTHI
INSTITUTE OF ENGINEERING & TECHNOLOGY
AUTONOMOUS

Fig.8 output of text detection block from a picture taken at main gate of VNRVJIET

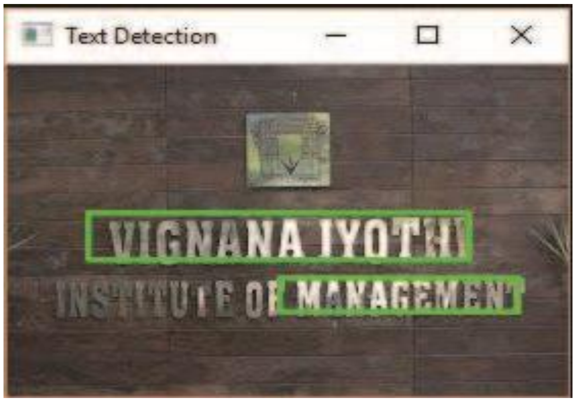
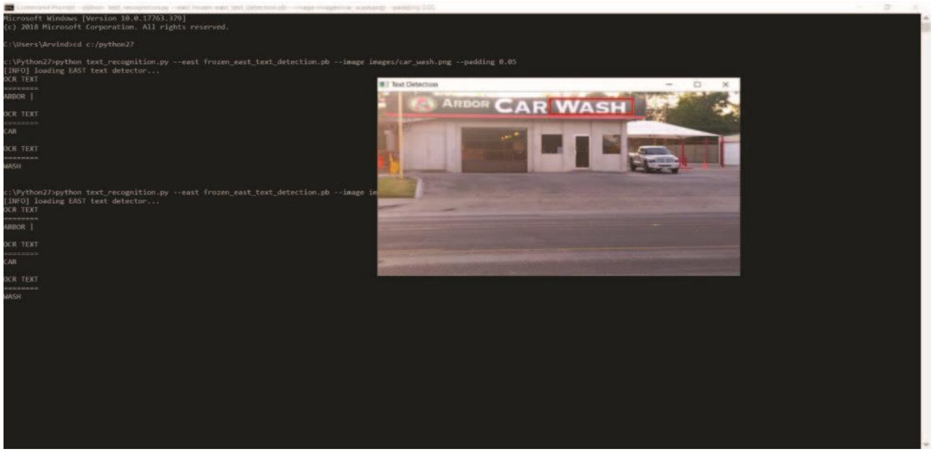


Fig.9 output of text detection block from another picture taken at VNRVJIET



```
Microsoft Windows [Version 10.0.17763.379]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\Arvind>cd c:/python27

c:\Python27>python text_recognition.py --east frozen_east_text_detection.pb --image images/car_wash.png --padding 0.05
[INFO] loading EAST text detector...
OCR TEXT
========
ARBOR |

OCR TEXT
========
CAR

OCR TEXT
========
WASH


c:\Python27>python text_recognition.py --east frozen_east_text_detection.pb --image im
[INFO] loading EAST text detector...
OCR TEXT
========
ARBOR |

OCR TEXT
========
CAR

OCR TEXT
========
WASH
```

Fig.10 output of text detection and recognition block for a sample picture

References

[1] Kwang In Kim, Keechul Jung "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm" - Pattern Analysis and Machine Intelligence, IEEE

[2] Chucai Li et.al "Portable Camera-Based Assistive Text and Product Label Reading From Hand-Held Objects for Blind Persons", IEEE Transactions on Mechatronics, June 2014

[3] R. Lienhart and A. Wernicke entitled "Localizing and segmenting text in images and videos," ,IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 4, pp. 256 –268, 2002.,

[4] Chucai Li and Ying Li Tian entitled "Text string detection from natural scenes by structure based partition and grouping," IEEE Trans. Image Process., vol. 20, no. 9,pp. 2594– 2605,Sep. 2011