Cognitive model for object detection based on speech-to-text conversion.

PAVULURI, J., SAI, T.V., MANNAM, R.K., MANIDEEP, R. and BANO, S.

2020

Cognitive Model for Object Detection based on Speech-to-Text Conversion

Pavuluri Jithendra
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
jithendrapavuluri43@gmail.com

Tummala Vinay Sai
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
vinaiinfo18@gmail.com

Raj Kumar Mannam
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
raj9948622139@gmail.com

Ramini Manideep
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
manideepramini@gmail.com

Shahana Bano
Department Of CSE
Koneru Lakshmaiah Education Foundation
Vaddeswaram, India
shahanabano@icloud.com

Abstract— **The goal of this paper is to develop a model which is the integrated version of both SpeechRecognition and Object detection. This model is developed after undergoing the literature survey and the existing models that are related to Object Detection and Speech Recognition. There are several types of Speech Recognition and Object Detection models available so far. In addition to the existing models, this paper proposes a new model named "**Cognitive Model for Object Detection based on Speech-to-Text Conversion," **which is an integrated version of both Speech Recognition and Object Detection models. Firstly, A speech command is provided as an input to the model, it takes the command and processes the data, and then it detects the specified object from a source of images. The detected object is represented with a rectangular box. This approach is implemented with the help of Google Speech Recognition and YOLO object detection models utilizing the Darknet and OpenCV frameworks.**

Keywords— SpeechRecognition, YOLO, Object Detection, Google cloud GPU, Darknet, LabelImg, OpenCV.

I. INTRODUCTION

Cognitive Models play a vital role in the enhancement of AI-enabled technologies, Speech is one of the finest means of communication. It is used for exchanging feelings and thoughts with one another. Object detection facilitates us to identify and locate objects in an image or video.

As mentioned earlier speech is the finest means of communication and the way of speaking needs to be clear to avoid ambiguity between each other. Few people can orally communicate but cannot type and interact with a computer or device. Considering this aspect, this model is developed to detect a specified object from an image that contains multiple objects based on the speech commands that are provided by the user. This approach can be implemented in several ways, but this model that receives the speech input from the user through a microphone and further, the SpeechRecognition will convert into text, and then the backend process is performed by the Darknet framework and YOLO object detection model which provides us the output as the object that is detected among the other objects that are available in the image. When a user provides the voice input to the system it converts the speech to text and the text is stored as the name of the object in a file named coco names. Later, the images will be trained using the Darknet framework on the cloud using the free GPU and then the weights file and configuration file are generated. Now, the YOLO object detection model comes into the picture. The model checks the confidence score of each pixel of the image. If the object is near to the score of the trained image then a boundary box is generated and the required object is detected.

The user needs to import few packages that are required for implementing the SpeechRecogntion and Object detection modules they are SpeechRecognition and PyAudio packages and for performing computer vision operations on an image the OpenCV library is to be imported. The training of the dataset of images is done with the Darknet Framework. This model of approach is done using the Google SpeechRecongnition model and Deep Neural Networks algorithm named YOLO object detection model.

II. LITERATURE SURVEY

MIT computer scientists [1] have developed a model that identifies and detects objects within an image, based on a spoken description of the image. The input of the model will be an image and an audio caption, the model will highlight in real-time the relevant regions or boundary boxes of the image.

Bishal Heuju, Bishal Lakha, Dipkamal Bhusal, and Kanhaiya Lal Shrestha [2] in 2016 have developed a voice-command-based object recognizing robot using speech and image feature extraction. The robot is developed to recognize the command 'identify' and then it will identify the object when the command is given as 'follow', the robot needs to track the object.

Swetha V Patil and Pradeep N [3] in 2019 have proposed a speech to speech system. Firstly, the user has to give the speech input in any of the four languages among English, Kannada, Hindi, and Telugu, and later the model converts the speech to text using the Google API.

Satoshi Nakamura, Konstantin Markov, Takatoshi Jitsuhiro, Jin-Song Zhang, Hirofumi Yamamoto, and Genichiro Kikui [4] are currently developing a speech-to-speech translation system at Advanced Telecommunication

Research Institute, Kyoto, Japan. It is a multi-lingual speech recognition system that supports Japanese, English, and Chinese languages. Although the better integration of speech to text and image recognition or identification gives a very accurate and better novelty in this research proposed method.

Sandeep Kumar, Aman Balyan, and Manvi Chawla [5] in 2017 have proposed an object detection model named as "Easynet model". Easynet model looks at the complete image during the testing phase so its predictions are informed by the global context. During the prediction time, their model generates confidence scores for the presence of the object in a particular category. The model makes predictions with a Single network evaluation.

Krishnaveni, G., Lalitha Bhavani, B., Vijaya Lakshmi, N.V.S.K. [6] have proposed an enhanced approach for object detection using a wavelet-based neural network. In their work, a novel characterization system called Affluence based Image Classification (AIC) is proposed utilizing a wavelet-based neural network system (WNS).

Inthiyaz, S., Ahammad, S.H., Sai Krishna, A., Bhargavi, V., Govardhan, D., and Rajesh, V. [7] have proposed a YOLO medical image model which computes network as relapse of an input image and builds individual bounding boxes for every associated class object with more accuracy. As the model convolves itself into a lone neural network, jumps straight to the image pixels to bounding box coordinates and object classes.

Mandhala, V.N., Bhattacharyya, D., Vamsi, B., Thirupathi Rao, N. [8] in 2018 have developed an Object Detection model to Assist Visually Impaired People. Their contribution focused on developing computer vision algorithms combined with a deep neural network to assist visually impaired individual's mobility in clinical environments by accurately detecting doors, stairs, and signages, the most remarkable landmarks.

J. Olabe; A. Santos; R. Martinez; E. Munoz; M. Martinez; A. Quilis; J. Bernstein. [9] have developed a Real-time text-to-speech conversion system for Spanish, that accepts a continuous source of alphanumeric characters (up to 250 words per minute) and produces good quality, natural Spanish output as described by the user.

Kishan Kumar; Shyam Nandan; Ashutosh Mishra; Kanv Kumar; V. K. Mittal. [10] have developed a Voice-controlled object tracking smart robot. The robot navigates its way as per the voice-command signal, and it also tracks the desired object. The voice-command signal processing is carried out in real-time, using an on-time cloud server that converts it to text format. The command signal text is then transferred to the robot via Bluetooth network to control its differential drive.

Speech Recognition:

Speech recognition is the ability of a device or system to recognize spoken words and convert them into readable text. Speech recognition is utilized in different fields of research in computer science, linguistics, Natural language processing, and computer engineering. Many modern devices are associated with speech recognition functions for enabling hands-free use of a device.

Existing Models for Speech Recognition are Speech Recognition Using Google Cloud Speech API, Speech Recognition Using Deep Neural Networks, and Speech Recognition Using Hidden Markov Models.

With these existing models, it can be utilized as an application of speech to text translation for further identification of objects.

Object Detection:

Object detection is a computer vision technique that allows us to locate or detect particular objects with a bounding box.

- Input: An image with multiple objects.
- Output: Bounding box around the specified object and a class label for the bounding box.

Existing Models for Object Detection are Custom Object Detection using TensorFlow, YOLO Object Detection using OpenCV, and SpeechYOLO: Detection and Localization of Speech Objects.

III. PROPOSED WORK

A. Importing all the packages :

To run this model, you need to import all the essential packages and libraries. To obtain the speech input, the user needs to import the SpeechRecognition package which contains many inbuilt methods like the listen method and the recognizer_google method. The recognizer method helps to recognize the speech and convert it into the desired text.

Later, you need to install the PyAudio package, this is used to record the voice data of the user through the microphone of the device. When you run the SpeechRecognition module it takes the input from speech and it is converted to text and stored in a file and the content in the file is used as a coco name.

YOLO object detection model helps us to detect the object and to generate a boundary box around the object. The Darknet framework helps to train the model. You

need to download and compile the darknet to work with the GPU in the cloud.

B. Methodology:

This algorithm for Object Detection based on Speech-to-Text conversion works when the user gives a speech command to the system to detect a particular object among the various objects that are available on the screen. Google speech recognition learns from real search engine strings and it can recognize different kinds of accents and it is currently working based on Long Short-term Memory Recurrent Neural Networks (LSTM RNNs). These LSTM RNNs have recurrent connections and memory cells that allow them to remember the data that is provided by the user.

Once the model collects your speech data it breaks down the whole audio data into individual sound waves and these sound waves are converted into a digital format and

then the model finds the most probable word fit in that language by looking at the different Treebanks that are available online, this entire process is done by using some models like Hidden Markov Model and Natural language processing techniques. The next step that is performed by the model is converting the speech into text. So, for converting the speech into text the recognize_google method is used which is available in the speech_recognition package.

After receiving the input and converting it into text the remaining process is done backend with help of the Darknet Framework. The converted text data is stored in a file and the text in the file is used as a coco name, which acts as training data for the Yolo object detection model. Later, the images are labeled using LabelImg software, which is an image annotation tool, the user needs to install this tool and for this model, the annotations should be saved in YOLO format instead of PASCAL VOC format and then you need to click and release the mouse to choose a region in the image to annotate or label the rectangular box as per your desired object name and the annotation will be saved to the specified folder. After labeling the images the user has to upload the labeled image into the google cloud along with the classes file and the file containing the coordinates of the Image. Now the image is trained using the Darknet framework in the cloud using a free GPU.

Firstly, the user needs to install the darknet framework and then mount the google drive so that the model can access the training data and the user needs to customize the configuration file based on the number of classes that they are having in their classes file. Now with all the training data and the configuration file, the User needs to train the model using darknet, and then after running the model for a few iterations the weights file is obtained. This weight file is utilized in running the YOLO object detection model. Now with the help of the coco names file, weights file, configuration file, and the trained image the user needs to run the object detection model. It checks the confidence score of each pixel of the image. If the object is near to the confidence score of the trained image then a boundary box is generated and then the required object is detected and represented with a rectangular box specifying the name of the object.

The dimensions of the rectangular boundary box is generated based on the below equations:

$$center\_x = int\ (detection\ [0] * width)$$
$$center\_y = int\ (detection\ [1] * height)$$
$$w = int\ (detection\ [2] * width)$$
$$h = int\ (detection\ [3] * height)$$

Step 1: Start.
Step 2: Place the microphone near the person or device by which the speech
input is going to be given.
Step 3: The user will give his speech data to the model.
Step 4: If the system recognizes the speech then it is recorded and used for translation of speech to text using Google Speech Recognition API.
Else ask the user to speak clearly.
Step 5: SpeechRecognition package performs the recognition of speech.
Step 6: Speech is converted to text using the recognize_google method.
Step 7: The obtained text is stored in a coco names file.
Step 8: Labelling the images using LabelImg software.

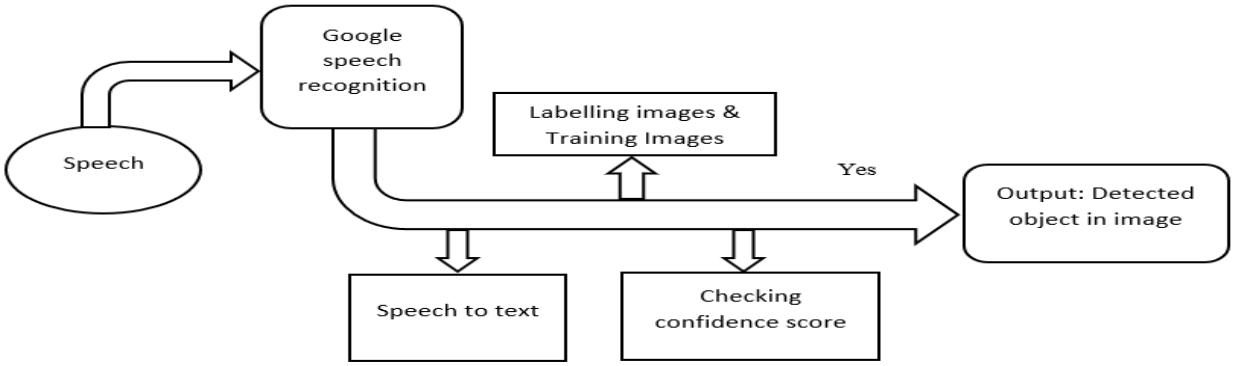

Fig 1: Block Diagram of the process.

Step 9: Training the image using the Darknet framework in the cloud using free GPU.
Step 10: Obtain the weights file after training the image using the darknet framework.
Step 11: Running the YOLO object detection model with all the training data.
Step 12: It checks the confidence score of each pixel of the image.
Step 13: If the object is near to the confidence score of the trained image then a bounding box is generated and the required object is detected.
Step 14: Else, Repeat the steps 1,2,3,4,5,6,7,8,9,10,11,12,13.
Step 15: End

C. Flow chart:

The workflow of this model begins when the user gives input to the system and the recognized speech is converted into text using Google Speech Recognition API. The obtained text is stored in a file named coco names. Now, the Labelling of the images is done using LabelImg software, and then the images are trained using the Darknet framework in the cloud using free GPU. The weights file is obtained after training the image using the darknet framework. Later, the user needs to run the YOLO object detection model with all the training data, the Configuration file. The model checks the confidence score of each pixel of the image. If the object is near to the score of the trained image then a boundary box is generated

and the required object is detected. Else you need to repeat the process. Fig.2.

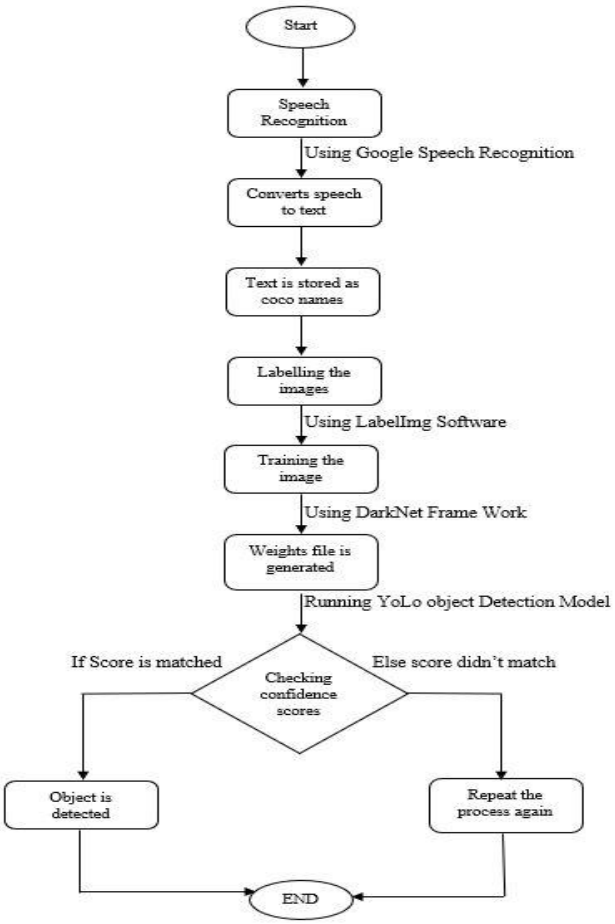

Fig 2: Overview of the process.

IV. RESULTS

Initially when you run the model. The process starts by receiving the speech input and then the speech is converted to text. This converted acts as the name of the object that is to be detected by the model. Here in fig.4, if the user gives the input as 'Clock' so the output should be bounded box named clock that is going to be displayed on the output screen.

Input: The user needs to specify the object name through the microphone.

```
Speak now:
voice recorded and converted into text!
```

Fig.3: Output provided by the model after Speech is converted to text.

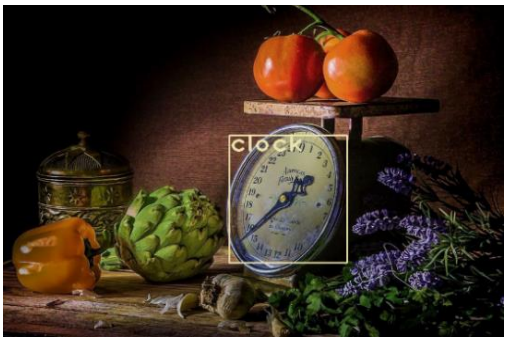

Fig.4: Output of Clock object If the given input is Clock.

A similar approach is followed for fig.5,6 'Cat' and 'Bottle' is given as speech input to identify those objects in the image. Here the speech to text is converted by utilizing Google Speech Recognition API. The Labelling of images is performed using LabelImg software, and then training the image is done with the Darknet framework in the cloud using free GPU. The weights file is obtained after training the image using the darknet framework. Later, you should run the YOLO object detection model with all the training data, the Configuration file. The model checks the confidence score of each pixel of the image. If the object is near to the score of the trained image then a boundary box is generated and the required object is detected.

Fig.5: Output of Cat object If the given input is Cat.

Fig.6: Output of Bottle object If the given input is Bottle.

During the training phase, the model is trained using the Darknet Framework with the labeled image, configuration file, and object name that is specified by the user. You will get different accuracy rates based on the number of iterations that you perform on the model. During every iteration, the model will compare the specified object with the existing or pre-trained YOLO weights and generates the accuracy percentage of the detected object. Table.1

Detected Object	Accuracy	
	Image Trained up to 100 Iterations	Image Trained up to 1000 Iterations
Clock	87%	98%
Cat	95%	100%
Bottle	89%	97%

Table 1: Accuracy metrics.

V. CONCLUSION

By implementing this integrated model, it has been concluded that there are many possible techniques to perform Speech Recognition and object detection. But this approach yields in providing better and accurate results as you are uploading all the images online on Google servers they can be processed for the training, this doesn't require a powerful computer or GPU because all the processing will occur online using the Google cloud GPU. This model will be helpful for the people who are illiterates and don't know how to operate a device, and this can also be utilized as a part of Virtual teaching.

REFERENCES

[1] "Machine-learning system tackles speech and object recognition" https://news.mit.edu/machine-learning-image-object-recognition-0918.

[2] Lakha, Bishal, Heuju, Bishal, Bhusal, Dipkamal, Shrestha, kanhaiya. (2016). "Voice Command Based Object Recognizing Robot Using Speech and Image Feature Extraction." 10.13140/RG.2.2.33893.81128.

[3] Swetha V Patil, Pradeep N, "Speech translation system for language barrier reduction," International Research Journal of Engineering and Technology (IRJET), Volume: 06, Issue: 08, August 2019.

[4] "Development and Application of Multilingual Speech Translation," Satoshi Nakamura', Spoken Language Communication Research Group Project, National Institute of Information and Communications Technology, Japan.

[5] Manvi Chawla, Sandeep Kumar, Aman Balyan, "Object Detection and Recognition in Images," International Journal of Engineering Development and Research, Volume: 05, Issue: 04, ISSN: 2321-9939, 2017.

[6] Krishnaveni, G., Lalitha Bhavani, B., Vijaya Lakshmi, N.V.S.K., "An enhanced approach for object detection using wavelet based neural network," International Conference on Computer Vision and Machine Learning 2018, ICCVML 2018, Volume: 1228, Issue: 1.

[7] Inthiyaz, S., Ahammad, S.H., Sai Krishna, A., Bhargavi, V., Govardhan, D., Rajesh, V., "YOLO (YOU ONLY LOOK ONCE) making object detection work in medical imaging on convolution detection system," International Journal of Pharmaceutical Research, Volume: 12, Issue: 2, April-June 2020, Pages: 312-326.

[8] Mandhala, V.N., Bhattacharyya, D., Vamsi, B., Thirupathi Rao, N., "Object detection using machine learning for visually impaired people," International Journal of Current Research and Review, Volume: 12, Issue: 20, October 2020, Pages: 157-167.

[9] Olabe, J. C.; Santos, A.; Martinez, R.; Munoz, E.; Martinez, M.; Quilis, A.; Bernstein, J., "Real time text-to speech conversion system for spanish," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84., vol.9, no., pp.85,87, Mar 1984.

[10] K. Kumar, S. Nandan, A. Mishra, K. Kumar and V. K. Mittal, "Voice-controlled object tracking smart robot," 2015 International Conference on Signal Processing, Computing and Control (ISPCC), Waknaghat, 2015, pp. 40-45, doi: 10.1109/ISPCC.2015.7374995.

[11] Bennilo Fernandes, J., Mannepalli, K.P., Saravanan, R.A., Kumar, K.T.P.S., "Fuzzy utilization in speech recognition and its different application," International Journal of Engineering and Advanced Technology,Volume: 8, Issue: 5, Special Issue: 3, July 2019, Pages: 261-266.

[12] Shahana Bano., Pavuluri Jithendra., Gorsa Lakshmi Niharika., Yalavarthi Sikhi., "Speech to Text Translation enabling Multilingualism," 2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India. Nov 6-8, 2020.

[13] Shahana Bano., Lakshmi Niharika Gorsa., Jithendra Pavuluri., Sikhi Yalavarthi., "Proposed Cognitive Model for Detection of Objects based on Speech Recognition," 2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India. Nov 6-8, 2020.

[14] Rajesh Kumar, T., Videla, L.S., Sivakumar, S., Gupta, A.G., Haritha, D., "Murmured speech recognition using hidden markov model," 7th International Conference on Smart Structures and Systems, ICSSS 2020, July 2020, Article number: 9202163.

[15] S. M. Chittajallu, N. Lakshmi Deepthi Mandalaneni, D. Parasa and Shahana Bano, "Classification of Binary Fracture Using CNN," 2019 Global Conference for Advancement in Technology (GCAT), BANGALURU, India, 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978468.

[16] P. Vishal, L. K. Snigdha, Shahana Bano, "An Efficient Face Recognition System using Local Binary Pattern", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5S4, February 2019.

[17] Shariff, M.N., Saisambasivarao, B., Vishvak, T., Rajesh Kumar, T., "Biometric user identity verification using speech recognition based on ANN/HMM," Journal of Advanced Research in Dynamical and Control Systems, Volume: 9, Issue: 12 Special issue, 2017, Pages: 1739-1748.

[18] Teju, V., Bhavana, D., "An efficient object detection using OFSA for thermal imaging," International Journal of Electrical Engineering Education, 2020.

[19] D. Choi, and M. Kim, "Trends on Object Detection Techniques Based on Deep Learning," Electronics and Telecommunications Trends, Vol.33, No.4, pp.23-32, Aug.2018.

[20] Mrinalini Ket al: Hindi-English Speech-to-Speech Translation System for Travel Expressions, 2015 International Conference On Computation Of Power, Energy, Information And Communication.

[21] Speech-to-Speech Translation: A Review, Mahak Dureja Department of CSE The NorthCap University, Gurgaon Sumanlata Gautam Department of CSE The NorthCap University, Gurgaon. International Journal of Computer Applications (0975 – 8887) Volume 129 – No.13, November2015.

[22] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," in Proc. Of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp.779-788, Jun. 2016.

[23] K. Kumar, S. Nandan, A. Mishra, K. Kumar and V. K. Mittal, "Voice-controlled object tracking smart robot," 2015 International Conference on Signal Processing, Computing and Control (ISPCC), Waknaghat, 2015, pp. 40-45, doi: 10.1109/ISPCC.2015.7374995.

[24] Das, Prerana & Acharjee, Kakali & Das, Pranab & Prasad, Vijay. (2015). Voice Recognition System: Speech-To-Text. Journal of Applied and Fundamental Sciences. 1. 2395-5562.

[25] Mohamed, A. R., Dahl, G. E., and Hinton, G., "Acoustic Modelling using Deep Belief Networks", submitted to IEEE TRANS. On audio, speech, and language processing, 2010.