International Journal of Emerging Trends in Engineering and Basic Sciences (IJEEBS)
ISSN (Online) 2349-6967
Volume 7, Special Issue 2 (March-April 2020), PP. 249-255

Development of an approach for Multiobject Detection & Recognition using Deep learning

Vivek Thakare, Animesh Tayal, Abhishek Jangitwar, Nikhil Borkar, Suraj Gupta, Aniket Kuralkar

Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur

vivek320thakare@gmail.com

Abstract: Object detection is a computer vision task which is widely used for face detection and recognition. There are many applications to detect an object and process it. If algorithms for object detection could be accurate and fast enough, the computers would be able to convey real-time scene information to users. We proposed an android application which is used for detecting and recognizing multiple objects provided by the live camera feed. YOLOv3 an object detection algorithm is used for this process. This application will also able to detect custom objects using the dataset trained for these custom objects. The input image or video obtained will be processed by the algorithm, used for detection and recognition of the objects in the camera frame. The algorithm divides the frame into SxS grids and performs the analysis. The objects in the frame are then recognized on the basis of class probability and confidence. The bounding box around the detected object is generated along with the text label specifying the name of the recognized object. The text label of the recognized object is extracted and converted to audio format using a text to speech algorithm. The object recognized is conveyed to the user via speech output. This application can also be used as virtual eye for the visually impaired person to help them to recognize objects.

1. INTRODUCTION

Object detection is a challenging and exciting task in Computer Vision. This task is comparatively difficult to perform for the machines as compared to Humans who perform object detection very effortlessly and instantaneously. Detection can be difficult since there are all kinds of variations in orientation, lighting, background and occlusion that can result in completely different images of the very same object. Now with the advance of deep learning and neural network, we can finally tackle such problems with less effort and time. If algorithms for image processing could be accurate and fast enough, the computers would be able to drive cars without specialized sensors, and assistive devices would be able to convey real-time scene information to users. Likewise, if these algorithms could complete Deep Learning (DL) tasks with high efficient and excellent performance like human beings do, it would be real Artificial Intelligent (AI).

The core tasks of image processing are a serial of recognition: classification, localization and object detection; and the key challenges are: accuracy, speed, cost and complexity. To achieve an ideal effect, algorithms of recognition have gone a long way before the significant boom of Convolutional Neural Network (CNN) brought by AlexNet [1] in 2012. Until 2012, AlexNet via CNN reduced the error rate from 26% to 15.3%, which dramatically fueled the development of DL, rekindled CNNs and GPU.

1.1. Objectives:

- To detect the multiple objects in a single frame.
- To recognize the detected object.
- To convert the text of recognized object into speech.

2. LITERATURE SURVEY

"YouOnlyLookOnce: Unified, Real-TimeObjectDetection" - This paper state that First, YOLO is extremely fast. Since frame detection as a regression problem we don't need a complex pipeline. We simply run our neural network on a new image at test time to predict detections. Our base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps [2].

"Understanding of Object Detection Based on CNN Family and YOLO (2018)" YOLOv2 provides state-of-the-art the best tradeoff between real-time speed and excellent accuracy for object detection than other detection systems across a variety of detection datasets and Furthermore, YOLO's better generalizing representation of object than other models making it ideal for applications that rely on fast, robust object detection [3].

"A review and an approach for object detection in images"This paper provides the details of the existing approaches based on the key concept which is used as the base for development of the approach. Apart from this, the summary of the available source codes and the datasets used for the evaluation of the object detection approach is presented. This paper also provides the idea to solve the multi class object detection problem based on the Steiner tree. This paper is useful for the study purpose as well as for the new researchers who want to explore the OD research area [4].

"Moving Object Tracking in Video" object detection in video image obtained from single camera with static background that means fixing camera is achieved by background subtraction approach. In this thesis, we tried different videos with fixed camera with a single object and multiple objects to see it is able to detect objects. Motion based systems for detecting and tracking given moving object of interest can be created. Using SIFT feature extraction first feature of the object and the frame has detected to match the interested object. Since for feature extraction, SIFT algorithm has been used so tracker is invariant to representation of interested object. In the future, we can extend the work to detect the moving object with non-static.

Background, having multiple cameras which can be used in real time surveillance applications [5].

"A review and an approach for object detection in images" a review on the topic of object detection (OD) has been carried out by Prasad (2012), Madaan and Sharma (2012) and Karasulu (2010). Prasad (2012) has discussed the problem of object detection in real images and addressed the various aspects like the feature types, learning model, object templates, matching schemes and boosting methods, where as Madaan and Sharma (2012) have considered the same problem of OD in remote sensing images and explored the concept of pre-segmentation for object detection . Karasulu (2010) has reviewed and evaluated the methods for moving object detection in videos. Though, the different review papers are available for OD, but, this paper is different from the existing papers. This paper provides the details of the existing approaches based on the key concept which is used as the base for development of the approach. Apart from this, the summary of the available source codes and the datasets used for the evaluation of the OD approach is presented. This paper also provides the idea to solve the

multi class OD problem based on the Steiner tree. This paper is useful for the study purpose as well as for the new researchers who want to explore the OD research area [6].

"Multi-Object Detection in Traffic Scenes Based on Improved SSD" YOLO algorithm is used for the purpose of detecting objects using a single neural network. This algorithm is generalized, it outperforms different strategies once generalizing from natural pictures to different domains. The algorithm is simple to build and can be trained directly on a complete image. Region proposal strategies limit the classifier to a particular region. YOLO accesses to the entire image in predicting boundaries. And also it predicts fewer false positives in background areas.Comparing to other classifier algorithms this algorithm is much more efficient and fastest algorithm to use in real time. [7]

"Object Detection CS725: Project Report of Department of Computer Science and Engineering Indian Institute of Technology Bombay India". An accurate and efficient object detection system has been developed which achieves comparable metrics with the existing state-of-the-art system. This project uses recent techniques in the field of computer vision and deep learning. Custom dataset was created using label image and the evaluation was consistent. This can be used in real-time applications which require object detection for pre-processing in their pipeline. An important scope would be to train the system on a video sequence for usage in tracking applications. Addition of a temporally consistent network would enable smooth detection and more optimal than per-frame detection. [8]

In this work, we pose a new problem of Interactive Question Answering for several question types ininteractive environments. We propose the Hierarchical Interactive Memory Network (HIMN) for thistask, consisting of a factorized set of controllers, allowing the system to operate at multiple levels oftemporal abstraction. We also introduce the Egocentric Spatial GRU for updating spatial memory maps.The effectiveness of our proposed model is demonstrated on a new benchmark dataset built upon ahigh-quality simulation environment for this task. This dataset still presents several challenges to ourmodel and baselines and warrants future research. [9]

We have performed an experimental comparison of some of the main aspects that influence the speed and accuracy of modern object detectors. We hope this will help practitioners choose an appropriate method when deploying object detection in the real world. We have also identified some new techniques for improving speed without sacrificing much accuracy, such as using many fewer proposals than is usual for Faster R-CNN [10].

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers---8x deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also presentanalysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation [11]**.**

VOC 2012 test	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
MR_CNN_MORE_DATA [11]	**73.9**	**85.5**	**82.9**	**76.6**	**57.8**	**62.7**	**79.4**	77.2	86.6	**55.0**	**79.1**	**62.2**	87.0	**83.4**	**84.7**	78.9	45.3	73.4	65.8	80.3	74.0
HyperNet_VGG	71.4	84.2	78.5	73.6	55.6	53.7	78.7	**79.8**	87.7	49.6	74.9	52.1	86.0	81.7	83.3	**81.8**	**48.6**	**73.5**	59.4	79.9	65.7
HyperNet_SP	71.3	84.1	78.3	73.3	55.5	53.6	78.6	79.6	87.5	49.5	74.9	52.1	85.6	81.6	83.2	81.6	48.4	73.2	59.3	79.7	65.6
Fast R-CNN + YOLO	70.7	83.4	78.5	73.5	55.8	43.4	79.1	73.1	**89.4**	49.4	75.5	57.0	**87.5**	80.9	81.0	74.7	41.8	71.5	68.5	**82.1**	67.2
MR_CNN_S_CNN [11]	70.7	85.0	79.6	71.5	55.3	57.7	76.0	73.9	84.6	50.5	74.3	61.7	85.5	79.9	81.7	76.4	41.0	69.0	61.2	77.7	72.1
Faster R-CNN [27]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
DEEP_ENS_COCO	70.1	84.0	79.4	71.6	51.9	51.1	74.1	72.1	88.6	48.3	73.4	57.8	86.1	80.0	80.7	70.4	46.6	69.6	**68.8**	75.9	71.4
NoC [28]	68.8	82.8	79.0	71.6	52.3	53.7	74.1	69.0	84.9	46.9	74.3	53.1	85.0	81.3	79.5	72.2	38.9	72.4	59.5	76.7	68.1
Fast R-CNN [14]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	**87.5**	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
UMICH_FGS_STRUCT	66.4	82.9	76.1	64.1	44.6	49.4	70.3	71.2	84.6	42.7	68.6	55.8	82.7	77.1	79.9	68.7	41.4	69.0	60.0	72.0	66.2
NUS_NIN_C2000 [7]	63.8	80.2	73.8	61.9	43.7	43.0	70.3	67.6	80.7	41.9	69.7	51.7	78.2	75.2	76.9	65.1	38.6	68.3	58.0	68.7	63.3
BabyLearning [7]	63.2	78.0	74.2	61.3	45.7	42.7	68.2	66.8	80.2	40.6	70.0	49.8	79.0	74.5	77.9	64.0	35.3	67.9	55.7	68.7	62.6
NUS_NIN	62.4	77.9	73.1	62.6	39.5	43.3	69.1	66.4	78.9	39.1	68.1	50.0	77.2	71.3	76.1	64.7	38.4	66.9	56.2	66.9	62.7
R-CNN VGG BB [13]	62.4	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7	82.0	74.8	76.0	65.2	35.6	65.4	54.2	67.4	60.3
R-CNN VGG [13]	59.2	76.8	70.9	56.6	37.5	36.9	62.9	63.6	81.1	35.7	64.3	43.9	80.4	71.6	74.0	60.0	30.8	63.4	52.0	63.5	58.7
YOLO	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
Feature Edit [32]	56.3	74.6	69.1	54.4	39.1	33.1	65.2	62.7	69.7	30.8	56.0	44.6	70.0	64.4	71.1	60.2	33.3	61.3	46.4	61.7	57.8
R-CNN BB [13]	53.3	71.8	65.8	52.0	34.1	32.6	59.6	60.0	69.8	27.6	52.0	41.7	69.6	61.3	68.3	57.8	29.6	57.8	40.9	59.3	54.1
SDS [16]	50.7	69.7	58.4	48.5	28.3	28.8	61.3	57.5	70.8	24.1	50.7	35.9	64.9	59.1	65.8	57.1	26.0	58.8	38.6	58.9	50.7
R-CNN [13]	49.6	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5	66.2	57.3	65.4	53.2	26.2	54.5	38.1	50.6	51.6

Table: YOLO compared with the full public leader board. Mean average precision and per-class average precision are shown for a variety of detection methods. YOLO is the only real-time detector. Fast R-CNN + YOLO is the fourth highest scoring method, with a 2.3% boost over Fast R-CNN.

3. PROPOSED WORK

3.1 Flow of the System:

The user will start the camera. After starting the camera the input image or video obtained will be processed by the image detection algorithm, used for recognition and detection of the objects in the camera frame. The algorithm divides the recognized frame into SxS grids and performs analysis. The objects in the frame are then recognized on the basis of class probability and confidence. The bounding box around the recognized object is generated along with the text label specifying the name of the recognized object. The text label of the recognized object is extracted and converted to audio format using a text to speech algorithm. The object recognized is conveyed to the user via speech output.

3.2 Functional Modules:

The whole system is divided into four modules. They are Dataset Gathering and Cleaning, Object Detection and Object Recognition, Text to speech conversion and GUI development.

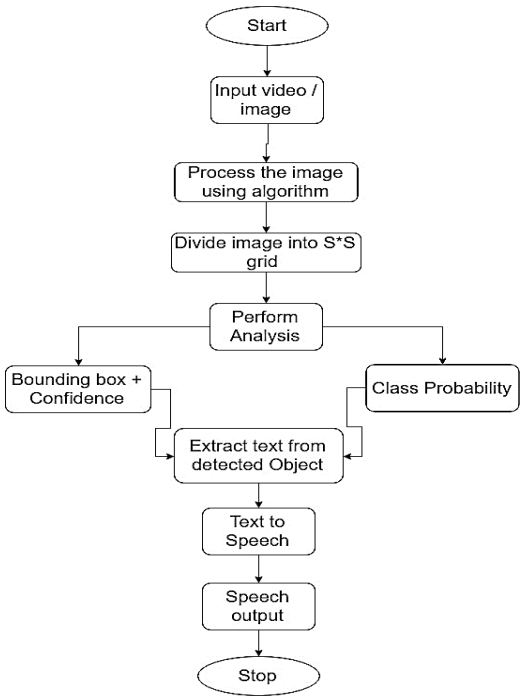

Fig.1: Flowchart for image recognition and detection

3.2.1 Dataset Gathering and Cleaning:

This involves training a dataset for a custom object. The dataset for custom object will be trained and the weight file or CFG file generated will be used by the YOLOV3 detection algorithm to recognize the custom object. We trained the data set for custom object by using supervisely.

3.2.2 Object Detection and Recognition:

This module focuses on detecting objects with the help of camera. The multiple objects detected in the camera frame will be highlighted by a rectangular box along with the text to indicate the detected object, this rectangular box is nothing but a Bounding box. The objects detected in the camera frame will be recognized on the basis of the class probability. The recognition part will be executed by the YOLOV3 library.

3.2.3 Text to Speech Conversion:

The text of the recognized objects willbe converted into speech output and it will be the final output which will convey the names of objects detected in the frame.

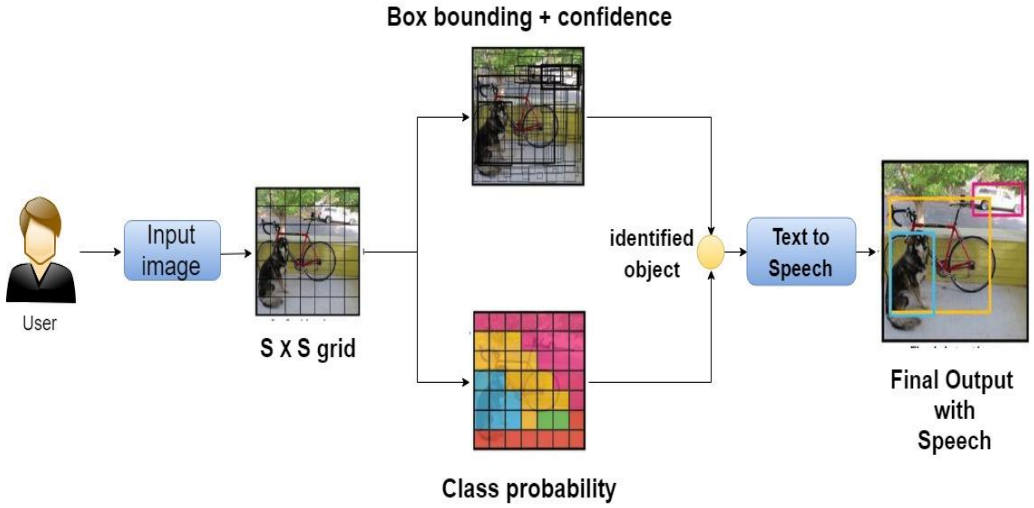

Fig. 2: Graphical flow of functional modules.

3.2.4 GUI development and App Integration:

This module is a mobile application. In this module we have integrated Object Detection and Recognition in mobile application.

4. CONCLUSION

This survey helps in developing an approach for multi-object detection and recognition. It has helped to explore the various approach that has been previously developed for single and multiple object detection and recognition. With this survey and study, we have proposed an efficient approach for detection and recognition using YOLO v3 from which we can increase the mAP (mean Average Precession). When we implement the YOLO v3 in combination with Fast R-CNN the mAP obtained is 70.7. Our approach also has a feature of converting recognized image into the speech that will be our final output. The proposed approach also has implemented in the mobile application. This system in future will help a person with having visual impairment as a helping hand by giving voice output of objects in front of them.

REFERENCES

Journal Papers:

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks.In Advances in neural information processing systems (pp.1097-1105).

[2] You Only Look Once: Unified, Real-Time Object Detection Joseph Redmon, Santosh Divvalay, Ross Girshick, Ali Farhadi, University of Washington, Allen Institute for AIy, Facebook AI Research.

[3] Understanding of Object Detection Based on CNN Family and YOLO (2018) Juan Du1, New Research and Development Center of Hisense, Qingdao, China. 266071.

[4] "A review and an approach for object detection in images" Int. J. Computational Vision and Robotics, Vol. 7, Nos. 1/2, 2017 KartikUmesh Sharma* and Nileshsingh V. Thakur

[6] "A review and an approach for object detection in images" KartikUmesh Sharma* and Nileshsingh V. Thakur Int. J. Computational Vision and Robotics, Vol. 7,Nos.1/2,2017.

[10] Speed/accuracy trade-offs for modern convolutional object detectors J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama and K. Murphy. In CVPR, 2017.

[11] Deep Residual Learning for Image Recognition.Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Projects:

[5] Yiwei Wang and John F. Doherty, Robert E. Van Dyck "Moving Object Tracking in Video" 2018.

[7] Multi-Object Detection in Traffic Scenes Based on Improved SSD Xinqing Wang, Xia Hua *, Feng Xiao, Yuyang Li, Xiaodong Hu and Pengyu Sun College of Field Engineering, PLA Army Engineering University, Nanjing 210007, China; wwwxxxqqq@126.com (X.W.); xiaofeng199377@163.com (F.X.); lyychqs@163.com (Y.L.); hxd3281008@163.com (X.H.); zzc91292@163.com (P.S.) * Correspondence: huaxia120888@163.com; Tel.: +86-176-2603-9818 Received: 9 September 2018; Accepted: 2 November 2018; Published: 6 November 2018

[8] Object Detection CS725: Project Report of Department of Computer Science and Engineering Indian Institute of Technology Bombay India.

[9] Iqa: Visual question answer in interactive environments. D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering In interactive environments.ArXiv preprint arXiv: 171203316, 2017