International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Feedback for Audio Object Detection using Deep Learning

[1]M Phani Vineel varma, [2]Dr. S. Sathyapriya
[*1]BTech Student, Department of CSE, HITS Chennai, India
[2]Associate Professor, Department of CSE, HITS Chennai, India

Article Info
Publication Issue :
Volume 8, Issue 6
November-December-2022
Page Number : 243-249
Article History
Accepted: 10 Nov 2022
Published: 23 Nov 2022

ABSTRACT

Object recognition is one of the challenging application of computer vision, which has been widely applied in many areas for e.g. autonomous cars, Robotics, Security tracking, Guiding Visually Impaired Peoples etc. With the rapid development of deep learning many algorithms were improving the relationship between video analysis and image understanding. All these algorithms work differently with their network architecture but with the same aim of detecting multiple objects within complex image. Absence of vision impairment restraint the movement of the person in an unfamiliar place and hence it is very essential to take help from our technologies and trained them to guide blind peoples whenever they need.

Keywords : Tensor flow, Yolo_v3, Web Speech API, Deep Learning.

I. INTRODUCTION

Humans almost by birth are trained by their parents to categorize between various objects as children self is one object. Human Visual System is very accurate and precise that can handle multi-tasks even with less conscious mind. When there is large data then we need more accurate system to correctly recognize and localize multiple objects simultaneously. Here machines comes into existence, we can train our computers with the help of better algorithms to detect multiple objects within the image with high accuracy and preciseness. Object Detection is the most challenging application of computer vision as it require complete understanding of images. In other words object tracker tries to find the presence of object within multiple frames and assigns labels to each object. There might be many problems faced by the tracker in terms of complex image, Loss of information and transformation of 3D world into 2 D image. To achieve good accuracy in object detection we should not only focus on classifying objects but also on locating the positions of different objects that may vary image to image. It is very important to develop the most effective real time object tracking algorithm which is a challenging task. Deep learning since 2012 is working in these kinds of problems and has revolutionized the domain of computer vision. This paper aims to test the performance of both the algorithms in different situations in real time using webcam and is made primarily for the visually impaired peoples. Blind peoples have to rely on

someone who can guide them or on their physical touch which is sometimes very risky also. Daily navigation of blind peoples in unfamiliar environments could be the frighten task without the help of some intelligent systems. They key concern

behind this contribution is to investigate the possibility of expanding the counts of objects at one go to expand the support given to the visually impaired peoples. Some common limitations of the previous techniques is less accuracy, complexity in scene, lightening etc. To overcome all those challenges two algorithms are analyzed on all possible grounds and from every perspective to achieve good accuracy. Recognizing objects and localizing them in images is one of the most fundamental and challenging problems in computer vision. There has been significant progress on this problem over the last decade due largely to the use of low-level image features, such as SIFT and HOG, in sophisticated machine learning frameworks. But if we look at performance on the canonical visual recognition task, PASCAL VOC object detection, it is generally acknowledged that progress slowed from 2010 onward, with small gains obtained by building ensemble systems and employing minor variants of successful methods.

CNNs saw heavy use in the 1990s, but then fell out of fashion with the rise of support vector machines. In 2012, Krizhevsky et al. rekindled interest in CNNs by showing a substantial improvement in image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Their success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on CNNs from the 1990s (e.g., max(x, 0) "ReLU" non-linearities, "dropout" regularization, and a fast GPU implementation). The significance of the ImageNet result was vigorously debated during the ILSVRC 2012 workshop. The central issue can be distilled to the following: To what extent do the CNN classification results on ImageNet generalize to object detection results on the PASCAL VOC Challenge? We answered this question in a conference version of this paper by showing that a CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to systems based on simpler HOG-like features. To achieve this result, we bridged the gap between image classification and object detection by developing solutions to two problems: (1) How can we localize objects with a deep network and (2) How can we train a high-capacity model with only a small quantity of annotated detection data? Unlike image classification, detection requires localizing (likely many) objects within an image. One approach is to frame detection as a regression problem. This formulation can work well for localizing a single object, but detecting multiple objects requires complex workarounds or an ad hoc assumption about the number of objects per image. An alternative is to build a sliding-window detector. CNNs have been used in this way for at least two decades, typically on constrained object categories, such as faces, hands, and pedestrians. This approach is attractive in terms of computational efficiency, however its straightforward application requires all objects to share a common aspect ratio. The aspect ratio problem can be addressed with mixture models where each component specializes in a narrow band of aspect ratios, or with bounding-box regression. Instead, we solve the localization problem by operating within the "recognition using regions" paradigm, which has been successful for both object detection and semantic segmentation. At test time, our method generates around 2000 category-independent region proposals for the input image, extracts a fixed-length feature vector from each proposal using a CNN, and then classifies each region with category-specific linear SVMs. We use a simple warping technique (anisotropic image scaling) to compute a fixedsize CNN input from each region proposal, regardless of the region's shape. Fig. 1 shows an overview of a Region-based Convolutional Network (R-CNN) and highlights some of our results. A second challenge faced in detection is that labeled data are scarce and the amount currently available is insufficient for training large CNNs from random initializations. The conventional solution to this problem is to use unsupervised pre-training, followed

by supervised fine-tuning. The second principle contribution of this paper is to show that supervised pre-training on a large auxiliary dataset (ILSVRC), followed by domain-specific fine-tuning on a small dataset (PASCAL), is an effective paradigm for learning high-capacity CNNs when data are scarce. In our experiments, fine-tuning for detection can improve mAP by as much as 8 percentage points. After fine-tuning, our system achieves a mAP of 63% on VOC 2010 compared to 33% for the highly-tuned, HOG-based deformable part model (DPM). Our original motivation for using regions was born out of a pragmatic research methodology: move from image classification to object detection as simply as possible. Since then, this design choice has proved valuable because RCNNs are straightforward to implement and train (compared to sliding-window CNNs) and it provides a unified solution to object detection and segmentation.

II. RELATED WORKS

Real time implementation of object tracking through webcam: Real time object detection and tracking is an important task in various computer vision applications. For robust object tracking the factors like object shape variation, partial and full occlusion, scene illumination variation will create significant problems. We introduce object detection and tracking approach that combines Prewitt edge detection and kalman filter. The target object's representation and the location prediction are the two major aspects for object tracking this can be achieved by using these algorithms. Here real time object tracking is developed through webcam. Experiments show that our tracking algorithm can track moving object efficiently under object deformation, occlusion and can track multiple objects.

Object Detection with Deep Learning: A Review: Due to object detection's close relationship with video analysis and image understanding, it has attracted much research attention in recent years. Traditional object detection methods are built on handcrafted features and shallow trainable architectures. Their performance easily stagnates by constructing complex ensembles which combine multiple low-level image features with high-level context from object detectors and scene classifiers. With the rapid development in deep learning, more powerful tools, which are able to learn semantic, high-level, deeper features, are introduced to address the problems existing in traditional architectures. These models behave differently in network architecture, training strategy and optimization function, etc. In this paper, we provide a review on deep learning based object detection frameworks. Our review begins with a brief introduction on the history of deep learning and its representative tool, namely Convolutional Neural Network (CNN). Then we focus on typical generic object detection architectures along with some modifications and useful tricks to improve detection performance further. As distinct specific detection tasks exhibit different characteristics, we also briefly survey several specific tasks, including salient object detection, face detection and pedestrian detection. Experimental analyses are also provided to compare various methods and draw some meaningful conclusions. Finally, several promising directions and tasks are provided to serve as guidelines for future work in both object detection and relevant neural network based learning systems.

Histograms of oriented gradients for human detection: We study the question of feature sets for robust visual object recognition; adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast

normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

Region-Based Convolutional Networks for Accurate Object Detection and Segmentation: Object detection performance, as measured on the canonical PASCAL VOC Challenge datasets, plateaued in the final years of the competition. The best-performing methods were complex ensemble systems that typically combined multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 50 percent relative to the previous best result on VOC 2012-achieving a mAP of 62.4 percent. Our approach combines two ideas: (1) one can apply high-capacity convolutional networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since we combine region proposals with CNNs, we call the resulting model an R-CNN or Region-based Convolutional Network.

Region-Based Convolutional Networks for Accurate Object Detection and Segmentation: Object detection performance, as measured on the canonical PASCAL VOC Challenge datasets, plateaued in the final years of the competition. The best-performing methods were complex ensemble systems that typically combined multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 50 percent relative to the previous best result on VOC 2012—achieving a mAP of 62.4 percent. Our approach combines two ideas: (1) one can apply high-capacity convolutional networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since we combine region proposals with CNNs, we call the resulting model an <italic>R-CNN</italic> or <italic>Region-based Convolutional Network</italic>. Source code for the complete system is available at http://www.cs.berkeley.edu/~rbg/rcnn.

III. Methodology

Proposed system:

We propose a system that will detect every possible day to day multiple objects on the other hand prompt a voice to alert person about the near as well as farthest objects around them. To get audio we will use web speech api to produce speech.

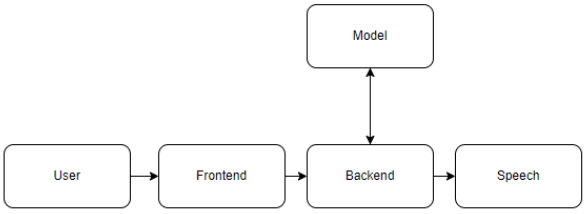


Figure 1: Block diagram

IV. Implementation

The project was carried out using the algorithms listed below.

CNN (Convolutional Neural Network):

In deep learning, a convolutional neural network (CNN) is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared-weight architecture of the convolution kernels that shift over input features and provide translation equivariant responses.

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks makes them prone to overfitting data. Typical ways of regularization, or preventing overfitting, include: penalizing parameters during training (such as weight decay) or trimming connectivity (skipped connections, dropout, etc.) CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.

The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolutional networks are a specialized type of neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

APPLICATIONS:

- Image recognition
- Video analysis
- Natural language processing
- Anomaly Detection
- Drug discovery
- Health risk assessment and biomarkers of aging discovery

YOLO:

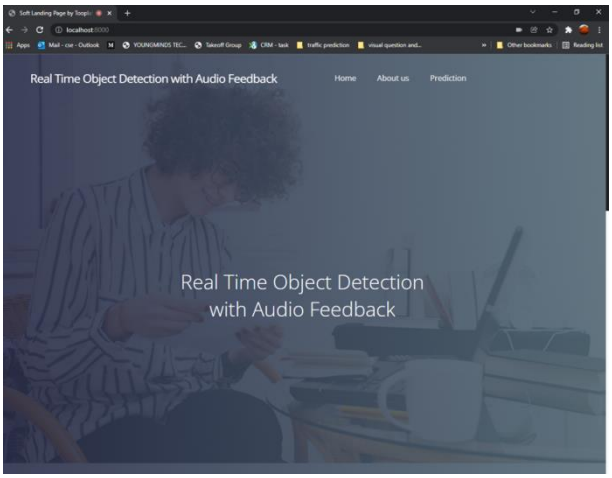
Yolo is a part of object detection, Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Every object class has its own special features that helps in classifying the class – for example all circles are round. Object class detection uses these special features. For example, when looking for circles, objects that are at a particular distance from a point (i.e., the center) are sought. Similarly, when looking for squares, objects that are perpendicular at corners and have equal side lengths are needed. A similar approach is used for face identification where eyes, nose, and lips can be found and features like skin color and distance between eyes can be found.

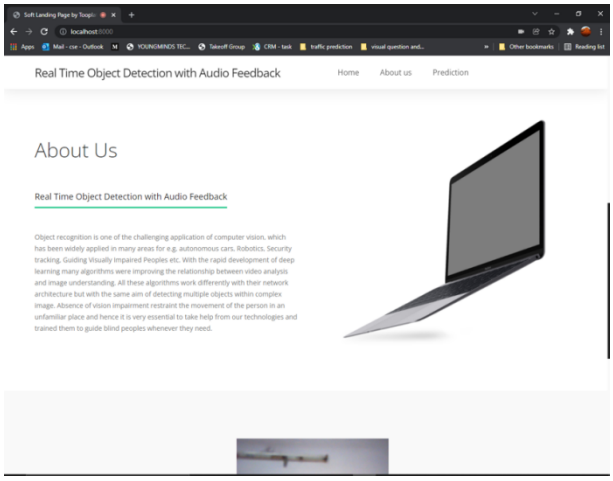
V. Results and Discussion

The following screenshots are depicted the flow and working process of project.

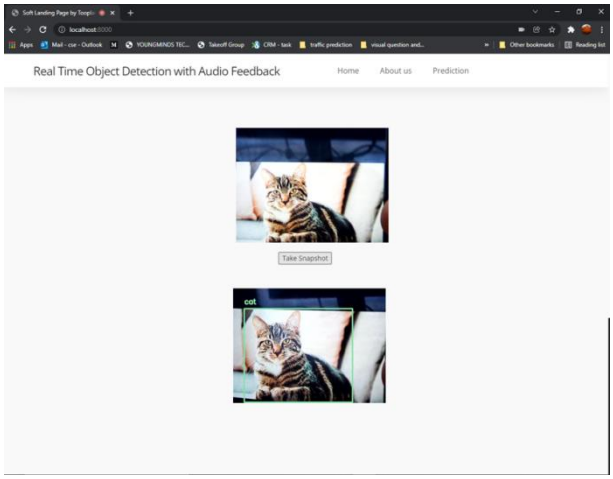
Home page: This is the home page of the real time object detection with audio feedback.



About us: This page will displays the brief introduction about the project.



Prediction: Here the prediction will be done by providing audio.



VI. CONCLUSION

We have developed a user friendly application called Object detection with audio feedback using deep learning techniques such as CNN (Convolutional Neural Network), YOLO. This will help out the blind people to identify the objects at unfamiliar places. These algorithms were improving the relationship between video analysis and image understanding.

VII. REFERENCES

[1]. S. Cherian, & C. Singh, "Real Time Implementation of Object Tracking Through webcam," Internation Journal of Research in Engineering and Technology, 128-132, (2014).

[2]. Z. Zhao, Q. Zheng, P.Xu, S. T, & X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, 30(11), 3212-3232, (2019).

[3]. N. Dalal, & B. Triggs, "Histograms of oriented gradients for human detection," In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE, (2005, June).

[4]. R. Girshick., J. Donahue, T. Darrell, & J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, 38(1), 142-158, (2015).

[5]. X. Wang, A. Shrivastava, & A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2606- 2615), (2017).

[6]. S. Ren, K. H, R. Girshick, & J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," In Advances in neural information processing systems (pp. 91-99), (2015).

[7]. J. Redmon, S. Divvala, R. Girshick, & A. Farhadi, "You only look once: Unified, real-time object detection," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788), (2016).

[8]. J. Redmon, & A. Farhadi, "YOLO9000: better, faster, stronger," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271) (2017).

[9]. J. Redmon & A. Farhadi, "Yolov3: An incremental improvement," ArXiv preprint arXiv: 1804.02767, (2018).

[10]. R. Bharti, K. Bhadane, P. Bhadane, & A. Gadhe, "Object Detection and Recognition for Blind Assistance," International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06, (2019).

[11]. T. Lin, Y. Maire, M. Belongie, S. Hays, J. Perona, P. Ramanan, D., & C.L. Zitnick, "Microsoft coco: Common objects in context," In European conference on computer vision (pp. 740-755). Springer, Cham, (2014, September).

[12]. Lowe D., "Distinctive image features from scale-invariant keypoints, " Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[13]. Dalal N. and Triggs B., "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 886–893.

[14]. Everingham M., van Gool L., Williams C. K. I., Winn J. , and Zisserman A., "The PASCAL visual object classes (VOC) challenge, " Int. J. Comput. Vis., vol. 80, no. 2, pp. 303–338, 2010.

[15]. Fukushima K., "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybern. , vol. 36, no. 4, pp. 193–202, 1980.

[16]. Rumelhart D. E., Hinton G. E., and Williams R. J., "Learning internal representations by error propagation," Parallel Distrib. Process. , vol. 1, pp. 318–362, 1986.

[17]. Krizhevsky A., Sutskever I., and Hinton G., "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1106–1114.

[18]. Sermanet P., Kavukcuoglu K., Chintala S., and LeCun Y., "Pedestrian detection with unsupervised multi-stage feature learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 3626–3633.

[19]. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," in ICLR, 2014.

[20]. J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," IJCV, 2013.

[21]. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," arXiv e-prints, vol. arXiv:1409.0575v1 [cs.CV], 2014.

Cite this article as :

M Phani Vineel varma, Dr. S. Sathyapriya, "A Feedback for Audio Object Detection using Deep Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 243-249, November-December 2022.
Journal URL : https://ijsrcseit.com/CSEIT228633