International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

A Review on Object Recognition for Blind People Based on Deep Learning

Rebeiro Sharlene Sara Carlton[1], Huda Noordean[2]

[1]M. Tech Student, Department of Computer Science and Engineering, College of Engineering and Management, Punnapra, Kerala, India

[2]Assistant Professor, Department of Computer Science and Engineering, College of Engineering and Management, Punnapra, Kerala, India

Abstract: ***Object recognition is an emerging area in the field of computer vision and it is rapidly maturing due to deep learning. Object recognition is a technique for detection of the object by creating a bounding box on the object and labeling it. The principle is to develop a model which is able to detect and recognize the object with maximum accuracy and improved performance. It finds its application in object recognition for the visually challenged by generating an audio output for the recognized object.***

Keywords: Object Recognition, Neural Network, Deep Learning

1. Introduction

The recent advancement in the architecture of neural networks has lead to the rapid development of deep learning. These networks are used in the various applications of computer vision such as object recognition and detection, automatic image captioning, face recognition, object tracking, semantic segmentation[1]. Deep Learning is the subfield of machine learning based on the deep neural network. Deep learning provides automatic feature extraction unlike machine learning where the features extracted are hand engineered. Thus coming up with appropriate features to extract is difficult, time consuming and requires expert knowledge. Object detection based on deep learning has a progressive evolution beginning from the R-CNN[2] to YOLO[3]. Object detection is the process to discover the presence of an object in an image which is notified by creating a bounding box around the object which is done using background subtraction technique. Object recognition is the process of identifying what the object is and it is notified by labeling on the bounding box. The blind people are benefitted by the developments in object recognition since the object recognized and labeled by the network can be output as audio which makes it easier to identify and locate the objects independently without any assistance from others.

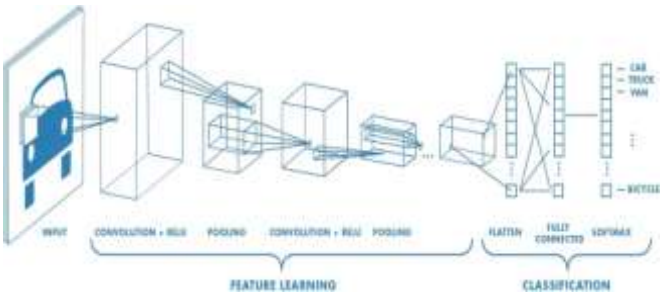

Figure 1: Object Detection based on deep learning

1.1 Deep Neural Network

Neural networks are a set of algorithms which is modeled after the human brain designed to recognize patterns. Deep neural network are composed of several layers and these layers are made of nodes. A neural network with more than three hidden layers is called a Deep neural network. Each layer of nodes trains on distinct set of features based on the previous layer of output. The further the layers pass, they aggregate and recombine features from previous layers deep neural networks helps to cluster and classify. Deep Learning maps input to output and finds correlation between any input x and output y. Classification is dependent on the labeled dataset and transfer of knowledge of the dataset by data scientists for the neural network to learn the correlation between labels and data is required. This is called as supervised learning. Clustering or grouping is the detection of similarities. Deep Learning does not require the labeled dataset to detect similarities and this type of learning is called as unsupervised learning. Deep Learning is able to process large quantities of unlabeled data which is a major advantage. Deep Learning network ends in an output layer which is generally a logistic or softmax, classifier which assigns the probability to a particular outcome or label. Given a raw data deep learning network may decide that the input data is 95% to represent the object.

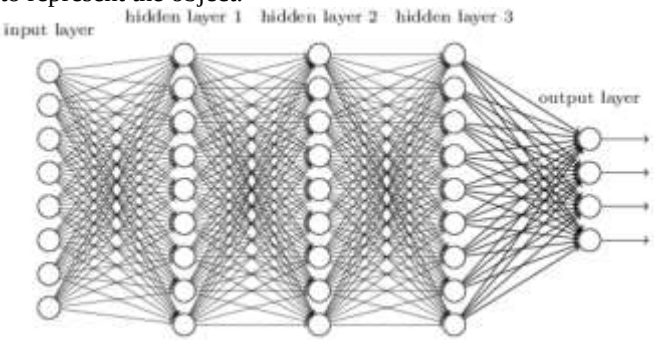

Figure 2: Deep neural network

2. Literature Survey

There has been advancement in the methods to detect objects in an image using hand engineered features to the recent development in deep learning. Since object detection is modeled as a classification problem the accuracy of a model is defined by how accurately it is able to classify the object. Initial solution of this problem was by using the classifiers

Licensed Under Creative Commons Attribution CC BY

such as the SVM, HOG based classifiers, Haar feature based cascade classifiers. After the development in deep learning the solution was to replace these classifiers with convolutional network based classifiers. The literature survey presents a description on the evolution of the object detection problem from the basic classifiers to the neural network based classifiers for object detection.

2.1 Object Recognition Techniques Using Hand Engineered Features.

There are five basic approaches of object detection namely:

A) Based on Template Matching
This method is used in target detection applications. When an image is given as input to a system, it is matched with the stored template images to identify the object in the image[4].

B) Based on Color
This method uses color attributes as an additional feature along with the shape features in order to detect objects. It provides better results where object detection using shape features alone fails. Object is detected by representing and matching the images based on color histogram [5].

C) Based on Scanning Method
This method involves scanning of the images to detect the object in an image. But the passive scanning method does not involve extraction of samples during scanning[6]. There are two passive scanning methods:

1) The sliding window approach: It checks if an object is present or not at all locations of the evenly spaced grid. A local sample is extracted at each grid and then classification is done based on whether it is an object or background of an image[7].
2) Part based approach: It determines the interest point in an image. It calculates the interest value for local samples at all points in an equally spaced grid and then at the interest points a new local sample is evaluated to check whether it belongs to the object or background[8].

Active scanning allows local samples to be used as a guide for scanning process. Here the image sample is extracted and is mapped to a shifting vector which gives an indication of the next scanning position. Successive samples are chosen, by skipping regions which is unlikely to contain the object. This saves the computational effort and gives a good quality detection.

D) Based on Shape
Shape features are used to detect objects in real world images they are often used as replacement or complement to local features. In [10], a new algorithm to find the correspondence between model and object is performed by providing an input image with an unknown object(shape) and compare it to a model thereby solving correspondence problem. In[11] the object is detected by extracting and clustering of edges using Gradient vector Girding method. In[12] a method for object detection is done based on the global shape based on elastic matching of contours.

D) Based on Local And Global Features
The most common technique of object detection is sliding of the window across the image and classifying the existence of an object or background in each local window . This approach successfully detects rigid objects. In[13] the object recognition and segmentation is performed using SIFT and graph cuts. Here the existence of an object is determined by SIFT key points and the object region is cut out using graph cut. In[14] the authors present object recognition with full boundary detection using ASIFT and region merging algorithm. In[15] each local window has its HOG feature calculated and it is fed to SVM to create classifiers. This model is computationally inexpensive and suitable for real world problems.

2.2 Object Detection Technique Based On Deep Learning

A) Region-based Convolutional Networks
The replacement of Hog based classifiers with CNN based classifier had one major problem. CNN were too slow and computationally expensive. Hence, there was difficulty in running patches generated by the sliding window detector. Thus, R-CNN[16] was introduced as a solution which does the object recognition by selective search algorithm thereby reducing the number of bounding box to be fed to the SVM classifier to close 2000 region proposals. Selective search uses the intensity, texture, color to generate all the possible location of the object. Then all the patches generated are transformed to fixed size patches and fed to the CNN. There are three important steps in R-CNN:

1) Run Selective Search to find the location of the object
2) Pass these patches to CNN, followed by SVM to classify the patches belong to which class.
3) Optimize the patches by training bounding box regressors.

B) Region-based Convolutional Networks
The replacement of Hog based classifiers with CNN based classifier had one major problem. CNN were too slow and computationally expensive. Hence,there was difficulty in running patches generated by the sliding window detector. Thus, R-CNN[16] was introduced as a solution which does the object recognition by selective search algorithm thereby reducing the number of bounding box to be fed to the SVM classifier to close 2000 region proposals. Selective search uses the intensity, texture, color to generate all the possible location of the object. Then all the patches generated are transformed to fixed size patches and fed to the CNN. There are three important steps in R-CNN:

1) Run Selective Search to find the location of the object
2) Pass these patches to CNN, followed by SVM to classify the patches belong to which class.
3) Optimize the patches by training bounding box regressors.

B) Spatial Pyramid Pooling
R-CNN was still too slow so to fix this SPP-Net [17] was introduced here, the CNN representation of an entire image is calculated then this is used to calculate the CNN representation of each patch generated by selective search. This is done by performing a pooling type of operation. Since, CNN requires a fixed size input SPP uses spatial

Licensed Under Creative Commons Attribution CC BY

pooling instead of max pooling. The SPP layer divides the region of any size into constant number of bins and max pooling is performed on each of the bin. However there is one problem it is not suitable to perform back propagation through spatial pooling layer. Hence it could be fine tuned only for the fully connected part of the network. But this led to the development of Fast R-CNN

C) Fast R-CNN

The disadvantage with SPP -Net is solved by Fast R-CNN[18] which borrows idea from R-CNN and SPP-Net making it possible to train end-to-end. In order to propagate the gradients through spatial pooling it uses a simple back propagation algorithm with the exception that pooling region overlap. Fast R-CNN added the bounding box regression to the neural network training, this reduced the overall training time and it increased the accuracy in comparison to SPP-Net due to end-to -end training.

D) Faster R-CNN

Faster R-CNN[19] is 10 times faster the Fast R-CNN this is achieved by replacing the selective search algorithm with a very small Region Proposal Network to generate the region of interest. Anchor boxes was introduced to handle the varying size of objects. The varying anchor boxes is passed along by applying spatial pooling similar to Fast R-CNN. The remaining network is similar to Fast R-CNN.

E) YOLO

In the early specified methods object detection is considered as a classification problem however, some models considered this as a regression problem and one of them is YOLO[20]. It predicts the bounding box and class probability in a single network evaluation. With earlier methods the bounding box usually contained objects but YOLO generates numerous bounding boxes which may not contain an object. The Non-Maximum Suppression technique is applied at end of network which merges many overlapping bounding box consisting of the same object into one. The difference with the above method is that it see the image at once rather then the regions generated by the Region Proposal Network in the above network. YOLO has a disadvantage it is unable to detect smaller objects.

F) Single Shot Detector

SSD uses a single deep neural network for object detection. This neural network is run once on the input image and it generate the feature map. The convolutional kernel is the run on the feature map to generate the bounding box and class probabilities[21].

G) YOLO9000 and YOLOv2

YOLOv2 was developed in order to improve the accuracy of detection while still being a faster detector. YOLOv2 uses a Res-Net like architecture to stack the low and high resolution feature maps to detect smaller objects. Batch normalization is used to prevent over fitting.YOLO9000[22] combines the ImageNet dataset with COCO dataset to be able to detect a precise objects. It is real time object detection model to detect 9000 categories of object.

H) Mask R-CNN

The similar approach to R-CNN but it adds a parallel branch to bounding box detection to predict object masks. The object mask is predicted as a segmentation of pixel in an image. It outperforms in the COCO challenge in instance segmentation, the bounding box detection, the object detection and the key point detection [23].

3. Proposed System

The implementation and training of neural networks require high computational hardware so in order to compromise to the low computational hardware a pre-trained model is used and transfer learning is applied. Transfer learning is a technique wherein the knowledge acquired from solving a problem is used to solve another but related problem. In order to be able to implement the object recognition in resource constrained devices such as a smart phone a neural network known as MobileNets[26] is used.

This model when combined with the SSD[21] gives a faster and efficient deep learning method for object recognition. Here fine-tuning is applied where the initial combined network of MobileNet and SSD which was pre-trained model with the COCO dataset was used. Later this model was fine-tuned with a similar dataset such as the PASCAL VOC dataset. Fine-tuning is done for much more accurate recognition depending on the application for which the model was fine-tuned.. Since, its application is implemented for the blind people the labeled output is converted to audio. The conversion of the labels to output is performed with the text to audio converter tools. Here the object recognition was performed in real time. The output displayed shows the object label along with the confidence value. The confidence value shows the probability of accuracy of detected object i.e If the object detected has a confidence value greater than a minimum threshold then the class label index is extracted and displayed.

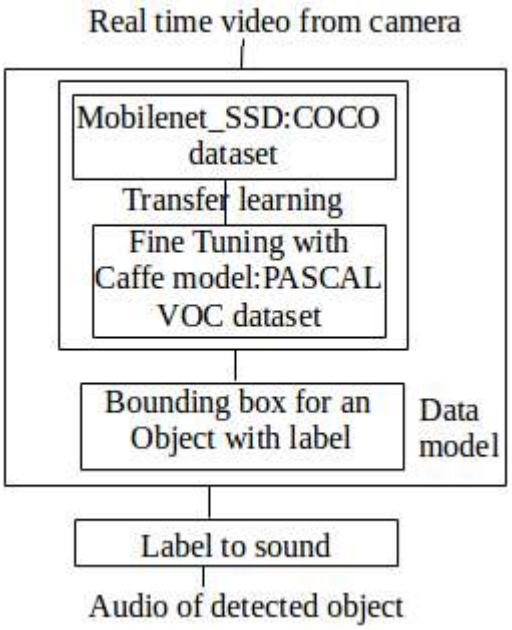

Figure 3: Proposed System

Licensed Under Creative Commons Attribution CC BY

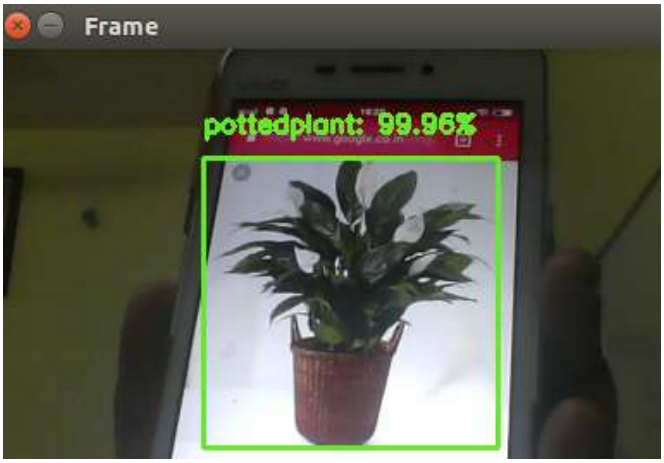

Figure 4: Proposed System Output

4. Challenges of Object Recognition

4.1 Variable number of objects

The machine learning models need to represent data in fixed size vector. As the number of objects in an image is not known beforehand there is an uncertainty on output number of objects thus this requires post-processing whichadds complexity. The sliding window approach is used to get the fixed size features in which some predictions are discarded while some are merged.

4.2. Size

The objects in an image are of varying sizes. So, generally an object which occupies the maximum area in an image is considered for classification. But, there exists object of very small size classifying these with sliding window is simple but it is inefficient.

4.3. Modeling

Object detection requires the process of location and classification so the problem arises on which mode to choose that is the one which includes hand engineered features or the deep learning model.

5. Dataset

There are several datasets released for the object detection. Researchers perform test on their algorithm with these dataset and give the predicted object accuracy as well as the spatial position of the object. The most commonly used dataset are as follows:

1) PASCAL VOC: Pascal Visual Object Classification[24] is a dataset used for object classification, detection and segmentation of objects. There are 10,000 images for training and validation with bounding boxes on object.
2) However it is able to classify only 20 categories. COCO dataset: Common Object In Context[25] dataset is developed by Microsoft. It consists of 80 categories. There are 1,20000 images for training and validation and more than 40,000 images for testing. The dataset changes each year. It is used for caption generation, object detection, object segmentation and key point segmentation.

6. Conclusion

This paper presents the overview on the evolution of the techniques used for object detection. Beginning from the hand engineered feature based classifiers to the convolutional network based classifiers in deep learning. It also presents the dataset which is commonly used in order to train, test and validate the object detected by the new models. The general overview of the proposed system and its implementation for the application of object recognition for the blind people is also introduced. The general challenges faced during object detection are also briefly described.

References

[1] Zhou, X.Y., Gong, W., Fu, W.L., Du, F.T,"Application of deep learning in object detection." IEEE/ACIS 16th International Conference on Computer and Information Science. pp. 631–634 ,2017
[2] Girshick, R., Donahue, J., Darrell, T., Malik, J." Region-based convolutional networks for ac te object detection and segmentation." PAMI ,2015.
[3] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: "You only look onc -time object detection."CVPR, 2016.
[4] Hu, Wiedo & Mohamed Gharuib, Ahmed & Hafez, Alaa," Template Match Object Detection for Inertial NavigationSystems.11.78-810.4236/pos.2011.22008.
[5] F. Khan, R. Muhammad , et.al., "Color Attributes for Object Detection," In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3306 – 3313, 2012.
[6] Khurana K, Awasthi R.,"Techniques for Object Recognition in Images and Multi-Object Detection". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 2(4):1383–1388
[7] P. Viola and M. Jones, "Robust real-time object detection,"International. Journal of Computer Vision, 57(2), pp.137–154, 2004.
[8] R. Fergus, P. Perona, A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition,"International Journal of Computer Vision, 2006.
[9] G. de Croon, "Active Object Detection," In 2nd International conference on computer vision theory and applications (VISAPP 2007), Barcelona, Institute for Systems and Technologies of Information, Control and Communication (INSTICC), pp. 97–103,2007.
[10] A. Berg, T.Berg , J. Malik, "Shape Matching and Object Recognition using Low Distortion Correspondences," In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp.26 – 33, 2005.
[11] H. Moballegh, N. Schmude, and R. Rojas, "Gradient Vector Griding:An Approach to Shape-based Object Detection in RoboCup Scenarios,"from:www.ais.uni-bonn.de/robocup.de/papers/RS11_Moballegh.pdf.

Licensed Under Creative Commons Attribution CC BY

[12] K.Schindler, D. Suter, "Object Detection by Global Countour Shape,"Pattern Recognition, 41(12), pp.3736–3748, 2008.
[13] A. Suga, K. Fukuda, T. Takiguchi, Y.Ariki, "Object Recognition and Segmentation Using SIFT and Graph Cuts," In 19th International Conference on Pattern Recognition ,pp. 1-4, 2008 .
[14] R. Oji, "An Automatic Algorithm for Object Recognition and Detection Based on ASIFT Keypoints," Signal & Image Processing:An International Journal (SIPIJ) Vol.3, No.5, pp.29-39, October 2012.
[15] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
[16] Liang, M., Hu, X.:"Recurrent convolutional neural network for object recognition. "In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3367–3375 (2015)
[17] K. He, X. Zhang, S. Ren, and J. Sun."Spatial pyramid pooling in deep convolutional networks for visual recognition. "In ECCV, 2014.
[18] R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015
[19] Ren, S., He, K., Girshick, R., Sun, J.:," Faster R-CNN: towards real-time object detection with region proposal networks.", NIPS 2015.
[20] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: "You only look once: unified, real-time object detection.", CVPR,2016.
[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, "SSD: Single shot multibox detector", 2015.
[22] J. Redmon and A. Farhadi."Yolo9000: Better, faster, stronger.",arXiv preprint arXiv:1612.08242, 2016
[23] K. He, G. Gkioxari, P. Doll ar, and R. Girshick. "Mask r-cnn.",arXiv:1703.06870, 2017.
[24] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zis-serman.,"The Pascal Visual Object Classes (VOC) Challenge.". IJCV,pages 303–338, 2010.
[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,P. Doll ar, and C. L. Zitnick.,"Microsoft COCO: Common objects in context. ",ECCV. 2014
[26] F. Khan, R. Muhammad , et.al., "Color Attributes for Object Detection," In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3306 – 3313, 2012.

Licensed Under Creative Commons Attribution CC BY