



A Literature Survey: Neural Networks for object detection

Aishwarya Sarkale¹, Kaiwant Shah¹, Anandji Chaudhary¹, Tatwadarshi P. N.²

¹(BE Computer Engg., VIVA Institute of technology, Mumbai University, Mumbai, India)

²(Asst. Professor Computer Engg., VIVA Institute of technology, Mumbai University, Mumbai, India)

Abstract: Humans have a great capability to distinguish objects by their vision. But, for machines object detection is an issue. Thus, Neural Networks have been introduced in the field of computer science. Neural Networks are also called as 'Artificial Neural Networks' [13]. Artificial Neural Networks are computational models of the brain which helps in object detection and recognition. This paper describes and demonstrates the different types of Neural Networks such as ANN, KNN, FASTER R-CNN, 3D-CNN, RNN etc. with their accuracies. From the study of various research papers, the accuracies of different Neural Networks are discussed and compared and it can be concluded that in the given test cases, the ANN gives the best accuracy for the object detection.

Keywords- ANN, Neural Networks, Object Detection.

1. INTRODUCTION

Artificial Neural Networks is a type of artificial intelligence that attempts to simulate the way a human brain works. Rather than using a digital model, in which all computations manipulate zeros and ones, a Neural Network works by creating connections between processing elements, the computer equivalent of neurons. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process[13]. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true for ANN's as well

Why Artificial Neural Networks?

1. Adaptive Learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. Self-Organisation: An ANN can create its own organisation or representation of the information it receives during learning time
3. Real time Operations: ANN computations may be carried out in parallel and special hardware devices are being designed and manufactured which take advantage of this capability.
4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

2. OBJECT DETECTION TECHNIQUES

Images of objects from a particular class are highly variable. One source of variation is the actual imaging process. Changes in illumination, changes in camera position as well as digitization of artifacts, all produce significant variations in image appearance, even in a static scene. The second source of variation is due to the intrinsic appearance variability of objects within a class, even assuming no variation in the imaging process. Object detection involves detecting instances of objects from a particular class in an image [14].

2.1 Object detection in images using artificial neural networks and improved binary gravitational search algorithm [1]

In this paper, Artificial Neural Network (ANN) and Improved Primary Gravitational Search algorithm (IBGSA) have been used to detect object in images. Watershed algorithm is used to segment images and extract

the objects colour, feature and geometric elements are separated from each question. IBGSA is utilized as a best technique to locate subset of components for array arranging coveted items. The reason for utilizing IBGSA is to diminish intricacy by choosing remarkable components.

Object recognition is an issue in clutter background, objects can be in various pose and lighting. Part base technology encode the structure by utilizing an arrangement of patches covering essential parts of an objects. In 3D ECDS, the edges of different objectives are segregated and the spatial relation of the same object is kept as well. A method of object detection that can combine the feature reduction and feature excerpt of PCA and Ada Boost.

Method:

In the current paper, Watershed, ANN and IBGSA are used for object detection. A lot of feature have been extracted from all these objects. Applying all these feature is time consuming and could grow calculation complexness of training ANN. Determining appropriate feature for knowledge can be used for this goal. For Example: there are some objects which automatically finds proper feature for object detection. In this methods selecting features from training objects are evaluated.

KNN classifier has low accuracy but high speed and recurrence of utilizing classifier in determination process. It is used as a part of this progression. By the point of upgrading the assessment work that is exactness of KNN classifier. In way of choosing highlights, because of its high effectiveness, ANN is utilized as a classifier, chosen highlights are utilized as a classifier, and chosen highlights are utilized for preparing ANN.

Advantages:

IBGSA is very useful in reducing extracting feature, which helps classifier for faster result.

Dis-advantages:

It uses KNN which have low accuracy as a classifier, but a good speed.

2.2 Comparison of Faster R-CNN models for object detection [2]

Object detection is a critical issue for machines. Faster R-CNN; one of the state-of-art object detection methods, approaches real time application. Moreover, computational impends on model and image crop size, yet accuracy is like-wise influenced; normally, time and accuracy have inverse relation. By altering input image size inspite downgrading performance, computation time meets criteria for one model.

In this paper, they have changed over a few best in class models from the Convolution Neural Network (CNN). At that point, we contrast changed over models and few picture edit estimate as far as calculation time and location accuracy. Examination information will be used for choosing an appropriate identification demonstration on the off chance that a robot needs to play out a question local assignment.

Method

CNN based feature extraction, features from RPN and CNN are taken by CNN. The CNN architecture from classification is used by extracting the feature from the image. Now CNN and RCNN is initialized by weights of CNN trained from image classification.

Region Proposal Network. CNN features pass small convolution network which perform a similar role to a hidden fully connected layer, and collectively thousands of anchors covering most region of image quality. Non-Maximum suppression is used to get regressed anchors before selecting ROI from anchors.

Region based CNN: Each ROI is classified and its box is regressed using the fast R-CNN. The feature from CNN are cropped by each ROI and only cropped features are pooled. Then pooled, features pass some hidden fully connected layers. Finally, they gather bounding boxes with scores. Additionally bounding boxes using Non-Maximum suppression to avoid duplicated detection.

Converting architecture

Exchange last pooling layer of CNN with ROI pooling layer. Last Classification layer of image classification with classifier and regression layer of Faster-RCNN.

Advantages:

Computation time has been rapid due to use of faster RCNN along with VGG16

Dis-advantages:

Enhancing time drastically diminishes performance.

Use of Faster-RCNN lead to lower in accuracy rate.

2.3 Detecting objects affordances with convolution neural networks [3]

A novel and real time method is shown to distinguish object affordances from RGBD pictures. This technique trains the Deep Convolution Neural Network (CNN) to learn profound features from the input data in an end-to-end manner. The CNN has an encoder-decoder design so as to get smooth label prediction. The information are represented to as various modalities to give the system a chance to take in component all more successfully.

Technique sets another benchmark on identifying order of object affordances enhancing the precision by 20% in correlation with cutting edge strategies that utilized hand-outlined geometric component. Besides this they apply direction strategies on a full size humanoid robot.

Human have a great capability to distinguish object by our vision. This helps in daily process of handling the objects. For a robot, detecting an object is essential to allow to interact with environment safely. Normally everybody used RGB-D images or point cloud data.

The benefit from this action leads to successful grasping action but fails in detecting other type of object affordances. Here unlike hand designed features are used, they treated this problem as pixel wise labelling task and use CNN to learn deep features from RGBD images. They show large CNN can be trained to detect object affordances from rich deep features. The affordances is studied quiet long time back in computer and robotics field.

Data representation:

Normally RGB-D images and cloud/depth images are used for training, but it is impossible to train a CNN by using limited dataset and having limited time. So a new methodology is being encrypted:

Horizontal disparity, Height above ground and Angle between each pixels surface and normal (HHA)

Advantages:

It is a novel method that has improved result than that of state-of-art method for object detection.

Dis-advantages:

Grasping method based on object affordances is limited to surfaces that fit the region.

2.4 3D Shapenets: A deep representation for volumetric shapes [4]

3D pattern is crucial but is heavily underutilized in todays computing system, mostly due to lack of good generic shape representation. With recent availability of inexpensive 2.5D depth sensors, it is becoming increasingly important to have a powerful 3D shape representation in loop.

Apart from this recognition, recovering full 3D physical body from persuasion based. 2.5D depth mathematical function is also critical part of visual understanding.

To this end, they propose to represent a geometric 3D shape as a chance distribution of binary variance of 3D Voxel Grid, using Convolution Deep Network. They have a 3D shape Nets, learns the distribution of complex 3D shapes across different objectives categories and arbitrary pores from raw CAD data and discovers hierarchical composition but representation automatically. It naturally support joint object recognition from 2.5D depth maps.

Usage of 3D shapenets

When provided with depth map of an object, it converts it into volumetric representation and identifies the observed surface and thus distinguishes it between free space and occupied space. 3D shape Nets can recognize object category complete all 3D shape and predict next best view if initial recognition is uncertain.

3D shape Nets to represent a geometric 3D shape as a probabilistic distribution of binary variables on a 3D vessel grid.

To train this 3D deep learning model, they construct Model Net, a large scale object dataset of 3D computer graphics CAD models.

Advantages:

3D representation for object and a convolution deep belief network to represent a geometric 3D shape as a probability distribution of a binary grid on a 3D voxel grid.

Disadvantages:

It is unable to jointly recognize and reconstruct object from single view i.e. RGB-D sensor.
A large dataset of 134M is used.

2.5 3D Object recognition from large scale point clouds with global descriptor and sliding window [5]

A novel strategy for object recognition has been proposed in this paper that mater given 3D model in large scale scene point. 3D model in large scale scene point. Since large scale indoor point clouds are greatly damaged by noise such as cluster, collusion, hole and points in a scene point cloud, based on similarities between local descriptor computed at key points on both point clouds. To avoid such problem they have come with idea to use sliding window with specific end goal to co-ordinate and pieces of scene points cloud.

They have used a bag-of-feature (BoF). A BoF representation if a window is efficiently calculated BoF vector. Though BoF is robust to partial noises it doesn't preserves any spatial information. Then global descriptor of a window which is almost invariant to horizontal rotation of object inside is been proposed. The task of 3D object recognition from unorganised point clouds has been studied widely from a long time. It is generally divided into two part, first estimates 6 degree of freedom poses of given specific models in environment scenes.

In first type, models are usually not contaminated by noises so that is easy to describe and exactly master their local shape around detected key points with local descriptor. In this, correspondence between models and scenes is calculated based on similarities of local descriptor. Then transition and rotation of input model are estimated from point to point matching by methods such as RANSAC methods of second type cut out individual object from at same point cloud at first and classify then with classifier obtained by supervised training using manually labelled data. In order to segment object from z background a clustering method like super voxels or plane removal by RANSAC is utilized. If the scene is simple like table top scene. It is easy to segment those pieces of point cloud that represents object from the scene.

Advantages:

Repetitive appearance of unhelpful primitive shapes and others is to detailed shape information due to noise is been tackled.

Disadvantages:

BoF is robust to partial noises, but it don't preserves any spatial information.

2.6 Scalable object detection using deep neural networks [10]

Deep convolutional neural networks recently demonstrated very impressive performance on a number of image recognition benchmarks. It has shown good performance on large scale visual recognition challenge. It was a winning model on localization subtask with the process by predicting single bounding box and identifying object category in the image. But the model cannot handle multiple instances of same object in the image. But the model cannot handle multiple instances of same object in the image. But now it can handle the same image having multiple instances and allows cross class generalization at highest level of network.

In this paper the computational challenge is addressed. Also this challenge becomes even harder when an object occurs more than once in the image. How they tackle this by generating a no of bounding boxes. For each box the output is a confidence score i.e. the likelihood of an image existing in that box. Various training exercises are performed for this. The predicted results and the real results are then matched for the learning purpose. They are capitalizing on the excellent learning abilities of DNN (Deep Neural Network). This approach has shown generalizing capability over unseen classes and can be used for other detection problems.

Now let us see the actual approach/methodology proposed in this paper. They use the Deep Neural Network which produces a fixed number of bounding boxes and then gives the output of each box as a confidence score.

Rounding box: The upper left and lower right coordinates are determined for the boxes. These boxes are adapted according to the dimensions of the image.

Confidence: The confidence score of each box is given as a single node value $C_i = 0$ or 1 .

After that they can combine the bounding boxes as a single layer. Similarly also the collection of the confidence scores can be treated as one output. In the algorithms the number of bounding boxes taken are between 100 and 200.

Training: The DNN predicts bounding boxes and confidence scores for each training image and then the highest scoring boxes are matched with actual values of the image. If M are actual number of images and K is the predicted amount. Then in reality the value of K is greater than M . Thus optimization is done of the predicted boxes which the ground truth ones.

Advantages:

It is able to capture multiple instances of same object in the image.

It is also able to generalize for categories was not trained on.

Disadvantages:

There are other methods showing better performance.

2.7 FPGA acceleration of Recurrent Neural Network based language model [11]

Recurrent neural network (RNN) based language model (RNNLM) is a biologically inspired model for natural language processing. It records the historical information through additional recurrent connections and therefore is very effective in capturing semantics of sentences. At architectural level the parallelism of RNN training scheme is improved and also reduces the computing resource requirement. Experiments at different network sizes demonstrates a great scalability of proposed framework.

RNN is a different type of neural network that can operate in time domain. RNN captures the long range dependencies using the additional recurrent connection. Then it stores them in hidden layer for later use. But the training costs in RNN was really high. So hardware up gradation was necessary to make it feasible. FPGA based accelerators have really caught the attention for tackling this problem.

Modern language models are based on statistical analysis. The n -gram model is one of the most commonly used model. What it does is it takes the probability of a word to exist after the word before it from the previous history. But when the value of n becomes more i.e. $n > 5$ then the computational costs really increase really increase. RNN comes to tackle this problem RNN uses its hidden layer to store historic information or previous information.

Most of the computational resources in RNN is spend on matrix vector multiplication. To overcome this or tackle it to some extent multiple cores are used for operations. But then this leads to high access memory requirement. Thus a proper balance between computation unit and memory bandwidth should be obtained be obtained by proper scalability.

Next comes the architectural optimization. It has various things to do in it. Like to increase parallelism between output layers and hidden layers. But it can only be done to a certain extent as there are limitations to it. Then there is the hardware implementations. The FPGA hardware design plays a huge role in supporting RNN.

Advantages:

Greater efficiency.

Extensive hardware tuning and modification is required.

2.8 An image matching and object recognition system using webcam robot [7]

Computer vision's vital steps is to find the relation among multiple images. Computer vision, is a science that makes machine capable to perceive the world around it in a similar way as human eyes and brain visually sense it.

This can be done if correspondence over consecutive frames in an image is tracked and matching among them is identified

This paper is based on image matching approach and is also based on the approach of field of ROBOTICS. Object Recognition involves identification, detection, and tracking. But, there are some challenges exist such as scale, view point variation, deformation, illumination etc. So, for best image matching and Object Recognition one of the optimal method named Chamfer matching is used. For best object recognition relevant features should be known. In this method, they have some local features such as point, edges and black and white points.

This paper can either be implemented through any hardware equipment to capture images or manually done by the user.

Step1:

All the nearest images of an object is stored manually in the database. Processing Algorithm is implemented after storing the images and are matched with current images taken by robot from different angles.

Step2:

Mobile Robot is fitted with CCD camera which controls through signalization. It is an eye of robot.

Step3:

Matching process within images using matching algorithm:

Here, they are using Chamfer Distance Transformation because of its simplicity and reliability. But before implementing 3-4 DT, the image is converted to grey scale & binarisation is performed to count the black & white points in the image. Also, Canny Edge Detector is used to detect the edge points.

Thus, this paper is based on the finding the matching percentage among two images that are exactly same, as well as slightly different and edited in some ways. Here, Chamfer Distance Transformation is used as it resulted efficient and high performance method for object detection due to its pixel based correlation approach.

Advantages:

The whole system is reliable & capable to match the two images in Digital Image Processing.

It uses Chamfer Distance Transformation that results in better performance for object recognition due to its pixel based correlation approach

Dis-advantages:

Here, Chamfer Distance Transformation, this algorithm is slightly time consuming because of the number of grey levels involved.

2.9 3D Convolutional object recognition using volumetric representation of depth data [8]

Convolutional Neural Network allow to extract features directly and automatically and produces better results in object Recognition. Here, RGB and Depth data are used in convolutional networks, volumetric information hidden in depth data are not fully utilized. So their system is proposed to utilize the volumetric information by 3D CNN. 3D CNN based approach is to exploit 3D geometrical structure of objects using depth data. Here depth data is used instead of RGB as RGB has rich colour, texture information while depth data has better ability representing 3D objects. Here, object can be recognized using only single depth images without having complete 3D model of object. There are 2 types of volumetric representation used.

Volumetric representation is used as it is providing simplicity to CNN and also good representation of 3D geometrical shape.

Volumetric Binary Grid

Volumetric Intensity Grid

In this method, input depth image is converted to a point cloud. The volumetric representation is found after de noising the point cloud to a 3D matrix space in which each cell represents a voxels. Volumetric Binary Grid represents the existence of surface point in voxel. 1 means present and 0 means absent & Volumetric Intensity Grid is to keep how many points a voxel represents. So, the voxel value is incremented by 1 for each projected point cloud value.

Now, this CNN architecture is composed of convolutional layers followed by leaky ReLU. This Convolutional layer have 32 filter with 5*5*5 and 3*3*3 sizes. The third layer is a pooling layer which down samples the input volume. The last two layers are fully connected layer. When they tested, they founded that, the proposed method handles the background problems without using masks and provides superior performance in the presence of background. This system has achieved higher accuracy than many state-of-arts approaches on the commonly used Washington RGB-D object Dataset. It is the first volumetric approach on this dataset.

So, 3D CNN on volumetric representation make it possible to learn rich 3D structural information of objects.

Advantages:

Higher accuracy.

First Volumetric approach in the Washington RGB-D object dataset

Volumetric Representation provides simplicity to CNN and good representation of 3D geometrical cues.

Dis-advantages:

Depth maps do not give enough information to build complete 3D model of objects.

2.10 A Shape Preserving Approach for Salient Object Detection Using Convolution Neural Network [12]

In computer vision what saliency does is, it identifies the most informative part of a visual scene. It also helps to reduce the computational complexity. This paper proposes a novel saliency object detection method which combines a shape preserving saliency prediction driven by a convolution neural network with low and middle-level region preserving image information. This model learns a saliency shape dictionary which is then used to train CNN. CNN then predicts the salient class of a target region and then estimates the full but coarse saliency map of the target image. Then the map is refined using image specific low-to-mid level data. The saliency map predicted by the CNN is further refined using the hierarchical segmentation maps by exploiting the global information such as spatial consistency and object boundaries. The proposed system outperforms the existing methods on popular benchmarks datasets.

2.11 Application of Deep Learning in Object Detection [6]

This paper mainly deals with the field of computer vision. The comparison between R-CNN, Fast R-CNN, and Faster R-CNN is the main focus of this paper. The above mentioned neural networks are similar to each other as the name suggests. Fast R-CNN and Faster R-CNN are the later versions of R-CNN. In this paper R-CNN, Fast R-CNN and Faster R-CNN are run across three different datasets i.e. Imagenet, PASCAL VOC and COCO. After the comparison the Faster R-CNN is the one that came out on top with most accuracy/precision. After determining that Faster R-CNN is the best amongst the three we tested it on the example of football field. Then its precision for various objects on the field is also mentioned.

2.12 Object Recognition and Detection by Shape and Color Pattern Recognition Utilizing Artificial Neural Networks [9]

A robust and accurate object recognition tool is presented in this paper. The paper introduced the use of Artificial Neural Networks in evaluating a frame shot of the target image. The system utilizes three major steps in object recognition, namely image processing, ANN processing and interpretation. In image processing stage a frame shot or an image go through a process of extracting numerical values of object's shape and object's color. These values are then fed to the Artificial Neural Network stage, wherein the recognition of the object is done. Since the output of the ANN stage is in numerical form the third process is indispensable for human understanding. This stage simply converts a given value to its equivalent linguistic term. All three components are integrated in an interface for ease of use. Upon the conclusion of the system's development, experimentation and testing procedures are initiated. The paper presents the following generalizations. The system's performance varies with the lighting condition with a recommended 1089 lumens with 97.93216% accuracy. Lastly the system contains a very high tolerance in the variations in the objects position or orientation, with the optimum accuracy at upward position with 99.9% accuracy rate.

3. ANALYSIS

The Table no.3.1 is a summary of studied research papers on object detection techniques and different classifier used. It enlightens on accuracy of various classifier from different papers:

Table no 3.1

Sr. No.	Paper Title	Classifier	Accuracy (In %)
1	Object detection in images using artificial binary gravitational search algorithm[1]	ANN & KNN (IBGSA)	91.70 & 61.4285
2	Comparison of Faster R-CNN models for object detection[2]	FASTER R-CNN (VGG-16)	68.1 & 80
3	Detecting objects affordances with convolution neural network[3]	CNN (HMP & SRF)	92.2
4	3D Shapenets: A deep representation for volumetric shapes[4]	Convolutional deep belief network (3D shapenets and model net)	80
5	3D Object recognition from large-scale point clouds with global descriptor and sliding window[5]	SVM(Adaboost) RANSAC	82.5
6	Object recognition and detection by shape and color pattern recognition using ANN[9]	ANN	99.9
7	3D Convolutional object recognition using volumetric representation of depth data[8]	3D-CNN	82
8	An image matching and object recognition system using webcam robot[7]	Chamfer distance transformation	70
9	Scalable object detection using deep neural network[10]	DNN ILSVRC	78.5
10	FPGA acceleration of recurrent neural network based on language model[11]	RNN FPGA	46.2
11	A shape preserving approach for salient object detection using convolutional neural network[12]	CNN SCSD	87.2
12	Application of deep learning in object detection[6]	R-CNN,FAST R-CNN,FASTER R-CNN, IMAGENET	66,66.9,73.2

4. CONCLUSIONS

In this survey extensive research and study of various neural networks was carried out. As time is progressing, the neural networks as well as the techniques for object detection are also progressing rapidly. Different neural networks have their own strengths and weaknesses. Some are a bit primitive like BPNN and others more advanced like ANN.

Like for example IBGSA is good for feature extraction, Faster R-CNN along with VGG-16 gives really good performance. This survey has described and compared various neural network very comprehensively and is providing a deep insight into the topic.

REFERENCES

- [1] F. Pourghahestani, E. Rashedi, "Object detection in images using artificial neural network and improved binary gravitational search algorithm", 2015 4th IEEE CFIS.
- [2] C. Lee, K. Won oh, H. Kim, "Comparison of faster R-CNN models for object detection", 2016 16th International Conference on Control, Automation and Systems, 16–19, 2016 in HICO.

- [3] A. Nguyen, D. Kanoulas, G. Caldwell, and N. Tsagarakis, "Detecting Object Affordances with Convolutional Neural Networks", *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 9-14, 2016
- [4] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, "3D Shapenets: A deep representation for volumetric shapes", *2015 IEEE*, 978-1-4673-6964-0/15.
- [5] N. Gunji, H. Niigaki, K. Tsutsuguchi, T. Kurozumi, and T. Kinebuchi, "3D Object recognition from large-scale point clouds with global descriptor and sliding window" , *2016 IEEE 23rd International Conference on Pattern Recognition (ICPR)*, December 4-8, 2016.
- [6] X. Zhou, W. Gong, W. Fu, F. Du, "Application of deep learning in object detection", *2017 IEEE ICIS*, May 24-26,2017.
- [7] J. Cruz, M. Dimaala, L. Francisco, E. Franco, A. Bandalá, E. Dadios, "Object recognition and detection by shape and color pattern recognition using ANN", *2013 IEEE 2013 International Conference of Information and Communication Technology*, 2013 IEEE.
- [8] A. Caglayan, A. Can, "3D Convolutional Object Recognition using Volumetric Representations of Depth Data", *2017 Fifteenth IAPR International Conference on Machine Vision Applications, MVA*.
- [9] S. Yadav, A. Singh, "An Image Matching and Object Recognition System using Webcam Robot", *2016 PDGC, IEEE*.
- [10] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, "Scalable Object Detection using Deep Neural Networks", *2014 IEEE Conference on Computer Vision and Pattern Recognition*.
- [11] Y. Wang, Q. Qiu, "FPGA Acceleration of Recurrent Neural Network based Language Model", *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*.
- [12] J. Kim, V. Pavlovic, "A Shape Preserving Approach for Salient Object Detection Using Convolutional Neural Networks", *2016 23rd International Conference on Pattern Recognition (ICPR),IEEE*.
- [13] S.N. Sivanandam, S.N. Deepa, *Introduction to neural networks using MATLAB 6.0* (Tata McGraw Hill Education, 2006).
- [14] Ramesh Jain, Rangachar Kasturi, Brain G. Schunck, *Machine vision*, (Tata McGraw Hill Education, 1995).