International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Deep Learning-Based Target Recognition and Localization

Mr.D Sanjeeva Reddy , Mr.A Srinivasan , Dr.R Vasanth Selva Kumar
Assistant Professor[1,2], Associate Professor[3]
Department of CSE,
Viswam Engineering College (VISM) Madanapalle-517325 Chittoor District, Andhra Pradesh, India

Abstract

Research and development of intelligent robots have seen increasing funding in recent years.

Whether it's the connected home, farming, manufacturing, or even

In the workplace, a human's superior cognitive abilities in recognizing and locating a goal far outweigh those of a robot. Studies in the past have used similar technology, such as computer vision and machine learning. While the intended result has been attained, the intelligence level is still rather low. One of the hottest topics in computer vision and AI right now is the development of robotic intelligent grasping systems that use deep learning algorithms and target location to investigate the speed and precision of target identification in a variety of settings. In this research, we use a single structured light camera to investigate how to determine the target's three-dimensional location using an enhanced deep learning technique called YOLO. The findings demonstrate its substantial real-time performance, accuracy, and other features. It can quickly recognize objects and adapt to new environments, and it can master human-level skills in picture recognition and target localization.

Keywords: monocular structured light, computer vision, recognition, and deep learning

INTRODUCTION

With the development of artificial intelligence technology, more and more people hope that machines human beings and be freely applied to various life scenarios. The realization recognition with the help of deep learning algorithm has more powerful adaptability to the environment, and it can overcome the change of environment and even task requirements to a certain extent. However, target positioning needs to determine the relative depth of the object and accurately measure the 3d information of the current object in the way of human processing image information. For a long time, people have tried to use machine learning to calculate the spatial position of the object in the scene, and promote two-dimensional image analysis to stereo vision analysis. Currently, monocular vision, binocular stereo vision and structured light have been widely used in 3d positioning algorithms.With the help of artificial intelligence, the research idea has changed from traditional machine learning algorithm to deep learning method. In 2015, Ian Lenz and others from Cornell university proposed a robotic grasping system based on RGBD,which is divided into two layers of network, using Stacked auto-encoder

[1]. The system shows better performance than

the old fashioned method of marking. Lerrel Pinto and Abhinav Gupta of Carnegie Mellon University created a framework for self-supervised learning in the same year. The concept of

Using RL to iteratively train the CNN network [2] is a kind of reinforcement learning. At the 2017 ICRA, UC Berkeley's Jeffrey Mahler, Brian Hou, Sherdil Niyaz, Florian T. Pokorny, Ramu Chandra, and Ken Goldberg .

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

[3] showed out Dex-Net 2.0, a method for training neural networks that gathers data in simulated environments by fusing depth maps with capturing postures.

It's worth noting that in a study published in 2016, the Deep Mind team addressed the transfer of knowledge from simulated to actual settings. As a proof-of-concept technique, progressive networks were presented.

[4] as the link between virtual and actual environments. Some researchers in the United States have begun using deep learning techniques to teach robots to grab in recent years. A robot grab discrimination approach using multi-model deep learning was suggested by Xungao Zhong and colleagues in 2016

A deep network model was constructed using a stacked automated noise reduction encoder, and network weight learning was achieved by the use of both automatic noise reduction coding and sparse restrictions.

[5] When training the network, they utilized RGB-D data captured with the Kinect motion camera. Xiru Wu used high-resolution industrial cameras to collect images of his targets

[6]. After that, they utilized a deep convolutional neural network to determine the targets' locations and types. Deep learning-based object identification and capture has emerged as a major topic in the study of AI.

Optimizing the YOLO technique, further investigating the tuning strategy of deep learning, and making use of the already-existing training database, this study aims to increase the speed and accuracy of picture identification. Next, we focus on how to pinpoint an object with great accuracy by discussing the mapping connection between a two-dimensional picture and stereo space. Our experiment turned out well in the end. This paper's sections are structured as follows: In the first section, we provide some context for the study and discuss its relevance. In Section 2, we compare and contrast the training outcomes with previous research methodologies, and investigate the underlying principles of the deep learning algorithm and monocular structured light. Section 3 employs the Orbbec camera to obtain precise target location and conducts experimental verification of the algorithm's efficacy. The study finishes with Section4 and suggests directions for further research.

RELATED WORK

A. The YOLO Algorithm: Its Basics and Future Development

The challenges of global scale variation in are generally brought on by the wide range of light intensities and the variety of target characteristics.

target, challenging feature extraction of tiny target, and partial occlusion of target, which calls for the highly adaptable target identification and positioning system. Important findings in the area of target identification have been published thanks to the deep learning algorithm's ability to more effectively complete a variety of classification tasks and extract features. At its heart, deep learning is the process of training a computer to mimic the learning behavior of a human brain. This is done by first creating a model of a neural network and then optimizing its parameters over the course of several training cycles. Deep learning has been the attention of academic circles and the intelligent manufacturing business thanks to its success in improving the accuracy and speed of target identification via AI research and investigation.

For prediction following feature extraction, the YOLO algorithm follows the traditional pattern of using the fully connected layer. The total network architecture is shown in Fig. 1, while the backbone network itself uses Darknet-53. In order to speed up evaluations and increase efficiency, this network design is able to make more

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

effective use of GPU. In most cases, a convolutional neural network would use a "candidate area box" to do the necessary classification and regression, with the highest score serving as the conclusive conclusion of the target identification process.

It is possible for candidate boxes to repeatedly find the same target, and there is some overlap between them. One-Stage Detection, or YOLO for short, is a kind of end-to-end network.

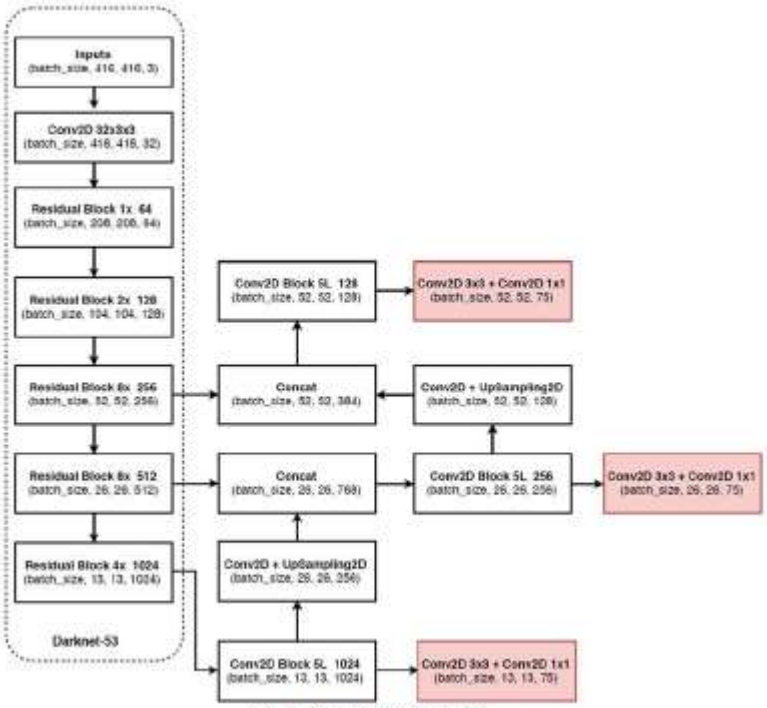

Fig.1. Network Structure

The method does not create candidate boxes but rather inputs the whole image into the model to produce a predetermined number of bounding boxes.

The whole image is broken down into SS grids using a neural network (using

multi-scale prediction with three output layers (13x13, 26x26, and 52x52 SS grids). Each grid's center and A and B radii serve as a prediction for the target. Each

box's confidence and position are restored. The model's fine-grained features come from a technique called superimposing the neighboring features of the shallow feature map to other channels, which improves the model's capacity to detect tiny objects. Coordinates (bx, by, bw, bh), confidence levels, and the total number of categories must all be predicted for each cell, for a total of 3(4+1+n). Tensor SS3(4+1+n) is the final predicted value for each layer (3 is the number of anchor boxes per layer). The threshold determines which target to prioritize in the results. In this approach, k-means clustering is used to determine the locations of 9 anchor boxes. Clusters 10, 13, 16, 30, 33, 23, 30, 61, 62, 45, 59, 119, 116, 90, 156, 198, and 373 may be found in the COCO dataset. It's been shown that earlier boxes of varying sizes match to feature maps of varying sizes. The category confidence value of each bounding box is calculated by multiplying the category information predicted by each grid by the confidence value of the bounding box prediction. After calculating the confidence score, a threshold may be used to exclude the prediction box that had a low score. The high-scoring box was processed

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

with non-maximal suppression to prevent producing redundant prediction boxes. I won't bore you with the specifics of the loss function, but to put it briefly: we utilize the IOU (Intersection over Union) for the coordinate values, and the cross loss entropy for the category probabilities.

A predetermined number of bounding boxes are used in the method to make predictions about the target's location, and the technique successfully finds the coordinates of both isolated and dispersed targets.repetitive placement, and maximizes the picture data contained inside the bounding box. As a result of this direct regression on the picture, the detection effectiveness of the target box is substantially enhanced.

B. Tuning strategy of training network

Deep learning has a longer training period compared to traditional machine learning, and the process of adjusting the training parameters is crucial. Final Outcomes results produced by using various methodologies for modulating network parameters to the same network configuration might vary widely [8, 9]. Methods for optimizing tuning used in this article include:

1. Data Transformations (by means of translation, rotation, and scaling)

(2) Activation function that is suitable (leaky ReLU function)

3.setting the beginning weight correctly.

4.the right optimization method (stochastic gradient descent, or SGD).

Reduce the pace of learning over time

(#5) Establish the amount of momentum

(#6)

C. Monocular Structured-light Scheme

It's challenging to pinpoint an object's exact location with monocular vision, extracting features from images is challenging, and creating dynamic fuzziness is simple. The use of binocular stereo vision,

matrix inversion and the linear triangle are employed

as a means of locating a certain location. Although the target's location is determined, no 3D point cloud containing the finer features of the target exists. Spotlights, striped lights, and other forms of structured light are all examples. and other forms of light, such as encoded structured light. The surface shape or depth information of the measured item may be obtained by projecting such a one- or two-dimensional picture onto it and analyzing the resulting size distortion. TOF produces a surface light source rather than speckle, therefore its intensity and depth accuracy are relatively unaffected by distance up to a certain point. It consumes more power and produces lower-quality depth images. Binocular stereoscopic vision involves the use of two or more cameras for simultaneous image capture and the use of an algorithm to derive depth information from analyzing the differences between pictures captured by each camera.

While binoculars work well in the great outdoors, they aren't suited for situations where texture change is subtle. TOF is more precise at great distances, but its usefulness is presently limited by poor picture quality and excessive battery needs. As can be shown in Fig. 2, the overall performance of the structured light scheme is the

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

greatest of the three 3d sensing technologies, hence we choose to conduct our experiment using a monocular structured light Astra- S camera.

Fig.2. Astra-s Camera

Orbbec integrates research and development, manufacturing, and sales for 3D vision Sensor solutions. It's just the fourth firm in the world capable of producing 3d sensors in large quantities. When compared to traditional cameras, 3D cameras

in-depth details. The depth value represents the height of the target above the plane of the camera. For applications including motion capture, 3D modeling, interior navigation, and positioning, the 3D camera's ability to acquire depth information, 3D scale, and spatial information in real time is crucial. Orbbec's OpenNI2.3 series SDK is built on top of OpenNI2, a cross-platform, multi-language framework that specifies how software and middleware should talk to each other and how 3D sensors should be read. To achieve this particular debugging method, we may utilize the OpenNi2 header file on vs2013.

D. Calculation of Target Coordinates

Each of the anchors from the preceding paragraph is made up of two numbers, one representing the anchor's height and the other its breadth. To make accurate predictions of future border boxes, the applied logistic regression. algorithm. Then, using the method in Fig. 3 [7], we can get the absolute coordinates (bx, by, bw, bh). Finally, we directly \sestimate the relative location and forecast the related coordinates of the target center point relative to the upper left corner of thegrid unit.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

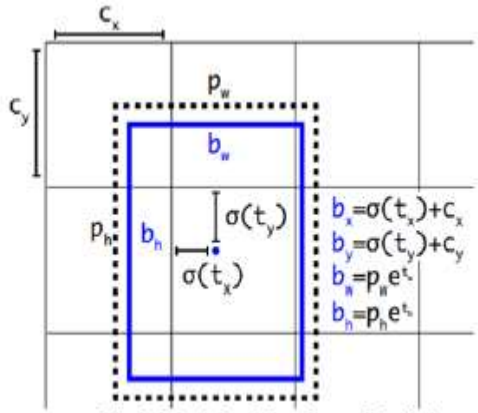

Fig.3. Principle of Coordinate Calculation

Here, tx, ty, tw, and th represent the model's forecasted results. The grid cell's Cx and Cy coordinates. For\sexample, if the feature map size of a layer is 13×13, then there

include coordinates for a 13-by-13 grid, with the 0th row and 1st column including (0,1). Predicted width and height, or pw and ph, respectively, are the bounding box sizes. bx, by, bw, and bh are the \scoordinates and size of the center of the anticipated bounding \sbox. The loss of coordinates is the square error loss. Perform \ssigmoid on tx and ty, then add the relevant offset, then \smultiply by the associated stride, that is 416/13, 416 ⁄ 26, \s416 ⁄ 52. Finally, sigmoid the object and classes confidence to \sget the probability (0-1). (0-1). Softmax will increase the highest category probability value while decreasing all other category probability values, which is why sigmoid is utilized.

EXPERIMEN

A. Target Recognition

this experiment, we test the updated Yolo algorithm using RTX2080ti hardware, Tensorflow as the experimental framework, and the Coco database. Considering the experimental Recognition

The COCO dataset is chosen for careful examination in the office situation. Simply input the relevant dataset and modify the network structure if you need to add a specific target for identification.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Fig.4. Result of target positioning

Fig 4 depicts the outcome of the implementation. The dataset comprises 90 categories, majority of the items are non-central, and there are many little targets and single picture targets.

distribution. It is harder to detect and more suited to the everyday surroundings. Good experimental results are still obtained even with modest experimental equipment, as illustrated in Fig 4. In comparison to RetinaNet's results, the algorithm can process 416x416 images at a speed of 29 FPS and mAP@0.5 up to 55.3%.

B. Three-dimensional Positioning

Target recognition has been accomplished using the deep learning technique from the previous section. We then find the target's 3D coordinate information in the recognition box.

The center point of the target in the contour is used as its location point after applying the coordinate transformation principle to translate the pixel coordinates into spatial coordinates.

We extract the target from the image, set a target-specific color threshold, then extract the target's contour based on the positioning frame's coordinates. Any point on the contour can have its 3D coordinates extracted using the camera image model's basic principal. We won't go into detail because camera calibration is not the main topic of this study. For information, kindly

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Fig.5. Result of Three Dimensional Localization

Our test results are shown in Fig 5. The three targets that have been detected are shown on the left, and their matching 3D coordinate value with millimeter accuracy is shown on the right.

The absolute average error of our depth distance is less than 9 mm when compared to the manual approach. The findings of the experiment indicate that it is possible to solve 3D spatial coordinates by calibrating the camera and imaging model. The high coordinate value of the solution in the 0.5–1.5 m field of view attests to the usefulness of adopting a structured light scheme for target location.

CONCLUTION

Target location and recognition based on deep learning algorithms has become a popular topic in pattern recognition in recent years. People don't just wish to be liberated from mundane and

repetitious tasks, but also more fervently believe that robotic intelligence will be able to meet people's needs on its own, lessening the load on families and society.

achieving a range of intelligent services. This work performs experimental research on target recognition and three-dimensional location using the grasping of the office service robot as its backdrop. In this article, we demonstrate that the improved YOLO method boosts recognition speed while maintaining accuracy by examining the fundamental theory behind the present advanced deep learning algorithm. We also carry out a theoretical investigation of how deep learning parameters are changed. Then, by contrasting several sensor methods, we selected the most relevant experimental approach. Finally, using the combination of a monocular and telescope, we finish the precise target positioning.

The robot arm mathematical model, MATLAB simulation, and building the robot grabbing mechanism are our next tasks. The inverse and forward kinematics

Analysis is done in conjunction with the target's three-dimensional data. The goal is to implement the exact gripping task for the office service robot by successfully applying the analysis of deep learning theory to the robot system.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

ACKNOWLEDGMENT

This research has been made possible by grants from the Anhui Natural Science Foundation (No. 1808085QF193) and the National Natural Science Foundation of China (No. 61703390).

Meiling Wang is the corresponding author of the technology support plan important projects of Jiangsu province (No. BE2017007).

REFERENCES

[1] Lenz, Ian, Lee, Honglak, Saxena, Ashutosh. Deep Learning for Detecting Robotic Grasps[J]. International Journal of Robotics Research, 2013, 34(4-5):705-724.

[2] Pinto L, Gupta A. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours[C]. IEEE International Conference on Robotics and Automation. IEEE, 2016:3406-3413.

[3] Mahler, Jeffrey, Liang, Jacky, Niyaz, Sherdil, et al. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics[J]. 2017.

[4] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive Neural Networks[J]. 2016.

[5] Xungao Zhong, Min Xu, et al. A Robot Grasping Discrimination Method Based on Multi-mode Feature Depth Learning[J]. Acta Automatica Sinica, 2016, 42(07):1022-1029.

[6] Xiru Wu, Guoming Huang, et al. Fast visual recognition and localization algorithm for industrial sorting robot based on deep learning[J]. ROBOT, 2016, 38(6):711-719.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016:779-788.

[8] Wang X. Research on robotic grabbing system based on fast SSD deep learning algorithm [D]. Wuhan University of Science and Technology, 2018.

[9] Liu J. Research on robot training simulation training technology based on deep learning [D]. Harbin Institute of Technology, 2018.

[10] Zhang Z. A Flexible New Technique for Camera Calibration[J]. TPAMI,

2000, 2000, 22(11):1330-1334.