**A REPORT**

**ON**

**Vision-Based Robotic Grasp Detection Using Deep Learning** **Algorithm**

**BY**

**MERIAM JOSEPH 2018A7PS0291U CS**

**AT**



**BITS, Pilani – Dubai Campus**

**Dubai International Academic City (DIAC)**

**Dubai, U.A.E**

**First Semester, 2021-2022**

**A REPORT**

**ON**

**Vision-Based Robotic Grasp Detection Using Deep Learning Algorithm**

## BY

**MERIAM JOSEPH 2018A7PS0291U CS**

**Prepared in fulfillment of the**

**Project Course: CS F376**

**AT**

****

**BITS, Pilani – Dubai Campus**

**Dubai International Academic City (DIAC)**

**Dubai, UAE**

**First Semester, 2021-22**

**BITS, Pilani – Dubai Campus**

**Dubai International Academic City (DIAC)**

**Dubai, UAE**

**Course Name: Design Project**

**Course No: CS F376**

**Duration: 4 Months**

**Date of Start: 5-08-2021**

**Date of Submission: 14-12-2021**

**Title of Report: Vision-Based Robotic Grasp Detection Using Deep Learning** **Algorithm**

**Name: Meriam Joseph**

**ID Number: 2018A7PS0291U**

**Discipline: Computer Science Engineering**

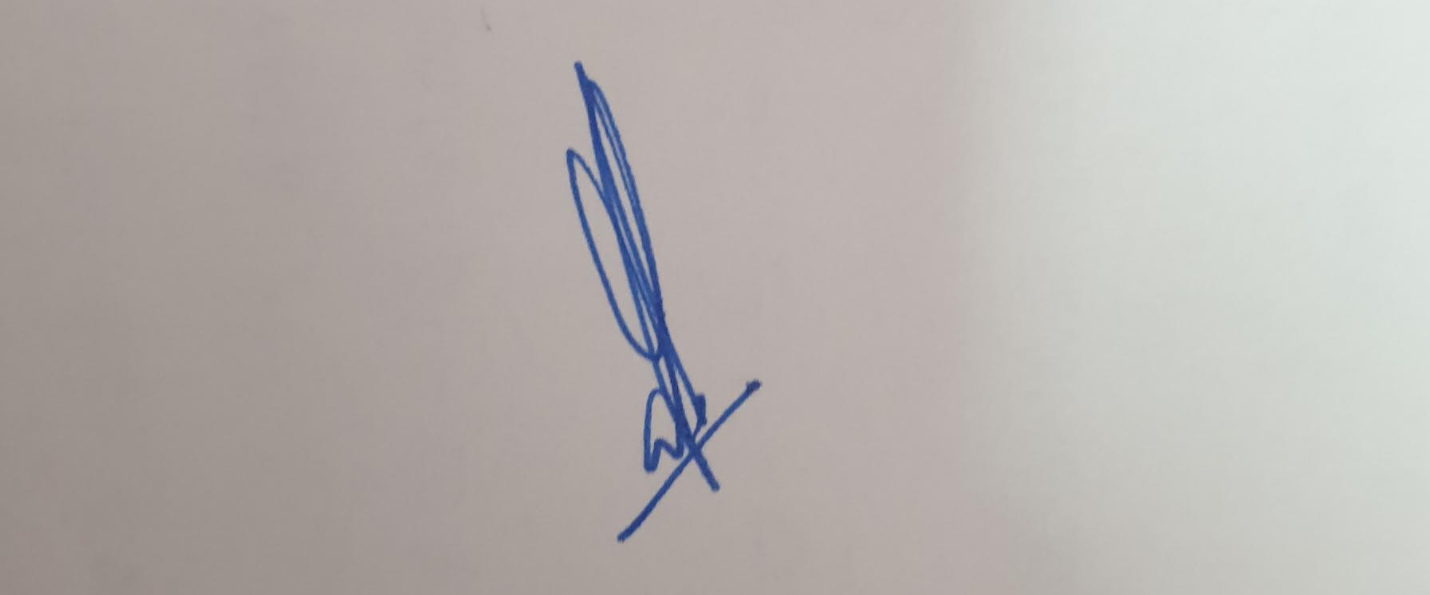
**Name of Project Supervisor:** Dr. V. Kalaichelvi

**Keywords:** GR-ConvNet, Grasp pose, IoU, Deep Learning, Grasp rectangle

**Project Area:** Machine Learning, Artificial Intelligence

**Abstract:**

Deep Learning has significantly advanced computer vision and natural language processing. It is based on artificial neural networks modeled on the human brain that aim to mimic the way humans learn and remember things. For example, humans are extremely capable of manipulation, they can see a few objects around them and immediately determine the optimal way to pick them up and do so with the utmost ease. However, for machines and robots, even after all the advances, this is still something that they cannot operate on the same level in. Therefore, robotic arms, that can grasp objects with the same level of dexterity as humans is a topic that is heavily researched and focused on, since it has a high demand for in everyday and industrial applications, especially due to the boom in e-commerce in the recent decade, where everything is stored in massive warehouses. As a solution to this, recent advancements in artificial intelligence have enabled vision-based robotic models that train robots to visually identify positive robotic grasps positions for objects through deep learning techniques. However, robots to this day still cannot identify and grasp objects around them to the dexterity that humans can. Hence, deep learning models like GR-ConvNet (Generative Residual Convolutional Network) aims to achieve optimal robotic grasp predictions for objects. The project aims to collect image data, train the model using the data, implement the model in real time and perform analysis work. Furthermore, in future work we can extend the project to optimization of the GR-ConvNet model and implementation of said optimal model to achieve optimal results.



**Signature of the Student Signature of Faculty**

**Date: 20-12-2021 Date:**

**Acknowledgements**

Firstly, I would like to express my heartfelt gratitude to the Director of BITS Pilani, Dubai Campus, Prof. R.N. Srinivasan Madapusi, who has ushered a new light on our college.

Next, I would like to thank Dr. V. Kalaichelvi, my Project Supervisor, for her faith in me even before I started working under her and for her motivation, support, guidance, and encouragement throughout the course of this project. She has been a great influence, urging me to learn and to be innovative, pushing me to excel.

I would like to thank Dr. M Raja, Instructor-in-Charge of Design Projects, for taking his time to investigate my work and for his support regarding my project.

Lastly, I would like to thank Ms. Abhilasha Singh, who worked alongside me and helped me through this project

A picture containing text, whiteboard

Description automatically generated

Meriam Joseph

2018A7PS0291U

**Contents**

Abstract

Acknowledgement

Table of Contents

List of Figures

**Chapter 1*:*  INTRODUCTION**

[1.1 INTRODUCTION](#_Introduction) …………………………………………….……………………1

[1.2 OBJECTIVES OF THE PROJECT](#_Specifications_–_General) ………………………………………….…..1

**Chapter 2*:* LITERATURE SURVEY………………………………………………………..2**

**Chapter 3*:* PROBLEM FORMULATION**

[3.1 INTRODUCTION TO THE PROBLEM](#_Lead_Compensators) ………………………………………….4

3.2 PROPOSED METHODOLOGY AND EXPERIMENTAL SET-UP………......4

**Chapter 4 SIMULATION, RESULTS AND DISCUSSION**

4.1 DATASET………………………………………………………………………….8

4.2 TRAINING RESULTS....…………………………………………………...…….8

4.3 TESTING RESULTS……………………………………………………………...9

4.4 PERFORMANCE ANALYSIS……………………………………………………11

**Chapter 5*:* SUMMARY AND CONCLUSION…………………………………………..…**.13

[**References**](#ref)

**LIST OF FIGURES**

[Figure 1 Flow-chart of proposed methodology 4](#_Toc90813820)

[Figure 2 Inference model for grasp prediction 5](#_Toc90813821)

[Figure 3 Proposed GR-ConvNet Model 7](#_Toc90813822)

[Figure 4 Images from Cornell Dataset 8](#_Toc90813823)

[Figure 5 Confidence Threshold vs IoU Measure 11](#_Toc90813824)

**LIST OF TABLES**

[Table 1 Training Results 8](#_Toc90813828)

[Table 2 Sample images from custom dataset for testing on GR-ConvNet model 9](#_Toc90813829)

[Table 3 Evaluation Results 10](#_Toc90813830)

[Table 4 Real-Time Grasp Pose of Image Obtained 10](#_Toc90813831)

[Table 5 Confidence Threshold vs IoU Measure of proposed GR-ConvNet model 11](#_Toc90813832)

CHAPTER 1

**INTRODUCTION**

# Introduction

In the recent years, machine learning, a type of artificial intelligence has made it possible for computers and machines to improve algorithms automatically using data and experience. Deep learning is based on artificial neural networks modeled on the human brain that aim to mimic the way humans learn and remember things. It is a type of machine learning. Humans have always had the ability to learn, re-learn and implement their learning with the utmost ease. Humans are capable of identification of objects, determination of optimal way to pick up said objects and manipulation. [1] However, despite decades of technological advancements, for machines and robots, this is still something that they cannot operate on the same level in. [2] Therefore, creating robotic arms, that can grasp objects with the same level of dexterity as humans isa topic that has been heavily researched and focused on.

Robotic grasping describes a robot’s ability to grasp and manipulate object in its environment. [3] Grasping is a canonical problem in robotics, and has been explored for many decades and hence, had yielded multiple techniques. [4] The greatest advantage of deep learning approaches is their ability to generalize to previously unseen objects. [5] Furthermore, allows for us to work directly on gnarly data in the form of image, video, and audio. Unlike traditional machine learning algorithms that must preprocess the data in some way, deep learning algorithms do the same without being programmed for in the algorithm. Another characteristic of deep learning is that the training process involves mathematical vector operations and results in billions of computations. To meet this need GPUs are necessary. [6]

The steadily heavy demand for customization in industries such as the mining industries, fashion industries etc. require robots can perform the specific grasping tasks with the utmost accuracy and precision. Furthermore, the boom of e-commerce and online trading in the recent decade has raised a demand for robots that can overcome the challenge of putting in and pulling things out of bins and containers and packaging them in warehouses. This is a huge demand in industry as there aren’t enough people to do these jobs and hence, an automated robotic solution is needed. Hence, vision-based robotic grasp detection using deep learning is a topic worth exploring since it allows automation of tasks without any human intervention and significantly reduces programming efforts. The ability to grasp arbitrary objects, if could be solved would have significant impact in factory warehouses and homes. [7] In this report, we explore how GR-ConvNet algorithm can be used to predict robotic grasps.

# Objectives of the project work

* Collection of data for object detection applied to robot grasping tasks.
* Deep learning techniques like GR-ConvNet are used for training the dataset.
* Real-time testing by using the live image captured from Intel RealSense camera.
* Performance analysis.

CHAPTER 2

LITERATURE SURVEY

A Generative Residual Convolutional Neural Network (GR-ConvNet) model was proposed in [8] for the purpose of predicting and producing antipodal grasps for a diverse range of household and adversarial items in real-time (∼20ms) using a 7-DoF Baxter robotic arm. This was tested against the Cornell and Jacquard datasets, resulting in an accuracy of 97.7% and 94.6% respectively. The depth-prediction module could be further improved upon to properly predict the depths of reflective items such as glass bottles to improve accuracy.

The model in [9] can predict multiple grasp coordinates for one or more items from an RGB-D image using a deep neural network. Grasp candidates are chosen based on a score that takes grasp sparsity and proximity to ground truth into account. ResNet-50 with an ROI pooling layer is chosen for the pre-training process and Nvidia Titan-X for the network training phase. The model achieves a 96% accuracy on the Cornell dataset and an 89% grasp success rate for household items.

A Generative Grasping Convolutional Neural Network (GG-CNN) can measure the grasp quality at every pixel in [10]. Its 50 Hz closed loop control, implemented via Position Based Visual Servo (PBVS) controller, helps in generating grasp candidates for moving items as well as dynamic clutter. It results in an 88% accuracy rate for moving household items and 81% for cluttered items. However, because grasp pose couldn’t be updated when the camera gripper is closer than 70mm from the item, the closed loop method falls short for small items like dice as well as adversarial items.

The system in [11] utilizes a Max-pooling Convolutional Neural Network (MPCNN) with each class having a different item pose assigned to it. It uses a Basler camera and a Schunk Dexterous robotic arm. The arm has a gripper, and the center of each pose class is used as the gripper’s target pose. This helps in pose estimating as it will gain a higher accuracy when closer to the center pose of each class. The database is made up of five items, each having different poses. This model can achieve a high accuracy for pose estimation.

Deep Convolutional Neural Networks (DCNN) make up the ROS framework for a real-time grasp detection system proposed in [12]. It can generate multiple grasp candidates using just one input in the form of an RGB picture. The authors create a loss computing function which relies on a Separate Axis Theorem (SAT) to counter any drawbacks such as the number of items in frame or the resolution of the images. It has achieved a 73% accuracy rate for single-grasp and 87.1% for multi-grasp network using the Cornell dataset.

From the above papers, we have observed the problem of robotic grasp from a state of world that is partially observed (a tabletop) or is stimulated. There is a huge gap between the simulation and the real-world. Therefore, there is a need for automatic adaptation to real world environments in the problem of robotic grasps. There is also a need for universal grasping- the ability of the model to predict grasp rectangles for any objects placed in front of the robot and not just be able to pick the same widget repeatedly. This is to be achieved with a deep learning model and not just use a complex gripper as we want to be able to use a parallel jaw gripper, a lightweight and inexpensive reliable option to grasp objects.

CHAPTER 3

**PROBLEM FORMULATION**

# Introduction to the problem

Our goal in this project is to ensure accurate prediction of grasping of objects with vision-based robotic grasp model using deep learning. For this, we have chosen the GR-ConvNet model and implemented it to display positive robotic grasps for offline images and in real-time.

The three most important parameters that determine the grasp in this model is the depth, angle, and quality of the image of the object. These three parameters are used to calculate the grasp pose.

**3.2 Proposed methodology and Experimental Set up**

Graphical user interface

Description automatically generated with medium confidence

Figure 1 Flow-chart of proposed methodology

For this project, as shown in Fig 1, we define our proposed methodology, which encompasses steps such as dataset collection, training, and testing. We start with identifying the datasets to be used for the purpose of our research and pre-processing the data before we implement the training functions on the dataset. We then start training the model, on our network which is the GR-ConvNet and save the model and record the loss at each epoch running for 50 epochs each. After training, we move on to test the trained models, and use IoU (Intersection Over Union) as our evaluation metric to evaluate the accuracy of the models. The evaluation and testing of the model will be done on both CPU and GPU and with the evaluation accuracy, we finally implement our grasping algorithm with object detection assess the image – (which can be either an offline image or real-time image obtained from a working camera) on the trained model and generate a plot of the positive grasp rectangle or grasp pose over the image along with the width, angle, and depth plots.

Diagram

Description automatically generated

Figure 2 Inference model for grasp prediction

The inference module consists of three different parts, as shown in Fig 2: image pre-processing, post-processing, and grasp rectangle. Firstly, the data that is input is pre-processed. This preprocessing of the input image includes cropping, resizing and finally normalization of the input image. This 224 × 224 n-channel input data is then fed to the proposed GR-ConvNet model. An advantage of this n-channel model is that, since it is generalized and not specific to a certain type of input image, we are able to choose the type of input e.g., Depth-only, RGB-only or RGB-D as per convenience.

The second part of the model generates three images from the input image which represent the grasp quality, grasp angle and grasp width. This is done by extracting the features from the preprocessed image by using the GR-ConvNet model. Finally, the third part of the model infers grasp poses from these three output images by proposing grasp rectangles and plotting them.

The model generates grasp rectangles based on rectangle metric proposed by Yun Jiang. According to the rectangle metric, a grasp is valid when it satisfies the following two conditions.

1. The IoU score between the predicted grasp rectangle and ground truth grasp rectangle is greater than 25%

2. There’s a 30 ° offset between grasp orientation of the ground truth rectangle and predicted grasp rectangle.

The IoU score is calculated according to Jaccard Similarity or Jaccard Index, which is used to calculate the similarity and diversity of sets. This is according to the given formula (1).

In Gr-ConvNet, A and B will represent the ground grasping rectangles and the predicting grasping rectangle. Thereby, we consider the pairs with a Jaccard index greater than 0.25 to be a successful predicted grasp.

A picture containing diagram

Description automatically generated

Figure 3 Proposed GR-ConvNet Model

**Fig. 3** shows the GR-ConvNet (Generative Residual Convolutional Neural Network) architecture. This is a generative architecture which takes in a n-channel input as per convenience and then generates a pixel-based grasp with three images as output which as quality, angle, and width. Quality is the grasp quality score, angle represents the angle required, and width, the image width value to grasp the object in the image. The convolutional layers perform extraction of features from the input image and the subsequent output for each layer is then fed in as the input for the following residual layers which are 5 in number for this architecture. After passing the input image, we find out that from the initial size of 224x224 it has been reduced to 56x56 which is very difficult to interpret. This is the reason why we up-sample the image to match the input size for it to be easy for interpretation. The network has 1,900,900 parameters in total, which is shorter than the other networks which have been used in previous studies. This makes the entire process computationally efficient and faster leading to a simpler process for grasping objects while using a smaller number of parameters.

For the real-time testing of the proposed model, we have made use of the Intel RealSense SR305 Camera with a serial number of 635201005367, which allows for input of 640x480 depth resolution at up to 60 frames per second. It is a short-range depth camera; hence we can obtain grasp positions of objects by plugging in either RGB images, depth images and even both.

Furthermore, data segmentation, can also be applied to the deep learning model. This is a technique that slightly modifies copies of existing dataset data to create synthetic data, and hence increases the overall amount of data at hand. It helps regulate overfitting while training a model. While testing, application of data augmentation allows for us to have a larger amount of data to evaluate your model. Hence, we can apply data augmentation to testing set and see how the model behaves on such images with respect to performance and IoU score.

**CHAPTER 4**

**SIMULATION, RESULTS AND DISCUSSION**

**4.1 Dataset**

For the implementation we have made use of Cornell Grasping Dataset, which is the most used dataset set for benchmark results. It is a hand labelled dataset and includes individual images of real objects found in everyday environment as shown in Fig 4**.** The dataset comprises of 1035 RGB-D images of real objects and 5110 positive grasps and 2909 negative grasps. However, as per our objectives we have only considered positively labeled grasps from the Cornell dataset during training.

|  |  |  |
| --- | --- | --- |
| Computer Mouse | Black Marker | Battery |
| Purple Block | Screwdriver | Bowl |

Figure 4 Images from Cornell Dataset

**4.2 Training Results**

The Cornell grasping dataset was trained using the GR-ConvNet model and the following IoU scores and losses were recorded. The trained model was then saved and epoch 31 was recorded to have the highest IoU score

Table 1 Training Results

|  |  |  |
| --- | --- | --- |
| **Epoch** | **Loss** | **IoU Score** |
| 00 | 0.0841 | 0.471910 (47.1%) |
| 31 | 0.0600 | 0.966292 (96.6%) |
| 49 | 0.0704 | 0.865169 (86.5%) |

Adam Optimizer, mini batch SGD technique and standard back propagation were used to train the GR-ConvNet model. The learning rate for the same was set to be 10−3 along with a mini-batch size equivalent to 8. The model ran for 50 Epochs and the best IoU Score was obtained at Epoch 31, where an IoU Score of 0.966 was observed as per Table 1.

**4.3 Testing Results**

The GR-ConvNet model was tested on a custom dataset of 18 images as given in Table 2, both on CPU and GPU. This custom dataset was made up of 6 different objects found in the lab. The model was able to accurately detect positive grasping poses for the unseen objects with an accuracy rate of 97%.

Table 2 Sample images from custom dataset for testing on GR-ConvNet model

|  |  |  |
| --- | --- | --- |
| Plier\_1 | Screwdriver\_1 | Spanner |
| Wooden Block | Blue Marker | Plier\_2 |

Table 3 Evaluation Results

|  |  |  |
| --- | --- | --- |
|  | CPU | GPU |
| Dataset | Cornell | Cornell |
| Validation Size | 89 | 89 |
| Evaluation Model | epoch\_31 | epoch\_31 |
| Average evaluation time per image (in seconds) | 1.396 s  With data augmentation - 0.887 s | 0.086 s |
| IOU (Intersection over Union) Results | 0.966 (96.6%)  With data augmentation-  0.943 (94.3%) | 0.966 (96.6%) |

The proposed GR-ConvNet was evaluated and tested on NVIDIA Quadro K2200 GPU and CPU Table 3 shows that the model was tested on an unseen RBG image and obtained a state-of -the -art accuracy of close to 97%. Furthermore, with data augmentation, the IoU results dropped to 94% accuracy, however it showed that the model with data augmentation showed better performance in GPU.

Table 4 Real-Time Grasp Pose of Image Obtained

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RGB Image** | **Grasp Quality** | **Angle** | **Width** | **Grasp** |
|  |  |  |  |  |
|  |  |  |  |  |

Table 4 displays the visual representation of the grasp quality score Q, at every point of the object image. This value Q is indicated by a score between 0 and 1, where 1 indicates the greatest chance of grasp success and 0 represents grasping failure. The results also display the angle for grasping required at each point of the object for it to be grasped. This is represented by angular rotation in a value range of e [ −π/2, π/2]. These three components are used to infer the grasp pose. It also includes the output, which is the rectangle grasp pose that is projected onto the RGB image. The model can be tested onto RGB-D images as well, but we have tested it onto an RBG image, a uni-modal input.

**4.4 Performance Analysis**

Table 5 Confidence Threshold vs IoU Measure of proposed GR-ConvNet model

|  |  |
| --- | --- |
| **Confidence Threshold** | **IoU Measure** |
| 0.1 | 0.996292 |
| 0.15 | 0.996292 |
| 0.25 | 0.996292 |
| 0.5 | 0.685395 |
| 0.9 | 0 |
| 1 | 0 |

Figure 5 Confidence Threshold vs IoU Measure

Table 5 and Figure 5 represents the confidence threshold versus IoU measure analysis. The confidence score threshold allows for the filter of false positives and ensures that a predicted bounding box has a certain minimum score. An IoU of 1 means your predicted bounding box perfectly matches the ground truth box, 0 means no part of your predicted bounding box overlaps with ground truth box. Hence, we can clearly observe that the highest IoU measure is observed when the confidence threshold is less than 0.25. Hence, why we set 0.25 as the optimal confidence threshold value, for most accurate predictions.

**CHAPTER 5**

**SUMMARY AND CONCLUSION**

**5.1 Conclusion**

Over the course of this research, we have delved into previous related works and have formulated an understanding on the recent possible methods and algorithms that perform vision based robotic grasp detection and how comparing them against each other would result in concluding which algorithm gives the most efficiency with accuracy.

We started our experimental analysis by implementing the GR-ConvNet model and VSCode by making use of Cornell Dataset to its size and simple complexity. The model was trained for 50 epochs and tested on the 31st epoch as it gave 97% IOU result. Then the model was used on both, an offline image and real-time image of an object to generate the grasp rectangles for the object in the image. The model successfully inputs an uploaded image and predicts the grasping coordinates for the same as well as runs the same in real-time for RGB, RGB-D and depth images. GR-ConvNet is an efficient model that predicts positive grasps for objects with an IoU score of 0.966, that is, close to 97%.

**5.2 Future Scope of Work**

Vision- based robotic grasp can be extended to image recognition where instead of just predicting the grasp rectangles, the model also classified objects into categories and hence sort them accordingly. This would be useful in many industries like the mining industry, retail industry, manufacturing industry etc. Furthermore, we can also implement successful prediction of grasping areas on an object using suction or vacuum grasps instead of the traditional grippers. Algorithms like  [Lightweight Convolutional Neural Network with Gaussian-based Grasping Representation, ResNet50 can be used for efficient grasping robotic grasping detection](https://paperswithcode.com/paper/lightweight-convolutional-neural-network-with).

# REFERENCES

|  |  |
| --- | --- |
| [1] | K. B. R. K. W. Kleeberger, "A Survey on Learning-Based Robotic Grasping," *Curr Robot Rep,* 2020,vol. 1, pp. 239–249, doi:10.1007/s43154020000216. |
| [2] | A. Zeng et al., "Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3750-3757, doi: 10.1109/ICRA.2018.8461044. |
| [3] | P. S. a. G. C. N. P. Goyal, "Regression based robotic grasp detection using Deep learning and Autoencoders," *2020 IEEE 4th Conference on Information & Communication Technology (CICT),* 2020,pp. 1-6, doi:[10.1109/CICT51604.2020.9312104](http://dx.doi.org/10.1109/CICT51604.2020.9312104). |
| [4] | D. Morrinson, P. Corke and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, 2020, vol. 39, no. 2–3, pp. 183–201, doi:[10.1177/0278364919859066](https://doi.org/10.1177/0278364919859066).019. |
| [5] | C. P. L. J. Morrison D, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research,* 2020, vol. 39, pp. 183-201, doi: [10.1177/0278364919859066](https://doi.org/10.1177%2F0278364919859066). |
| [6] | M. L. C. G. G. M. K. S. P. A. S. S. M. Schwarz, "Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing," *IEEE International Conference on Robotics and Automation (ICRA),* 2018, pp. 3347–3354, 2018, doi: 10.1109/ICRA.2018.8461195. |
| [7] | K. K. L. A. H. M. El-Shamouty M, "Simulation-driven machine learning for robotics and automation," *Technisches Messen,* 2019, vol. 86, no. 11, p. 673. |
| [8] | S. Kumra, S. Joshi and F. Sahin, "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020,pp. 9626-9633, doi: 10.1109/IROS45743.2020.9340777. |
| [9] | F. -J. Chu, R. Xu and P. A. Vela, "Real-World Multiobject, Multigrasp Detection," *IEEE Robotics and Automation Letters*, 2018, vol. 3, no. 4, pp. 3355-3362, doi: 10.1109/LRA.2018.2852777. |
| [10] | P. C. a. J. L. Douglas Morrison, "Closing the Loop for Robotic Grasping:A Real-time, Generative Grasp Synthesis Approach", *The International Journal of Robotics Research*, 2020, vol. 39, no. 2–3, pp. 183–201, doi:[10.1177/0278364919859066](https://doi.org/10.1177/0278364919859066). |
| [11] | J. Yu, K. Weng, G. Liang and G. Xie, "A vision-based robotic grasping system using deep learning for 3D object recognition and pose estimation," *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2013, pp. 1175-1180, doi: 10.1109/ROBIO.2013.6739623. |
| [12] | Bergamini L., Sposato M., Pellicciari M., Peruzzini M., Calderara S., Schmidt J, "Deep learning-based method for vision-guided robotic grasping of unknown objects", *Advanced Engineering Informatics*, 2020, vol. 44, pp. 1, doi: 10.1016/j.aei.2020.101052. |

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application

Description automatically generated