Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Localization and Grasping of Objects by a Robot Arm Covered with Sensitive Skin

Karel Bartůněk

Supervisor: doc. Mgr. Matěj Hoffmann, Ph.D.
Supervisor–specialist: Ing. Lukáš Rustler
Field of study: Cybernetics and Robotics
May 2025

Acknowledgements

I would like to thank my supervisors, Mgr. Mat¥j Ho mann, Ph.D and Ing. Luká² Rustler for their guidance, advice, and for maintaining the student laboratory that I was allowed to utilize for the purposes of this work.

Abstract

Localization and grasping of objects by robots is typically performed while relying on some form of visual sensors. Haptic feedback from contacts with the environment is only secondary if present at all. In this work, an extreme case of localization and grasping in complete absence of visual input is explored. One such case is a robot that relies on a whole-body electronic skin. I designed a system for localization and grasping of objects for this robot in an environment with many target objects. I evaluated this system using several handmade objects on a table. The system achieved a grasp success rate of 81.67% and average execution time of 207s per object. A variant that does not use haptic feedback during grasping has shown a success rate of 64.14% while decreasing the execution time by 6.5% to 194s The platform and system design places severe limitations on general application, both in object type and pose. Future work should focus on the design of systems that predict object model using more information. A large number of contacts per object, a more detailed tactile sensor or both could be used for this purpose.

Keywords: tactile localization, electronic skin, collaborative robot, whole-body tactile feedback

Supervisor: doc. Mgr. Mat¥j Ho mann, Ph.D.

Abstrakt

Lokalizace a uchopování p°edm¥t· roboty obvykle závisí na uºití obrazových £idel. Hmatová zp¥tná vazba (HZV) s prost°edím se uºívá druhotn¥, nebo se nepouºívá v·bec. V této práci je zkoumán krajní p°ípad, kdy nejsou dostupná jakákoliv obrazová data. Jedním takovým p°ípadem je robot s elektronickou k·ºí po celém svém t¥le. Navrhl jsem soustavu pro lokalizaci a uchopení p°edm¥t· pro tohoto robota v prost°edí, kde se nachází vícero cílových objekt·. Testoval jsem tento systém v prost°edí s n¥kolika ru£n¥ vyrobenými p°edm¥ty na stole. Systém dosahoval úsp¥²nosti 81.67%, p°i£emº doba úchytu £inila pr·m¥rn¥ 207s Dále byla otestována varianta systému, která nepouºívá HZV p°i samotném úchytu. Tato varianta dosahovala úsp¥²nosti 64.14% se sníºením doby úchytu o 6.5% na 194s Platforma a návrh systému kladou omezení na druh p°edm¥tu a jejich umíst¥ní, která je robot schopen uchopit. Budoucí práce by se m¥la zam¥°it na stavbu robustn¥j²ích model· p°edm¥t·. Data z velkého mnoºství kontakt·, z p°esn¥j²ího dotykového senzoru, nebo oboje mohou být uºita k tomuto ú£elu.

Klí£ová slova: hmatová lokalizace, elektronická k·ºe, kolaborativní robot, celot¥lová hmatová zp¥tná vazba

P°eklad názvu: Lokalizace a uchopení p°edm¥t· robotickou rukou pokrytou citlivou k·ºí

Contents

Chapter 1

Introduction

In humans, the ability to grasp objects using visual input is rooted in receiving structured visual input early in development [1]. Innately blind individuals (also known as congenitally blind) have to rely on other forms of input, such as tactile feedback. Visually sighted individuals might not be able to rely on visual input in certain scenarios, especially when low light, deep fog, or occlusion is introduced. Stone et al. [2] show that hand use and hand preference in innately blind people are similar to that of visually blind people and are more influenced by factors such as the scale and shape of objects. They also show that the distribution in the direction of grasps and hand preference is different in the innately blind to an age and gender-matched group of blindfolded sighted individuals but is close to the group of visually unimpaired. Piller et al. [1] suggest that sighted individuals build structured models of object weight and plan correct grasping movements based on visually obtained information. They observed the grasping ability of innately blind individuals with restored sight. Even years after sight restoration, the subjects showed difficulties in grasping novel objects. Remarkably, they significantly improved the ability to predict the features of the object using sight. These works suggest a difference in grasping habits and approaches between blindfolded sighted individuals and the innately blind.

Robotics realize grasping as pick-and-place tasks. The robot's ability to find and grasp objects is crucial for applications such as bin picking [3]. The robotic arms used for pick-and-place could be a standalone unit or part of a more complex assembly, e.g., a humanoid [4]. In traditional pick-and-place operations, the operator is required to teach the robot the positions of all objects to be picked and placed. This approach suffers from faults when the pose of the objects changes due to an error. Vision-based intelligent robotic localization and grasping systems offer the ability to perform pick-and-place tasks on arbitrary objects in space without additional operator intervention or teaching [4]. Vision-based object localization and pose estimation is performed on a variety of objects, but adversarial materials, such as transparent plastics, glass, or specific metals, require specialized (and often complex) approaches to achieve satisfactory results [5]. The robot's various poses might occlude the workspace, preventing any effective visual feedback. In addition, vision-based methods depend on the background and lighting of the scene [6]. Some robots might be equipped with haptic sensors, sensitive skin, or force sensors that allow them to detect direct contact with the environment and retrieve the 6D pose of the contact. These tactile sensors can be used for active online exploration of unknown objects and to determine their shapes and contours [7]. In a review conducted by [6], the idea of using these sensors as replacements for vision in intelligent localization and grasping was shown to be viable using a variety of small-scale tactile sensors attached to a collaborative robot (cobot) platform or on humanoids.

Using human skin as an analogy, making the tactile sensor area as large as possible could result in better results. The increased contact area should greatly speed up object localization by allowing it to explore more space at any given time. But does it? In an attempt to explore this hypothesis, a similar task is performed on a robot covered in pressure-sensitive electronic skin. I propose a closed-loop system for tactile-based localization and grasping that uses only a whole-body arti cial skin and a 6-axis force/torque (F/T) sensor. The system locates every suitable object in the workspace, grasps it, and moves it to a predetermined bin. The objects for the task are manufactured and chosen to t speci c criteria. The platform and technologies used are thoroughly described in Chapter 3. The problem of localization and grasping individual objects is broken down into separated subtasks in Chapter 4, and for each subtask one or more strategies are proposed in Section 4.1, Section 4.2 and Section 4.3. Each strategy is described and analyzed on the basis of its operation principle, theoretical limitations, and theoretical worst-case performance. The performance of the system is tested on a physical cobot platform using multiple manufactured objects in Chapter 6. For subtasks in which multiple strategies are developed, their performance (success rate, speed of execution) is compared. Interesting ideas and approaches that were proposed or even implemented but are not part of the system for various reasons are described in Chapter 5. Finally, the bene ts and limitations of the system are discussed in Chapter 7

Chapter 2

Related Work

2.1 Vision-based grasping

Survey conducted by G. Du et al. [4] provides a brief taxonomy of tasks required for an intelligent vision-based robotic grasp detection system. Various methods of 2D and 3D object localization are described. These methods are categorized as classificatory or not. Non-classificatory approaches often involve fitting input data to primitive 2D/3D shapes or using saliency maps[1]. A 3D point cloud is a very commonly targeted input of object localization for both shape fitting and saliency generation [4].

Object detection targets workpieces of an initially chosen class, for which it yields its location in the input data. The technique of creating bounding boxes has become popular due to inaccuracies while generating salience regions, or when the shape of the workpiece differs between samples[2]. Methods of object detection target 2D input (typically 2D images) and 3D input (typically RGB-D data). Two-stage methods[3] are popular in both classical classification using Support Vector Machine (SVM) and deep learning. They are often implemented with the sliding-window strategy and labeled data [9] [4].

Object pose estimation is the identification and verification of the position and orientation of the workpiece [10]. The expected result is a 6D pose. 2D pose might be sufficient in cases where the workpiece is constrained to a planar workspace (such as a worktable). Some examples of common approaches include correspondence finding, template matching, and feature extraction. Correspondence-based methods are based on finding correspondences on the features of objects. Template matching algorithms search Ground Truth object templates and attempt to choose the closest one in terms of object pose [4]. Feature extraction often relies on machine learning to choose relevant features. They are effective on objects with partial occlusion or a lack of texture. A popular feature extraction algorithm is the random forest. The random forest algorithm consists of several decision trees (classifiers) each voting on randomly selected hypotheses. The hypothesis with the highest number of votes is then chosen [10].

[1] Saliency based coding separates the input image into regions of interest (ROI) containing the object of interest. In the context of machine learning, a saliency map is constructed as an output of a machine learning model such as Convolutional Neural Network (CNN) and used to generate segmentation masks [8]

[2] This occurs when classes are not rigidly defined or do not have strictly defined shapes, such as when classifying animal species or broad groups of objects.

[3] Two-stage methods of classification consist of object proposal generation (guessing plausible locations) and conduct classification (verifying location guesses) [4].

Grasp estimation is responsible for determining a viable gripper pose in relation to the identified and localized workpiece [11]. Grasp contact points are determined using methods similar to those for object pose estimation. All such methods must be adjusted to a specific manipulator based on its size and capabilities. Grasp poses can be 6D or 2D depending on the workspace and the specifics of the manipulator [4]. 2D grasp poses might prove problematic for objects in certain poses. One such case is the use of vertical grasp poses to grasp horizontally placed flat objects [11].

2.2 Tactile perception in selected tasks

Tactile input has already been shown to be useful in robotic regrasping. Regrasping modifies the gripper pose while grasping the object currently held, usually its rotation [12] [13]. Regrasping is often performed to improve grasp stability or in industrial applications, where it replaces a sequence of arms performing individual rotations [14]. Hogan et al. [12] introduces a tactile-based grasp adjustment system to increase grasp stability based on a grasp adjustment policy. The policy relies on a special grasp quality metric consisting of a pre-trained ResNet network. Calandra et al. [13] proposes a similar policy using a grasp quality metric derived from a deep residual network.

Abraham et al. [15] uses a robot with only a binary collision probe to explore and estimate the shapes of flat, rigid objects on a plane. The choice of object class allowed them to search for a decision boundary between shapes with a Gaussian kernel statistical model. Sun et al. [16] devises an algorithm for the identification of objects based on haptic features captured by a colliding end effector. They constructed a test platform consisting of a 2-joint robot and a silicone hemisphere containing a 6-axis force torque sensor. The platform is programmed to travel a path on the workpiece while retaining a constant force. This creates deviations in friction and traveled path as the robot is forced to adjust to the shape and material of the workpiece. The workpiece is then chosen as the one whose features are closest to the recorded features. Jamali et al. [17] have created a sampling strategy to map the object surface using an iCub fingertip for tactile data. The plane of the workspace is linearly scanned and the contact height and point cloud of the haptic sensor are collected in a grid of points with a grid spacing of 5 mm. Upon initial contact, the next exploration position is chosen near the location of the guessed workpiece to increase the probability of determining an object model.

Kaboli et al. [18] proposes a framework for the identification of objects in an unstructured workplace. They propose a regression algorithm that minimizes uncertainty in a contact point cloud obtained by linear scans in the planar workspace. The data obtained from the point cloud are then mean-shift clustered[4], and the localization is complete with the construction of oriented bounding boxes on the clusters. They then prepare several experiments to determine the stiffness, texture, and center of mass of the workpiece. These experiments are then used for active object recognition. Kaboli et al. also built a framework for multimodal tactile object discrimination. They use a combination of a force sensor,

[4] Mean-shift approaches in point cloud clustering consist of a point search whose criteria are not increasing the movement of the cluster's geometric center. A point is added to a cluster when the distance between the geometric centers with and without the point is less than a given threshold [19].

an accelerometer, a proximity sensor, and a temperature sensor to learn object properties. They define active touch as a system that combines the features of exploration actions (static contact, sliding, and pressing with the end effector of the arm) with learnable observation models. They localize objects using a pre-touch procedure that combines hexagonal grid exploration with a radar-like object search. Recursive Bayesian estimation is used to cyclically choose the most probable exploration pose until all objects on the workspace are found [20].

2.3 Contact-based grasping

The above mentioned work on tactile perception and its use displays the potential of haptic feedback and tactile sensors for manipulation with workpieces. Robots with tactile sensors seem capable of determining the location of an object, its shape, and pose. Robots could also benefit from haptic feedback to find a stable grasp of the workpiece using methods such as regrasping.

Felip et al. [21] presents a complete system for grasping novel objects based on haptic and force-torque sensors using a 25 DOF humanoid robot torso. They perform object localization using a linear scanning motion of the scanning area. A specific configuration of the end effector is used to detect hand-crafted combinations of drag forces. Object pose estimation is not performed and is instead a part of grasp estimation. Grasp estimation uses the sliding grasp approach. Optimized metrics are changes in contact forces when the fingers apply a constant force to the workpiece. Murali et al. [22] developed a system combining contact-based object localization and tactile regrasping. They perform object localization using a contact sensing algorithm utilizing a particle filter approach. A motion model and sensor model are devised for the given scenario. An encoder-decoder model is chosen for the extraction of tactile characteristics and the estimation of grasp quality. A classifier model is then used for policy estimation using the aforementioned grasp quality metrics. These models are then trained on a dataset of objects in various poses.

The review conducted by Li et al. [6] describes trends in robotic grasping using tactile perception. The review introduces the term "tactile-visual fusion methods" that describes methods that use tactile data as a feedback mechanism for visual-based grasping. The review claims that tactile data improve traditional object pose and grasp estimation by providing data on grasp forces and properties of workpieces occluded in a typical visual grasping scenario. The review separates approaches to the estimation of the pose of the tactile target into model matching, filter-based[5], optimization, machine learning classification, and tactile-visual fusion. The review shows uses of tactile sensors for surface characteristic determination using feature extraction on a dataset of surfaces. This extraction is performed either using traditional optimization of sparse Gaussian kernels or machine learning. The review claims that machine learning methods have greater generalization and adaptability from a single data set compared to traditional methods.

[5] The particle filter is a very common approach for increasing the accuracy of data obtained from online interaction of the robot with objects. These methods often rely on specific scenarios and environments [6].

2.4 Whole-body tactile skin

The majority of used tactile skins in tactile perception are tactically placed on the robot hand either as a part of the end e ector or as a part of a combination of sensors [23] [6]. Visual-tactile sensors are particularly poorly suited for full-body applications as the required large coverage area would require the use of a large number of individual sensors with a small activation area each [24]. This trend is especially visible in robotic grasping. The vast majority of object localization and grasping research focuses on robotic hands with tactile sensors or small patches of arti cial skin [6].

In contrast, only a small amount of research focuses on arti cial skins that span larger areas of the robot. Jianget al. [25] proposes a hierarchical safety framework and applies it to a collaborative teleoperation system. The system consists of a danger avoidance planning system, a robot movement restriction system that adapts to danger, and a motion retraction triggered by tactile stimulus. For the purposes of tactile perception, they construct a modular electronic skin consisting of ex-printed circuit boards with a force-sensing resistor array. Each module uses a wireless interface to communicate with the main controller to transmit contact events. Albini et al. [26] propose a control system for navigating a robot arm partially covered in an arti cial robot mesh skin through an unstructured environment with obstacles. They construct a closed-loop controller that minimizes Cartesian and angle distance from the target pose and minimizes contact forces perceived by the skin. The skin provides 3495 taxels, each calibrated for normal force readings. The controller does not address reaching the global target distance minima due to its reactive nature.

Chapter 3

Materials and Methods

3.1 Used materials and their speci cations

Presented additions to the existing corpus depend on the exploitation of particular hardware and software for the purposes of localization and grasping of objects in a workspace. The goal is to allow the manipulator to locate and grasp objects in unknown poses within speci ed limits. This section aims to describe the materials used for this thesis. It also describes what objects will be located and grasped. The programming environment, the methodology behind certain design decisions, and certain limits on the properties of objects and their poses are described.

3.1.1 Robotic manipulator

The system is designed for and all experiments were performed on the Universal Robots UR10e standalone manipulator arm equipped with a parallel gripper OnRobot RG6 as its end e ector, shown in Figure 3.1. The robot has 6 degrees of freedom, the maximum TCP speed of 4 m/s with each joint rotating at a maximum of 120 s^{1} [27]. The gripper is designed for safe grasping of objects without exceeding the designed grasp strength [28]. Parts of the manipulator body are covered with industrial-grade arti cial skin AIRSKIN —. The ange contains a 6-axis force/torque (F/T) sensor. The F/T sensor has a range of 100 N for force and 10 N m for torque on each axis (three for force and three for torque) with an accuracy of 5:5 N for force and 0:5 N m for torque [27]. The F/T sensor is sampled at a rate of f_s = 500 Hz.

3.1.2 Arti cial skin AIRSKIN —

AIRSKIN — is a commercial solution to improve safety with collision detection in industrial robotic manipulators. The manufacturer advertises the system as full-body protection for both collaborative and industrial robots. According to the manufacturer, AIRSKIN — reduces the required oor space occupied by a robot by o ering a soft and safe human collision detection. The system consists of several AIRSKIN— soft and pressure-sensitive safety sensors [29]. Sensor construction is based on soft chambers lled with pressurized air. The electronics of each sensor consist of a pressure sensor, a small air compressor, a thermometer, and control and communication circuitry [29]. For each sensor, its internal pressure is interpreted as a single taxel. These units are in standard conditions connected to the AIRSKIN — Connection Box that exposes push-in contacts for the manipulator or a suitable PLC (Programmable Logic Controller) [29]. The sampling frequency of the skin pressure reading was measured at 35 Hz for a value update of all pads[1].

In this installation, the UR10 S2W AIRSKIN — kit is connected to a proprietary HAT [2] for the Raspberry Pi 3 Model B+. The HAT allows the operator to retrieve unprocessed pressure values from each unit. From now on, each pad will be referred to by its index, starting at 0 (the pad covering the shoulder joint) and increasing towards the end effector. The numbering of each pad is also described in Figure 3.1.

Figure 3.1: Robotic manipulator with artificial skin, the numbers represent the ordering of pads that the artificial skin consists of, starting with 0. This order is the same as the one used for pressure data packets. The pad that would be present on the end effector has been removed in this installation.

3.2 Objects for grasping

For the purposes of grasping, these non-trivial assumptions are placed about objects and their pose on the workspace:

- Objects are not stacked on each other.

[1] Each pad is queried in sequence by the controller, the measured sampling frequency is for two independent updates of the same pad.

[2] Expansion hardware for the Raspberry Pi platform conforming to electrical and mechanical constraints of the platform [30].