Soft robotic manipulation in agrotechnical procedures based on machine and deep learning

Polić, Marsela

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Marsela Polić

SOFT ROBOTIC MANIPULATION IN AGROTECHNICAL PROCEDURES BASED ON MACHINE AND DEEP LEARNING

DOCTORAL THESIS

Supervisor: Associate Professor Matko Orsag, PhD

Zagreb, 2022

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Marsela Polić

PODATNA ROBOTSKA MANIPULACIJA U AGROTEHNIČKIM ZAHVATIMA ZASNOVANA NA STROJNOM I DUBOKOM UČENJU

DOKTORSKI RAD

Mentor: izv. prof. dr. sc. Matko Orsag

Zagreb, 2022.

Doctoral thesis was written at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Control and Computer Engineering.

Supervisor: Associate Professor Matko Orsag, PhD

Thesis contains 104 pages

Thesis no.:

ABOUT THE SUPERVISOR

MATKO ORSAG is an associate professor at the University of Zagreb Faculty of Electrical Engineering and Computing (UNIZG-FER). He has been involved as a researcher in various projects financed by the government and industry. In 2011/2012, he worked as a visiting researcher at the Drexel University, Philadelphia, USA as a recipient of the Fulbright exchange grant.

As a researcher, he participated in national and international research projects in the field of robotics, control, and automation. Currently, he is working as the principal investigator of the Croatian Scientific Foundation project Specularia - Structured Ecological CULtivation with Autonomous Robots In Agriculture. He co-authored over 60 scientific and professional journal and conference papers, a book chapter, and a monography in the area of aerial robotics.

He serves as a reviewer in journals and conferences, and as an editor and guest editor of several journals (Automatika, JINT, etc.). He is a member of IEEE, currently serving as IEEE Robotics and Automation Society Chapter Chair of IEEE Croatia Section, and a member of the Aerial Robotics Topic Group. He is a member of Scientific and Professional Council of Innovation Centre Nikola Tesla (ICENT). In 2019 he received a Croatian Academy of Engineering young scientist award „Vera Johanides".

O MENTORU

MATKO ORSAG izvanredni je profesor na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Sudjelovao je u različitim istraživačkim projektima financiranim od strane vlade i industrije. Godine 2011./2012. radio je kao gostujući istraživač na Sveučilištu Drexel, Philadelphia, SAD, kao stipendist Fulbright programa razmjene. Kao istraživač sudjelovao je u domaćim i međunarodnim istraživačkim projektima iz područja robotike, upravljanja i automatizacije. Trenutno je glavni istraživač na projektu Hrvatske zaklade za znanost, Specularia - Strukturiran ekološki uzgoj primjenom autonomnih robota u staklenicima. Koautor je više od 60 znanstvenih i stručnih radova u časopisima i skupovima, poglavlja knjige i monografije iz područja zračne robotike.

Angažiran je kao recenzent u časopisima i na konferencijama, te kao urednik i gostujući urednik nekoliko časopisa (Automatika, JINT i dr.). Član je IEEE-a, trenutno služi kao predsjedavajući ogranka Društva IEEE Robotics and Automation Society Sekcije IEEE Hrvatska i član tematske grupe za zračnu robotiku. Član je Znanstvenog i stručnog vijeća Inovacijskog centra Nikola Tesla (ICENT). Godine 2019. dobio je nagradu za mladog znanstvenika Hrvatske tehničke akademije „Vera Johanides".

ACKNOWLEDGEMENTS

First and foremost, my biggest thanks to my supervisor Asst. prof. Matko Orsag for taking me as one of his first students, for his guidance in research, all the time and patience he has dedicated to making me a researcher I am, all the invaluable advice, suggestions, ideas and solutions he has happily shared with me. More importantly, I want to thank him the for the paternally endless patience, support and help he keeps giving. My gratitude also goes to Prof. Stjepan Bogdan who taught me so much, but most importantly the art of acquiescence. A loving thanks to Prof. Zdenko Kovacic for founding the lab and enjoying his status as a father figure to all of us.

Most of my work results would not exist without the efforts and hard work of my dear LARICS colleagues, especially Tomislav, Ivan, Marko, Bruno, Damjan, Juraj, Jelena, Dario and Ivo. Thanks to prof. Lepora for accepting me into the team and the fruitful collaboration we had.

I will be forever grateful to have met, befriended, traveled, and had fun with my (LARICS) friends Barbara, Marko, Karlo, Antun, Frano, Goran, who made my life. Working with them has also been ok (mostly). Immense thanks to all of my friends outside the lab, without mentioning any names in fear of missing someone out.

Forever thanks to my parents, brothers, Marko, Medo and those looking after us for all the love.

Zagreb, March 11 2022.

ABSTRACT

Robotising organic agriculture has recently been growing in an attempt to alleviate the cost of sustainable large scale food production under decreased use of pesticides. One of the main goals is to reduce human labour input by replacing it with a robotic workforce. The focus of this thesis is on a collaborative manipulator arm capable of conducting various plant treatments autonomously. In order to achieve sufficient versatility with respect to different crops and procedures, existing robotic solutions need to overcome their limited capabilities in terms of perception, motion planning, or dexterous manipulation. This thesis aims to explore and expand the state of the art robot capabilities in these domains. Replacing human labour requires robot perception to imitate human capabilities, even though here it is not possible to rely on millennia of evolutionary progress. Imitation strategies are deployed instead, both in the inference machine design through the concept of neural networks, and on the level of learning strategies through concepts such as transfer and sim2real learning. As shown in this thesis through the development of a perception module for robotic manipulation in indoor farming conditions, these approaches manage to imitate and occasionally even outperform human capabilities. Furthermore, this thesis demonstrates how manipulation modules that use these perception results as input in planning phases can execute safe and compliant manipulation procedures with limited feedback. Finally, the thesis explores and extends the capabilities of novel tactile sensing technologies, and their potential in enriching contact feedback during manipulation of fragile objects.

KEYWORDS: Convolutional neural networks, transfer learning, sim2real, rgb-d, compliant manipulation, collaborative robot, impedance control, agricultural robotics, soft sensing, optical based tactile sensing

SAŽETAK

Podatna robotska manipulacija u agrotehničkim zahvatima zasnovana na strojnom i dubokom učenju

Moderna poljoprivreda rezultat je brojnih prilagodbi koje su omogućile masovnu proizvodnju hrane. Uz tešku mehanizaciju, jedna od najznačajnijih takvih prilagodbi jest upotreba pesticida, usprkos višestruko utvrđenim činjenicama o negativnom utjecaju na okoliš i na sav živi svijet, uključujući biljke, vodene i kopnene divlje životinje, kukce oprašivače, ali i ljude. Održiva alternativa upotrebi pesticida u proizvodnji hrane je organska poljoprivreda, u kojoj je glavno oružje u borbi protiv štetnika i bolesti mehanički rad. Već je sada teško zadovoljiti zahtjeve za radnom snagom potrebne za održivu proizvodnju hrane velikih razmjera, a u kombinaciji s rastućom potražnjom za hranom diljem svijeta zbog rasta ljudske populacije ovaj problem postaje sve izraženiji. Doda li se tome činjenica da ovakva vrsta mehaničkog rada odgovara opisu dosadnih, ponavljajućih i opasnih poslova, jasno je da je idealan kandidat za zamjenu ljudskog rada robotskim. Ako je suditi prema rezultatima uvođenja robotizacije u druga područja ljudskog rada, kao primjerice industrijsku proizvodnju, korištenjem robota istovremeno se mogu osigurati i veća učinkovitost, bolja kvaliteta, i lakša skalabilnost, uz primarni cilj i benefit rasterećenja ljudske radne snage.

Robotizacija poljoprivrede nije novi koncept, već velika istraživačka tema u kojoj se ugrubo mogu prepoznati dva pristupa. Stariji pristup, već uvelike primjenjen u komercijalnoj proizvodnji, fokusira se na razvoj strojeva za tradicionalni uzgoj na otvorenom. U ovakvim se primjenama teži postizanju što višeg stupnja autonomije velikih strojeva koji su inicijalno projektirani za ljudskog operatera. Ovakvi strojevi uglavnom se optimiraju za robusnost i efikasnost na velikoj skali, to jest cilj je obrada što većih površina u što kraćem vremenu. Pritom postizanje visoke preciznosti nije jedan od primarnih ciljeva.

Noviji smjer istraživanja i razvoja poljoprivredne robotike koncentrira se na sustave za uzgoj u zatvorenim prostorima, kao što su staklenici i plastenici. Zahvaljujući činjenici da je u ovakvoj vrsti uzgoja radno okruženje u određenoj mjeri strukturirano, ovaj se pristup često smatra prikladnijim za implementaciju autonomnih robotskih rješenja izvorno dizajniranih za industrijska okruženja. Drugim riječima, umjesto pretvaranja snažnih strojeva u autonomne robote, ovdje se postojeće sustave autonomnih robota, razvijene za industrijske zadatke, prilagođava i programira za provođenje poljoprivrednih aktivnosti. Nadogradnjom postojećih rješenja razvoj se značajno ubrzava, a dobivena rješenja su efikasnija i pouzdanija. U ovoj vrsti poljoprivrednog uzgoja, umjesto kvantitete fokus je na preciznosti obavljanja agrotehničkih zadataka, budući da se autonomnim strojevima emulira ljudski mehanički rad i biljke se obrađuju pojedinačno.

Istraživanje opisano u ovom radu, provedeno u sklopu projekta SpECULARIA, bavi se problemima iz domene poljoprivredne robotike. Cilj projekta je razvoj heterogenog robotskog sustava za autonomni uzgoj biljaka u staklenicima. Sustav se sastoji od bespilotne letjelice, mobilnog robota i robotskog manipulatora. Svaki robot u ovom heterogenom sustavu ima ograničen specifičan skup sposobnosti, ali kada rade zajedno, mogu se primijeniti za postizanje raznolikih ciljeva u strukturiranom okruženju kao što je zatvorena organska farma. Uloga letjelice je nadzor plantaže i identifikacija biljaka kojima je potreban tretman. Opremljena je manipulatorom s više stupnjeva slobode koji nosi senzore za nadzor kao što su kamera i multispektralna kamera. Zahvaljujući manipulatoru omogućen je precizan nadzor uz let na sigurnoj udaljenosti, izvan područja u kojem bi utjecaj letjelice bio štetan za biljke. Budući da je uzgoj organiziran kontejnerski, drugi dio robotskog sustava čine mobilni roboti opremljeni mehanizmom za transport uzgojnih kontejnera između njihovih uzgojnih pozicija u stakleniku i stanice za obradu. Stanica za obradu je radni prostor robotskog manipulatora čiji je zadatak delikatno rukovanje biljkama. Manipulator obavlja razne aktivnosti biljne higijene i agrotehničke zahvate kao što su orezivanje, branje plodova, oprašivanje, kontrola uroda ili kontrola vlage u zemlji. Upravo ove aktivnosti i njihovo ostvarenje kroz preciznu senzoriku i podatnu aktuaciju robotskog manipulatora u fokusu su ovog istraživanja.

Kako bi se osiguralo pažljivo rukovanje s osjetljivim biljkama koje se ne bi smjele ozlijediti prilikom obrade, rješenja razvijena za robotski manipulator na nekoliko razina emuliraju ljudske mogućnosti percepcije i manipulacije. Pritom se koriste principi podatnog upravljanja, mekane robotike i senzorike, i umjetne inteligencije. Metode strojnog i dubokog učenja koriste se za izgradnju trodimenzionalnog semantičkog i funkcionalnog modela biljke. Takav model biljke može se koristiti za autonomnu identifikaciju, planiranje i provođenje potrebnih agrotehničkih zahvata. Za provedbu agrotehničkih zahvata poput branja plodova, kidanja zaperaka, mjerenja vlage i drugih sličnih aktivnosti razvijeni su algoritmi planiranja gibanja i podatnog upravljanja manipulatorom. Pod podatnim upravljanjem podrazumijevamo upravljanje pozicijom robotskog manipulatora tako da se pritom u obzir uzimaju mjerene dodirne sile. Mjerenje dodirnih sila može se dobiti iz različitih izvora, a jedan primjer je meki taktilni senzor s dizajnom i programskom podrškom izvedenom u skladu s principima mekane robotike.

Ova disertacija je organizirana u osam poglavlja. Prvo poglavlje disertacije daje uvod u temu kroz motivaciju i opis otvorenih pitanja u području istraživanja. Također se definiraju hipoteze istraživanja i znanstveni doprinosi disertacije. Drugo poglavlje daje pregled područja za različite domene kojima se disertacija bavi. Sljedeća tri poglavlja bave se glavnim doprinosima i rezultatima disertacije. Tako treće poglavlje razmatra vizualnu percepciju u poljoprivrednoj robotici, četvrto se bavi podatnim upravljanjem robotskim manipulatorima, a peto domenom taktilnih senzora. Šesto poglavlje iznosi zaključke donesene kroz rezultate disertacije. Sedmo poglavlje daje popis objavljenih radova koji čine disertaciju, dok osmo poglavlje opisuje doprinos autora na svakome od njih. Potom je izložen popis literature korištene u disertaciji te su priloženi radovi na kojima se disertacija zasniva. Disertacija je izrađena po skandinavskom modelu te ju čine tri časopisna i jedan konferencijski članak. Glavni doprinosi disertacije su izloženi i opisani u nastavku poglavlja.

#1 *Senzor dodira prilagođen za podatnu manipulaciju osjetljivih predmeta korištenjem umjetne inteligencije.*

Manipulacija krhkih objekata kao što su biljke ili razni osjetljivi predmeti iz svakodnevnog života zahtjeva visoku razinu spretnosti. Evolucijom su ljudi razvili šake kao vrlo kompleksne manipulacijske sustave, daleko složenije od konvencionalnih robotskih sustava, i prema svojim su mogućnostima organizirali životne i proizvodne aktivnosti. Pokušaj zamjene ljudske radne snage robotskom, makar i u dosadnim, ponavljajućim aktivnostima, stoga zahtjeva imitaciju ljudskih mogućnosti manipulacije i razine spretnosti. Pritom je ključan korak reprodukcija sposobnosti ljudskog osjetila dodira, kako bi robot prilikom gibanja i manipulacije raspolagao povratnom informacijom o dodiru. Ova povratna veza važna je ne samo u zadacima precizne manipulacije, već i za sigurnu interakciju s čovjekom ili rad u zajedničkom radnom prostoru s ljudskim operaterima.

U okviru ovog istraživanja dio znanstvenog doprinosa odnosi se na razvoj senzora dodira koji robotu pruža informaciju visoke razine kompleksnosti o ostvarnom kontaktu s okolinom. Među različitim tehnologijama koje se koriste u pokušajima razvoja adekvatnih senzora dodira, odabran je senzor iz obitelji TacTip. Ovi optički taktilni senzori dizajnirani su prema principima mekane robotike, prvenstveno zahvaljujući zadovoljavajućoj uravnoteženosti kvalitete i upotrebljivosti dobivenih senzorskih informacija s mehanički podatnom strukturom tijela. Podatno tijelo senzora poželjna je karakteristika jer poboljšava sigurnosne aspekte kontakta robota i okoline. Naime, podatno tijelo pasivno će adaptirati svoj oblik i time apsorbirati dio energije prilikom kontakta, za razliku od konvencionalnih, rigidnih senzora koji su dizajnirani za vrlo precizna mjerenja prilikom kontakta, nauštrb rizika od oštećenja tijela senzora ili predmeta s kojim je u interakciji. Kvaliteta informacije sa senzora ovisi o konkretnom manipulacijskom zadatku, ali općenito se može kvalitativno razmatrati u kontekstu mogućnosti postizanja željenog cilja, primjerice prepoznavanje ili eksploracija oblika objekta, procjena stabilnosti hvata, detekcija klizanja ili upravljanje dodirnom silom.

Odabrana senzorska tehnologija svojim sklopovskim dizajnom imitira ljudsko osjetilo dodira, a znanstveni doprinos ove disertacije tu imitaciju prenosi i na razinu ekstrakcije korisne informacije na algoritamskoj razini. Meko tijelo senzora inspirirano je jagodicama ljudskih prstiju, a sferična dodirna ploha se podatno prilagođava okolini prilikom ostvarivanja kontakta. Unutrašnjost mekane opne senzora obogaćena je trodimenzionalnim vizualnim oznakama koje, inspirirane Merkellovim stanicama u ljudskoj koži, prenose i pojačavaju signal o deformaciji površine tijela (kože ili opne senzora). Ta se deformacija snima ugrađenom kamerom, te se različitim metodama obrade iz ovih sirovih senzorskih podataka (slika kamere) mogu dobiti bogate informacije o ostvarenom kontaktu. Druga razina imitacije ljudskog zaključivanja i percepcije, ona na algoritamskoj razini, ostvaruje se upravo u ovom procesu pomoću metoda dubokog učenja.

Kao značajan znanstveni doprinos ove disertacije, konvolucijska neuralna mreža u arhitekturi autoenkodera istrenirana je za ekstrakciju nezavisnih značajki, čime se omogućuje postizanje značajnog smanjenja dimenzionalnosti podataka sa senzora. Nakon faze učenja, ulazni dio istrenirane mreže, enkoder, može se smatrati dijelom sustava za akviziciju podataka. Time efektivno taktilni senzor daje visoko kvalitetnu informaciju komprimiranu

u malom broju nezavisnih varijabli. Baratanje podacima niže dimenzije omogućuje brže treniranje percepcijskih modela zahvaljujući manjoj složenosti potrebnoj za podjelu ulaznog prostora. Dodatna je prednost što takvi modeli općenito ne iziskuju velike skupove podataka za učenje. Osim prilikom treninga, prednosti ulaznih podataka niske dimenzionalnosti su značajne i prilikom korištenja, zahvaljujući znatno smanjenim zahtjevima za resursima. Kompleksniji dio ekstrakcije podataka zajednički je različitim percepcijskim problemima i provodi se samo jednom, te se posljedično omogućuje korištenje više percepcijskih modula u paraleli bez gubitaka na brzini izvođenja.

#2 *Metoda izgradnje semantičkog trodimenzionalnog modela biljke pogodnog za planiranje robotske manipulacije u agrotehničkim zahvatima.*

Za autonomni uzgoj u staklenicima potrebno je identificirati funkcionalne i semantičke značajke biljaka, kako bi robotski sustav mogao prepoznati potrebu za intervencijom i valjano djelovati. Pritom je važno razlikovati različite dijelove biljke i uzgojnog sustava, a među njima i razlikovati njihove karakteristike. Uspješnost provođenja ovakvih autonomnih aktivnosti ovisi o različitim okolišnim čimbenicima. Jedan od važnih faktora koji narušava strukturiranost radnog okruženja u uzgojnom sustavu jesu promjenjivi uvjeti osvjetljenja, budući da mogu značajno promijeniti izgled značajki kod vizualne percepcije. Nadalje, objekti detekcije u fokusu ovog rada, odnosno organske strukture poput voća i povrća, općenito su podložni visokom stupnju varijabilnosti i zahtijevaju robusne metode percepcije koje mogu podnijeti takvu varijaciju. Zahvaljujući visokoj preciznosti u detekciji, robusnosti na varijabilnost i razvoju računalnih resursa, primat u ovakvim zadacima percepcije posljednjih godina imaju modeli dubokog učenja.

Problem percepcije u ovoj se disertaciji razmatra vezan uz primjene u robotičkoj manipulaciji nestrukturiranih osjetljivih objekata. Priroda problema i cilj manipulacije, odnosno planiranje gibanja u tri prostorne dimenzije oko predmeta generalno nepoznatog oblika zahtijevaju percepciju u tri dimenzije. Budući da slika daje dvodimenzionalnu infomaciju, rješenja u ovom radu oslanjaju se na RGB-D kameru, odnosno kameru koja osim slike u boji daje informaciju o dubini. Fuzijom sirovih trodimenzionalnih podataka i mjerenja iz različitih percepcijskih modula dobivamo precizan trodimenzionalni model biljke s označenim relevantnim funkcionalnim i morfološkim dijelovima biljke. Pritom se oslanjamo na dva pristupa u percepciji, odnosno detekciju i semantičku segmentaciju. Kod detekcije, modul percepcije daje nam informaciju o poziciji i veličini objekta određene klase unutar slike. Pritom se projekcija objekata u ravninu 2D slike generalno aproksimira minimalnim omeđujućim pravokutnikom, što znači da dio slike na kojem je pronađen predmet također sadrži dijelove scene koji ne odgovaraju tom predmetu. Kod semantičke segmentacije nema takve pretpostavke, odnosno moduli identificiraju one piksele slike koji odgovaraju određenoj klasi objekata. U primjenama razmatranim u ovoj disertaciji ova se dva pristupa percepciji koriste simultano za različite dijelove biljke. Tako se primjerice plodovi i cvjetovi detektiraju kao individualni objekti, budući da je za njihovu manipulaciju nužno razlikovanje instanci, a dovoljna je okvirna informacija o poziciji i veličini. S druge strane, lisnati dio biljke ili površina zemlje u uzgojnoj jedinici segmentiraju se semantički jer u ovom slučaju nije potrebna (ili ne postoji) informacija o pojedinim instancama tih

klasa. Primjerice, u slučaju da lišće zaklanja određeni plod, informacija o točnom broju listova nije krucijalna za izvođenje procedure branja ploda.

Druga komponenta znanstvenog doprinosa ove disertacije jest metoda za izradu trodimenzionalnog funkcionalnog modela biljke pogodnog za planiranje i izvođenje manipulacije. Metoda se oslanja na fuziju izlaza percepcijskih modula s mjerenjima 3D kamere. Tako su dijelovi radnog prostora robota semantički opisani i ta se informacija koristi u planiranju cilja i putanje robota za sigurno i precizno izvršavanje zadanog agrotehničkog zahvata.

#3 *Algoritam upravljanja robotskom rukom za podatnu manipulaciju u agrotehničkim zahvatima zasnovan na mjerenjima sile dodira.*

Kod manipulacije osjetljivih predmeta klasično upravljanje pozicijom robotskog manipulatora rijetko je dostatno za uspješno i sigurno izvršenje zadataka. Jedan je razlog nepoznavanje predmeta manipulacije, koje onemogućuje idealno planiranje izvođenja procedure. Drugi je razlog je nemogućnost prilagodbe robota neplaniranim preprekama (kolizijama) zbog prirode pozicijskog upravljanja. Kao alternativa, ovisno o dostupnim resursima nameću se dva pristupa. Prvi pristup prilagođava klasične pozicijski upravljane robotske sustave za podatnu interakciju pomoću senzorskog sustava za mjerenje dodirnih sila i momenata. Pritom se upravljački sustav može osloniti samo na povratnu informaciju o kontaktu na ograničenom području, u skladu s njom adaptirati svoju pozicijsku referencu i tako reagirati na prepreke u radnoj okolini. Drugi pristup oslanja se na razvoj pozicijskog upravljanja za kolaborativne robote. Ovakvi roboti inherentno prilagođavaju svoje gibanje prilikom kontakta s okolinom, ali ih ta osjetljivost čini manje preciznima u pozicioniranju.

U sklopu istraživanja i razvoja upravljačkih rješenja za agrotehničke zahvate u ovoj disertaciji korišten je potonji pristup. Treću komponentu doprinosa ove disertacije predstavljaju upravljački sustavi za autonomno i sigurno provođenje različitih agrotehničkih aktivnosti pomoću kolaborativnog robota. Pritom se upravljački sustav oslanja na algoritme podatnog i reaktivnog upravljanja robotskom rukom, a zadatci se planiraju i izvode korištenjem ranije opisanog funkcionalnog i semantičkog modela biljke.

Razvijena rješenja konačno omogućuju provedbu raznih agrotehničkih zahvata, poput branja plodova, kidanja zaperaka, mjerenja vlage i zalijevanja, robotskog oprašivanja, pinciranja, i drugih, uz lako proširenje na cijeli niz aktivnosti iz drugih domena. Kao primjer, provedena je eksperimentalna validacija u industrijskim zadacima poput zaglađivanja neravnih ploha, u servisnoj robotici kroz razvoj robota kuharskog pomoćnika, i u medicinskim primjenama kroz pomoć kirurgu tijekom operacije.

KLJUČNE RIJEČI: Konvolucijske neuronske mreže, preneseno učenje, sim2real, rgb-d, podatna manipulacija, kolaborativni roboti, impedantni regulator, agrikulturalna robotika, meki senzori, senzori dodira, optički taktilni senzori

CONTENTS

1

Introduction

A global consensus has finally been reached regarding the negative impacts of pesticide use on the environment [1], gravely influencing plant, aquatic and terrestrial wildlife, pollinators, and humans [2]. In combination with the growing demand for food worldwide, it is increasingly more difficult to meet the labour demands required for a sustainable large scale food production. Robotising organic agriculture has recently been growing in an attempt to alleviate some of these issues. One line of robotics research has for decades been developing machinery for traditional outdoor farming. Another research direction concentrates on indoor agriculture systems. Since in the latter approach the working environment can to some extent be structured, it is often considered better fit for deployment of autonomous robots that were originally designed for industrial environments. The work presented in this thesis, conducted within the SpECULARIA project, tackles some of the problems in the domain of indoor agricultural robotics.

Both the project and the acronym, standing for Structured Ecological CULtivation with Autonomous Robots In Agriculture, are inspired by the ancient roman proto-greenhouses, known as specularia. In his Book 19, the Roman agricultural writer Pliny the Elder describes how, during the reign of Emperor Tiberius, the vegetables were grown in beds mounted on wheels, which enabled moving them out into the sun and back inside for protection on wintry days. Our modern day SpECULARIA deals with reducing human labour input in agriculture, with specific focus on small indoor farms. The goal is to automate cultivation with a heterogeneous team of robots, and replace humans in the dull and difficult tasks, numerous in this application field.

The proposed heterogeneous team consists of three types of robots. An unmanned aerial vehicle (UAV) plays a surveillance role. It is equipped with a multi degree of freedom (DoF) manipulator carrying sensors, which enable monitoring plant status and health. A fleet of Unmanned Ground Vehicles (UGVs) is deployed for transport of growth unit containers. Those plants in need of manipulation, as identified by the expert system, are transported in their containers from their growth positions to the manipulation station. There, the third and final robot type is situated, a multi DoF manipulator, equipped to perform delicate handling of plants. Together, this robotic team will be capable of autonomously maintaining and harvesting an indoor farm, under the guidance of an artificial intelligence (AI) powered expert system. The focus of this thesis is on enabling the collaborative manipulator for autonomous execution of various plant treatment tasks and procedures.

The scope of this thesis includes perception and control techniques for robot manipula-

tors deployed in agricultural procedures. The remainder of this chapter aims to justify the relevance of the problem, and to identify the key challenges in achieving an autonomous robotic manipulator that can replace arduous manual labour in agriculture as efficiently as in industrial applications.

1.1 MOTIVATION

The development of industrial robots mostly followed the classic paradigm of perfect repetition of dull, dangerous or difficult tasks. These robots, usually designed with high positioning precision and repeatability were intended for use in very structured workspaces, where the manipulated objects as well as all the possible events are well known and can be predicted and accounted for a-priori. The benefits of replacing human labour with machine power was a strong driving force for further development in robotics, solving ever more complex industrial tasks. The successful deployment in these industrial settings was also an inspiration for investigating other possibilities of deployment, even in unstructured environments, through numerous efforts in the research community. Contemporary robotics challenges largely deal with extending the robotic capabilities for deployment with fragile environments, such as human everyday life settings in home, transportation and services, or outdoor settings such as search and rescue missions and agriculture. The complexity of robot tasks and deployment increases due to several issues, such as safety, unpredictability and uncertainty.

When considering agriculture in particular, the fast growing worldwide human population, and the corresponding increase in demand for food and natural produce are the main reasons for automating the production. Coupled with the increasing labour shortages, the motives and reasons for robotising the field are pretty self-explanatory. Not surprisingly, the development of agricultural robotics started in more developed parts of the world, as is clearly reflected in the several prevailing crops in focus of the existing research and technical solutions, specific to these geographic areas. The developed world is characterised with an ageing society, resulting in the lack of a work force in general. Moreover, urbanisation and globalisation provide better opportunities for the young people, leaving these parts of the world with a significantly reduced rural population. Furthermore, even more manual work load is needed when switching back to organic food production. Namely, reducing the use of various chemical solutions such as pesticides or artificial fertilisers in an attempt to meet a more sustainable growth of agricultural production means introduction of additional mechanical work load. The need for agricultural workers combined with the fact that manual labour in agriculture fits the description of difficult, dull, and dangerous jobs, make this field perfect candidate for automation. This potential was long ago recognized in the research community, but the development of commercial solutions was only enabled with the development of the supporting technology.

Deployment of robots on big farms is already a fast growing industry that focuses on big machines applied for specific crops and use cases [3]. Although task-specific solutions exist, the limited capabilities of various agricultural robots are preventing them from becoming a standard in food production. The main challenges can be identified within three categories: perception, motion planning, and dexterous manipulation. Depending on the crop and the working environment, different problems may occur in each of the domains. Throughout this

thesis, the existing solutions are analysed, with their important characteristics, advantages and drawbacks. These are used as guidelines in development of novel, more generalised solutions meeting the indoor agriculture applications criteria.

1.2 PROBLEM STATEMENT

The work presented in this thesis attempts to develop an expert system that can provide a robotic manipulator with inputs for a desired goal or behaviour. We consider the system at a complexity level that allows decision making within particular robotic actions, such as detecting and choosing a particular target among multiple detected objects. An example in the harvesting application would be for this system to decide which fruit is to be picked next. In classical industrial setups, such expert systems are most often based on observations and rationale of a human operator. Increased complexity of robotic applications, as well as the desire to automate entire production processes have led to development of expert systems based on artificial intelligence. This popular term, artificial intelligence, can represent a wide variety of methods and solutions of variable complexity and capabilities. The expert systems developed and analysed in this work all rely on the same sensory setup in the planning phase of manipulation, namely on a commercial RGB-D camera that simultaneously provides visual (colour), depth, and infra-red information of the recorded scene. Such a system is not a decision module that schedules or chooses activities, but a unit in the control framework responsible for information processing within procedure execution.

A first intuitive approach to implement an artificial thought process is usually in the form of a simple model, such as a decision tree or fuzzy rules. The most basic workflow is to detect an object of interest, plan robot motion with respect to the detected object, and finally to execute the manipulation procedure. Over time, a large number of experimental repetitions generates large amounts of data, which can be stored and processed thanks to continuous development of hardware. Data processing enables extraction of patterns from the historical data, which can then be used for decision making in the future. Theoretical contributions enabled with novel technologies yield ever more complex machine learning models that can find statistical relations between many decision variables, and imitate them within their figurative black boxes.

Finally, another layer of cognitive complexity is introduced into artificial expert systems with deep learning. These methods rely on iterative search for optimal representation of relations within the data, and have been shown to outperform even the most complex machine learning models. Furthermore, special forms of neurons enable imitation of natural inference, for example by processing temporal and spatial data, instead of their vectorized representations. The versatility of deep learning systems, built upon these novel neuron types (e.g. convolutional filters, or neurons with memory), enables development of both particular stand-alone detection modules, and of complete end-to-end solutions. Regardless of the particular application, generation of an extensive dataset is usually the most costly and tedious part in the development of these methods. In an attempt to alleviate these costs, many research groups have recently turned to exploration of sim2real approach, a promising, growing field, still with many open questions.

The 2D detection methods have recently been extended into the third dimension, thanks

to the extensive work in depth imaging that resulted in good availability of consumer depth cameras. This opens new possibilities in 3D perception for robotics, calling for development of novel planning and control methods as well. These robot control strategies should enable precise positioning of the end effector during target approach, while ensuring safety for the surrounding fragile environment. One way to improve safety in robotic agriculture is to include more domain knowledge, for instance in providing information about the structure of the working environment, the presence and absence of objects, that enable planning with active obstacle avoidance. Another way is to extend the sensing capabilities beyond the common machine-vision, for example by active sensing in compliant control frameworks.

Tactile sensing is a novel approach both in compliant manipulator control and within dexterous manipulation that is only beginning to receive attention in the agricultural robotics community. The requirements of safe dexterous manipulation heavily depend on the particular plant manipulation task. Most procedures involve some kind of physical contact, and require application of external force for successful task execution. Some examples include harvesting, topping, pruning, pest removal, weeding, etc. Compliant behaviour is crucial here for successful and harmless execution. Current state of the art robotic tools mostly rely on rigid solutions. These rigid technologies have traditionally been used in order to satisfy the requirements regarding short cycle times with precise positioning. This is in conflict with the necessity of compliant interaction with objects of high variability in size and softness. A solution that would reconcile these requirements could be in compliant actuators and end-effectors based on tactile sensing, soft structure and compliant control.

Finally, the execution of the agricultural procedure depends on the manipulator actuation capabilities and the deployed control strategies. The control of robot arms is a complex problem based on kinematic and dynamic modelling, in which different fundamental control approaches can be deployed depending on the basic actuation units in the robot joints. These approaches are either based on position/velocity, or on force/torque control, and are constantly improved, extended and optimised according to advancements and increased capabilities of the controlled systems. Manipulation in agricultural applications usually deals with fragile, dynamic and non-uniform objects, where compliance is often more important than high positional precision. However, during free space motion the position reference tracking is essential, regardless of the final manipulation objective. The control problem is then to reconcile the application requirements, relying on the available sensory data and actuation system.

1.3 HYPOTHESES

The problems considered within this thesis arising in the attempt of robotization of modern indoor agriculture are investigated experimentally, defining the hypotheses of the proposed scientific research as follows:

1. It is possible to obtain contact information with a novel optical tactile sensor through machine learning models.

2. It is possible to develop a compliant control algorithm for a collaborative robot that manipulates a living plant without harm.

3. Certain plant hygiene procedures can be conducted with a collaborative robot fitted with an appropriate gripper.

4. A robot manipulator equipped with a sensory apparatus can be programmed to autonomously generate a 3D plant model fusing multiple measurements taken from different sensors and robot poses.

These hypotheses are designed to solve the investigated problems in particular applications, having a general solution in mind in the form of a robot skillset. The goal is ultimately to supplement the human work force, and to reduce the workload of human workers where applicable, since complete robotization can hardly be realised even in much more structured setups. The hypotheses are also used to validate the realisation of the original contributions of the thesis, through development of various expert systems for particular procedure challenges.

1.4 ORIGINAL CONTRIBUTIONS

1. Tactile sensor adapted for compliant manipulation of fragile objects based on artificial intelligence

2. Method for generation of semantic 3D models of plants suitable for soft robotic manipulation planning in agrotechnical procedures

3. Control algorithm based on contact force feedback for soft robotic manipulation in agrotechnical procedures.

1.5 OUTLINE OF THE THESIS

This thesis is divided into x chapters, discussing the existing state of the art, and the original contributions of this thesis. The short summary of the following chapters is as follows:

Ch 2 This chapter provides a review of the state of the art and recent technological advancements in robotic agriculture. Special focus is given to three domains: perception, manipulation, and tactile sensing.

Ch 3 This chapter discusses the contributions of this thesis related to the visual perception for robotic applications in indoor agriculture. A brief introduction is given into calibration details for a depth camera in eye-in-hand configuration. Then, the developed components of the perception module are presented along with main experimental results. The scientific contribution regarding robot perception for generation of an object model suitable for manipulation is discussed.

Ch 4 This chapter discusses the contributions of this thesis related to the compliant manipulator control in agrotechnical procedures. Brief introduction is given into the main control concepts used in development of manipulation modules. Results of compliant manipulation procedures relying on perception results in planning and only limited feedback in execution are presented. The thesis contribution in robotic manipulation for agricultural tasks is discussed.

Ch 5 This chapter discusses the contributions of this thesis related to the domain of tactile sensing. A brief theoretical background is given about the inference model used in development of a feature extraction technique for an optical tactile sensor. The same principles with visual cues in the sensor are deployed with example manipulation tools, showing the technology versatile enough for extension into new domains, with promising prospects in enriching the available feedback during robotic manipulation.

Ch 6 this chapter gives concluding remarks of the thesis. The extent of particular contributions reaches are discussed along with identification of potential shortcomings and room for improvement as a guidance in choosing future work directions.

Ch 7 This chapter lists all the publications contributing to the main results of the thesis.

Ch 8 This chapter states the author's contribution to each of the included publications.

A list of referenced bibliography is given following the main body of this thesis. Afterwards, all the publications related to the main results of the thesis, previously published in peer-reviewed journals and proceedings of international scientific conferences are attached. A short biography of the author followed by a complete list of publications is given in the end.

2

State of the Art

In an attempt to achieve autonomous agricultural production, an entire line of research has emerged, dedicated to development of crops suitable for autonomous cultivation. This approach, lately known as *breeding for robotics*, has recently come into the focus of plant breeders, which aim to select commercially and nutritionally productive cultivars that are physically better suited for robotic treatment [4, 5]. This can be achieved by reducing variation in the plant population, in particular plant parts such as fruit, and in the growth environment, thereby reducing the uncertainties in the control framework development. For example, many woody and shrubby crops can be grown on nets in a planar structure, enabling deployment of simple solutions for robot motion planning, without handling complex problems of occlusions and obstacle avoidance. Other research directions might lead to selective breeding for plants that yield uniform fruit, or use the environmental conditions such as humidity, light spectrum and CO_2 to control the plant growth [5]. On the other end of the spectrum opposed to the *breeding for robotics* is the development of complex methods that can robustly handle organically dynamic natural structures. The best solution is often somewhere in between, combining the best of both approaches.

The thin line between machines and robotics is usually in the re-programming ability for realisation of various novel procedures. Following this definition, the solutions for increasing the level of autonomy of existing machines could be considered early attempts of agriculture robotization. One example is autonomous navigation of tractor vehicles [6], with one example approach based on geo-fencing for autonomy in navigation in large fields, combined with path planning for complete surface coverage. As opposed to the mature and commercial technology behind the GPS based autonomous navigation, more recent approaches rely on visual feedback for local planning. This approach, still mostly in the research phase, was made possible with the advancements of technological components that allow complex online computation. For example, the on-board camera and processing units can be used to detect and track crop lines. This is especially applicable to crops growing close to the ground, such as lettuce or strawberries. Most early systems rely on well known machine vision, such as Hough transforms for line detection. These methods are rarely robust to variations such as illumination, environmental conditions, or natural obstructions such as inter-row weeds. Similar to most other vision related problems, a solution might be in deep learning [7]. It should be mentioned that in addition to the conventional RGB cameras, near-infrared and multi-spectral cameras are widely used as well in the agricultural applications.

The commercial availability of more complex hardware has introduced novel methods, such as localisation in 3D maps generated by lidars. This approach is suitable for cultures grown in canopies, such as trees and bushes, though it should be mentioned that simpler RGB vision solutions have been shown successful as well, especially in combination with deep learning [8, 9]. The detection and tracking of the tall rows can be used in vehicle navigation and localization through the field. Similar to early vision based row detection, basic algorithms such as Hough transform can be used in combination with 2D lidar sensory data. Further technological advancements lead to wide availability of 3D lidars. The recorded 3D scans of the environment can be segmented using algorithms such as random sample consensus (RANSAC). For example, certain parts of the scene can be described using geometric primitives such as lines or planes. RANSAC can then be used for scene segmentation, to identify tree rows [10] or the ground plane [11].

The autonomous navigation problem as one of the first robotic steps in agriculture automation is not an end in itself, but a single step towards the complete autonomy in activities such as harvesting, seeding, weeding, precision farming in pesticide and fertiliser use and for irrigation, or remote inspection [12]. Among these and others, harvesting might be the most interesting to the research community, both for its challenging nature, and for the commercialization potential. A thorough review of the existing commercial technologies can be found in [13], divided based on the softness of fruit they are handling. Such classification implies that the variation in this research field is still too large to overcome with a unique solution, mostly due to challenges in two domains: perception and manipulation. [14]. The perception problem is related to precise detection and localization of the target object, often impaired due to organically cluttered operating scenes due to leaves, fruits and other plant parts obstructing clear view. Since the sensory hardware is at a very complex technological level, further development is relatively slow, and most of the new scientific breakthroughs are based on software advancements. The robustness of detection and localization improvements are lately heavily dependent on sensor fusion and deep learning [15]. The results presented within this thesis are built upon RGB-D sensory information. In-depth overview of the RGB-D state of the art is given later in this chapter. The software side is considered in the remainder of this thesis, along with discussion of the particular contribution.

The manipulation challenge in robotic execution of agricultural procedures relates not only to the task execution success rate, but also to the manner in which the task is executed as well. One of the most important objectives is to minimise the damage both to the manipulated object such as harvested fruit, and to other plant parts as well. Compliant control techniques are used along with adequate sensory setup. This topic is discussed in the Chapter 4. The other part of the solution is development of intelligent, controllable, compliant or reconfigurable tools [16]. While various industrial commercial solutions are available, none are close to widespread application in the agricultural field where the workspace is highly unstructured and the objects are delicate, easily bruised, adhesive and/or slippery [16]. In order to achieve the necessary compliance, these rigid tools are equipped with sensory components, with the prevailing majority based on visual or tactile feedback. Visual feedback is mostly used for geometric planning in object recognition, localization, and 3D model estimation. Despite a myriad of solutions across different application fields,

examples in the agricultural domain are sparse [17]. The tactile feedback provides more in terms of compliant control, since the feedback models contact forces and torques. Even though this technology has not yet reached market maturity, some examples have been developed for agricultural purposes, such as tomato harvesting gripper with two rigid fingers capable of force measurements with resistance based tactile sensors [18].

The rapid-prototyping techniques and new materials have enabled development of hardware solutions that can improve the compliance of manipulation inherently, by mimicking the natural, human grasping tools, i.e. fingers. Soft robotics is emerging as a fast growing research field, dealing with conflict-free manipulation by developing soft end effectors. A number of multiple-fingered soft grippers for harvesting can be found in recent reviews [16, 19], with most solutions, still in the developmental stage, either crop specific, or applicable only for a limited group of physically similar fruit. Some examples are shown in Figure2.1. The physical durability (wear and tear), and a reliable, robust feedback control despite inconsistent manufacturing of such tools are the main open questions in this field. A control system of such devices should enable precise motion reference tracking taking into account the deformations of the soft body during manipulation. While most current solutions such as those shown in Fig.2.1operate in open-loop, one solution could be in using information-rich soft bodied optical tactile sensors deployed as end effectors. This domain is considered in more detail within Chapter5.

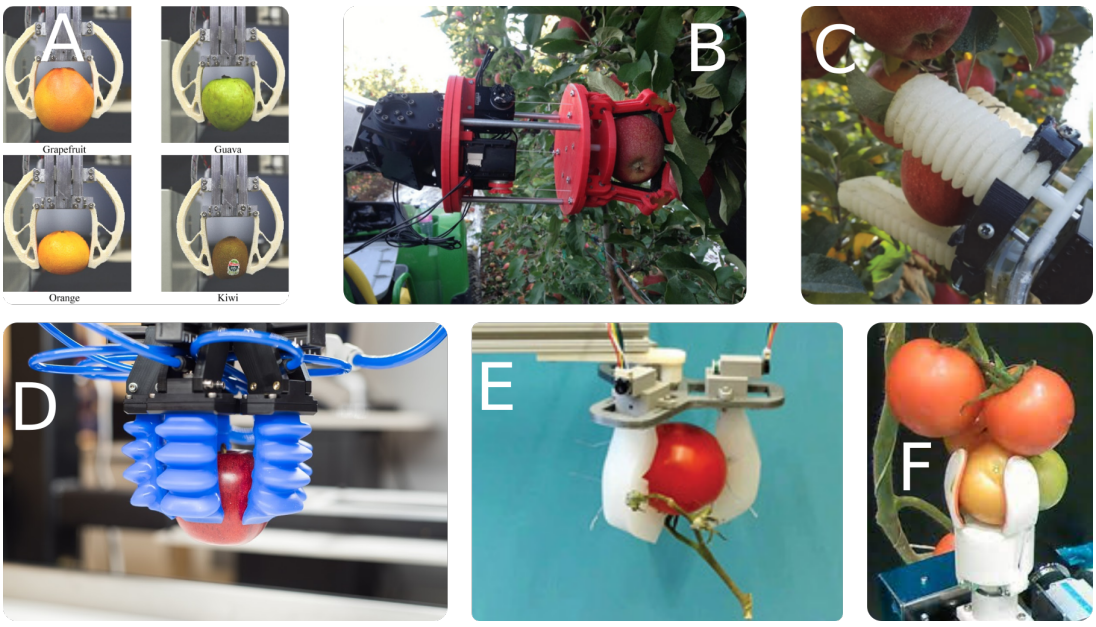


Figure 2.1: Examples of novel fruit harvesting grippers developed over the last several years. (A) a soft gripper with 3D printed compliant fingers [20], (B) 3D printed rigid robotic gripper for apple harvesting [21], (C) a soft-robotic gripper for apple harvesting [22], (D) commercially available food gripper by Soft Robotics Inc., (E) Smart soft actuator picking a tomato [23], (F) Rigid tomato picking gripper [17].

The research conducted as a part of this thesis follows the ideas behind the SpECULARIA project that focuses on indoor agriculture. This setup can be considered less complex in certain aspects thanks to the partially controlled environment conditions and relatively more structured cultivation when compared to open-field cultivation [15]. A team of ground

robots is imagined to autonomously navigate the indoor farm, and carry the plants in their growth containers to the manipulator workstation. The particulars of such a setup are considered in the remainder of this chapter, starting from the controlled conditions of the manipulator workstation. The sensory setup based on the eye-in-hand configuration is considered next, followed by a state of the art overview regarding perception.

2.1 CONTROLLED ILLUMINATION

One source of disturbance causing low detection rate in many precision agriculture applications is the uncontrolled dynamic working environment. For the most part, this relates to variable illumination conditions, such as changing weather conditions, sun direction, or shades generated by artificial or natural objects in vicinity. Various case-specific solutions have been proposed, either attempting to control the environment, to develop more robust methods, or a combination of both. In some cases, the researchers propose absolute approaches such as flooding the workspace with artificial light, or even operation exclusively at night [4]. One recent example of a more active approach to controlling the environment that attempts to generate controlled conditions in the inherently dynamic workspace is proposed in [24]. This novel and promising approach relies on a controlled illumination acquisition protocol called Flash-No-Flash (FNF). In this method, two images are acquired consequentially for the same setup, with and without strong artificial illumination, and the detection is based on the difference between these images, effectively achieving control over the ambient light.

As opposed to controlling the external variations, the perception method can be developed with the ability to adapt to external conditions. One example is the method described in [25]. The main working principle of this method relies on tuning threshold parameters for the three image channels, with which the object of interest can be detected against the background. The method works over various representations of the image colour space, depending on the specific use case. These threshold parameters can be automatically tuned for optimal object detection, depending on distribution of light and intensities in the image. Many other methods, similarly developed with the objective to generalise over variation in fruit size, shape, colour, etc., manage to generalise over environmental variation up to certain extent as well. These methods are further discussed in the following chapter.

The approach designed within SpECULARIA tries to take advantage of both presented approaches. Similar to classical industrial robotic manipulation where the controlled working conditions enable precise and successful task repetitions regardless of outer factors such as time of day, the project setup proposes a fixed, controlled manipulator workstation, where multiple conditions, including lighting, are controlled and isolated from the external variations. The workstation is imagined so that the robot camera system only captures the plant and a blank background. The background panels also block the lighting sources from the outside, and the most prominent illumination is the artificial light source of the workstation. However, in development of detection modules, great care is taken in designing them with an ability to adapt to and generalise over certain variations as well.

2.2 RGB-D SENSING

Numerous machine vision system have been developed for fruits detection and localization, mostly based on the RGB cameras in the earlier research, thanks to their affordability, versatility, and the availability of various computer vision and machine learning techniques [26, 27, 28, 29]. However, they have widely been replaced with RGB-D sensors ever since the introduction of the Microsoft Kinect camera in 2010 in many agriculture applications including detection and localization [30]. These sensors, in addition to the 2D colour information also provide depth measurements, and in some cases other pixel values such as infrared as well, inherently providing a high-density, textured three-dimensional (3D) point cloud of the recorded scene.

There are three most commonly used depth measurement principles in the available consumed RGB-D cameras [30]. The first one, called structured light (SL), calculates depth information based on the deformation of the structured light pattern they emit. This is the underlying working principle of the Microsoft Kinect V1, a popular choice among the research community when it was first released, with applications in fruit detection [31, 32] and yield estimation [33]. These types of sensors are sensitive to ambient illumination, and suffer from multi-device interference problems. The second common working principle of the RGB-D cameras is based on the time of flight (ToF) calculation, i.e. it estimates the distance by measuring the time it takes for an emitted light signal to hit the target and return to the internal light detector. An example of a commercial camera with this working principle is the Microsoft Kinect V2, the successor of the V1. This camera is one of the most popular cameras in the existing state of the art, thanks to the more accurate 3D measurements, higher frame rates, and conveniently smaller housing at a comparatively affordable price, with applications found across many different crops and cultures over the recent years [34, 35, 36, 29]. However, the technology again suffers from interference, and requires high power consumption for operation.

Finally, the third technology used in the consumer RGB-D cameras is called active infrared stereo (AIRS) technique, and it combines pattern projection of SL devices with passive stereo camera pairs. The work conducted within this thesis was done using the latter technology, inside the Intel RealSense D435 sensor. As a relatively newer commercial sensor, previous work is relatively limited, especially in comparison to the Kinect sensors, with some applications found for grapevine (using R200) [37] and apple orchards [38], and an SL based model F200 for pepper detection F200 [39]. Thanks to the more robust performance both in indoor and outdoor conditions, and a better working range in terms of close range measurements, these sensors are, at the time of writing this thesis, beginning to gain momentum in the research community. Simultaneously, Intel has announced that the RealSense series will be "discontinued" for unknown reasons, potentially meaning that a large body of recent research results might become obsolete quicker than expected.

2.3 AI IN VISUAL PERCEPTION

The inherent variability in fruits' size, shape, texture, and location is an important source of increased complexity of detection problems. Most early detection systems relied on

colour filtering for ripe fruit detection [40, 41, 42]. These pixel-based detection methods are developed by parameter (threshold) fine-tuning for maximum detection accuracy on the training data. Various crop-specific solutions have been proposed for cultures such as peppers [43, 44], citrus fruit [45], or apples [31, 46, 47]. However, as the results show in e.g. [48, 31, 49], this kind of fine tuning is a source of significant sensitivity to variations in either environment or fruit characteristics. Different attempts were made to improve generalisation of these segmentation methods especially in combination with 3D or depth data, relying on more advanced algorithms such as machine learning (K-means, support vector machines (SVM), and artificial neural networks (ANN) [50]. However, only with the advent of deeper neural networks, especially based on convolution filters, i.e. the CNN models, has the detection generalisation performance been immensely improved [30].

2.3.1 *Object detection*

According to some review articles, since the initial explosion in interest for deep CNN architectures in 2015, the widespread use of deep learning-based models in agricultural solutions has only kept growing and improving [50]. This coincides with the appearance of the three major CNN architectures that transformed the object detection approach, namely, Faster R-CNN (Region Proposal CNN) [51], YOLO (You Only Look Once) [52], and SSD (Single Shot MultiBox Detector) [53]. The main difference in the working principles of these networks is in the number of prediction stages. For the Faster R-CNN, built upon Fast R-CNN (built upon R-CNN), the prediction is conducted in two stages: first the region proposal, and then classification stage. The region proposal module is a network that proposes Regions of Interest (ROIs), on which the second stage, i.e. the classification is conducted, along with localisation. On the other hand, the other two approaches, YOLO and SSD, work as single shot detectors, skipping the proposal stage and solving the detection as a regression problem directly, optimising both bounding box location and confidence at the same time. As a rule of thumb, the Faster R-CNN models are more accurate and precise, but the computation times are significantly longer than with the single shot detectors, making them impractical for real-time deployment. The YOLO architecture is generally faster than the SSD, but the advantages of SSD are the shorter training times, multi-object detection capabilities and higher accuracy.

The CNNs have been found deployed in various agricultural tasks, including plant phenotyping, weeds detection, and yield estimation. Most of the applications include object detection, and use this either as the end result (e.g. in harvesting applications), or as an intermediate step in estimation of other variables. One of the most widely used architectures is the Faster-RCNN, mostly thanks to its longest presence on the scene, and its high accuracy in prediction. Namely, the early research considered detection problems that need not be conducted in real time. A classic deployment scenario would include detection of an object of interest, or multiple objects, and a consequential robot action. The list of existing state of the art solutions is constantly growing, as the applications are widened to new cultivars, predictions improved regarding 3D localisation and robustness to external conditions, and combined with novel robot tools in new agricultural applications. Recent examples include detection of sweet peppers [54], apples [55], kiwi fruit [56], passion fruit [57, 35], and many

other cultures [58].

Over time, with the improvements on the hardware level regarding the size, power consumption and cost, real-time computation came into focus as well, with the research community attempting to deploy neural networks for detection in closed-loop robot control as well. One straight-forward example is to use the position of the detected object as a feedback link in visual servo control. As mentioned, the YOLO approach is characterised with much faster computation times when compared to the SSD networks. The initially lower accuracy is now approaching, and in some cases even out-performing the SSD approach, with the intense development of newer improved versions of this architecture. However, the training time costs of the SSD networks are still significantly lower, and therefore much better fit for iterative development. With development of more efficient models MobileNets [59], and especially the improved version MobileNetV2 [60], the deep CNN inference was immensely accelerated. Though initially envisioned in embedded and mobile applications, deploying the MobileNet as the backbone of the SSD models enabled high speed inference and use of these complex CNN models in the closed loop robot control.

2.3.2 *Segmentation*

One of the drawbacks of object detection methods is met when a precise 3D location of the object of interest has to be extracted using some combination of depth data and detection bounding box. To properly extract the object from the background and obstacles encompassed within the bounding box, different techniques are developed, often task specific. In such cases, pixel-wise classification in images could be more beneficial. Namely, the detection bounding box encompasses other scene components along with the object, especially in case of occlusions. The working principle of early pixel-wise classifications in the image colour space is better able to overcome this issue, but suffers from the already mentioned sensitivity to variations. The most recent breaking point in semantic segmentation development that marked the beginning of the deep learning era in this field was the appearance of the fully convolutional neural networks (FCNs) [61]. The idea behind the FCNs was to adapt the early classification networks, such as AlexNet, for the task of semantic segmentation, based on the fully convolutional networks. The novel factor in these networks are the skip connections, i.e. links between the nonadjacent convolutional layers. These connections can to some point reduce the information loss due to max-pooling layers or dropouts. Though revolutionary, the FCNs had significant drawbacks such as loss of label localisation. Various developments managed to overcome some of the FCN challenges, such as the Encoder-Decoder concept within the U-Net architecture [62], that introduced skip connections between the corresponding layers of the encoder and the decoder.

Another novel improvement into the FCNs was proposed by a group of Google researchers in the DeepLab architecture [63]. This architecture enables computing feature maps at a higher rate thanks to the atrous or dilated convolution. The atrous convolution is achieved by upsampling the convolution filters by inserting zeros between non-zero filter elements. Several iterations of improvements of the DeepLab architecture have been conducted, resulting with the most recently developed DeepLabv3+ [64]. Another interesting concept to mention is the instance segmentation. As opposed to semantic segmentation,

where the pixels are classified into object categories, the problem of instance segmentation additionally distinguishes between different instances of the same object class in the image. This means that the instance segmentation will provide separate labels for, e.g. two leaves in the same image, whereas in the semantic segmentation, all of the pixels corresponding to leaves would be classified in the same category. The most prominent solution to instance segmentation is based on the Mask R-CNN network architecture, built upon the Faster R-CNN network. This network is developed as a two-stage inference machine that outputs the bounding box of the detected object, and the pixel-level semantic mask within.

2.4 SIM2REAL FOR DEEP LEARNING

The advent of CNNs in 2012 undoubtedly marked a new era in the field of machine vision, showing outstanding performance in various perception tasks, such as image classification, object detection and segmentation. However, the use of deep learning perception methods is associated with the cost of large training dataset generation due to the labor intensive data acquisition and manual labeling procedure [65, 66]. In an attempt to reduce the cost of the tedious labeling procedure, researchers have recently turned to the generation of synthetic datasets. Several independent results have shown that the inference accuracy can be improved by augmenting real datasets with synthetic examples, in cases where the amount of the annotated real data is insufficient [67, 68]. On the other hand, recent studies show that bridging the domain gap remains an open research problem for models trained on the synthetic data exclusively [69, 70]. For example, the synthetic datasets designed to perfectly match particular workspace conditions, aiming to achieve perfect domain adaptation, often results in poor generalization to even slight perturbations in environmental conditions.

Though neural networks are biologically inspired, some recent studies have shown that CNNs differ from humans in many crucial aspects [71], including tendency to rely on textural cues and local features heavier than on global features such as object shape [72, 73], even in classification tasks on well known large scale datasets, such as ImageNet [74, 75]. This texture-bias seems to be the main cause of the domain gap of the sim2real approach [76], and two basic approaches have emerged as a response: domain adaptation and domain randomization. Within the domain adaptation approach, the aim is to minimize the difference between the synthetic and the real data. One recently proposed approach relies on an intermediate domain, Content Invariant Representation (CIR), in which the synthetic images retain the same content, but with reduced difference in style when compared to the real images [77]. This can be realized in numerous ways, including for example the generative adversarial networks (GANs).

The domain randomization, on the other hand, attempts to increase the robustness of the trained models by maximizing the diversity in the synthetic dataset. Initially, it was proposed as a technique of randomized rendering [78], using random backgrounds, random lighting conditions, and randomly placed distractor objects of random shapes and sizes. The idea behind is to teach the model to perceive the real

2.5 SIM2REAL FOR DEEP LEARNING

The advent of CNNs in 2012 undoubtedly marked a new era in the field of machine vision, showing outstanding performance in various perception tasks, such as image classification, object detection and segmentation. However, the use of deep learning perception methods is associated with the cost of large training dataset generation due to the labour intensive data acquisition and manual labelling procedure [65, 66]. In an attempt to reduce the cost of the tedious labelling procedure, researchers have recently turned to the generation of synthetic datasets. Several independent results have shown that the inference accuracy can be improved by augmenting real datasets with synthetic examples, in cases where the amount of the annotated real data is insufficient [67, 68]. On the other hand, recent studies show that bridging the domain gap remains an open research problem for models trained on the synthetic data exclusively [69, 70]. For example, the synthetic datasets designed to perfectly match particular workspace conditions, aiming to achieve perfect domain adaptation, often results in poor generalisation to even slight perturbations in environmental conditions.

Though neural networks are biologically inspired, some recent studies have shown that CNNs differ from humans in many crucial aspects [71], including tendency to rely on textural cues and local features heavier than on global features such as object shape [72, 73], even in classification tasks on well known large scale datasets, such as ImageNet [74, 75]. This texture-bias seems to be the main cause of the domain gap of the sim2real approach [76], and two basic approaches have emerged as a response: domain adaptation and domain randomization. Within the domain adaptation approach, the aim is to minimise the difference between the synthetic and the real data. One recently proposed approach relies on an intermediate domain, Content Invariant Representation (CIR), in which the synthetic images retain the same content, but with reduced difference in style when compared to the real images [77]. This can be realised in numerous ways, including for example the generative adversarial networks (GANs).

The domain randomization, on the other hand, attempts to increase the robustness of the trained models by maximising the diversity in the synthetic dataset. Initially, it was proposed as a technique of randomised rendering [78], using random backgrounds, random lighting conditions, and randomly placed distractor objects of random shapes and sizes. The idea behind is to teach the model to perceive the real world as just another variation of what it has already seen, based on the variations inside the synthetic dataset. The approach was tested on object localization tasks, achieving accuracy high enough for robotic grasping in a cluttered environment. A convenient tool that stands out in this field is the NVIDIA Deep Learning Dataset Synthesizer (NDDS), a custom plug-in to the Unreal Engine 4 (UE4) [79]. This dataset was used in state-of-the-art non-realistic rendering research work, for example in generating a synthetic dataset of 100K images of cars, with randomised distractors on random backgrounds [68], and for the Falling Things dataset consisting of 61.5K images of household environment [80], which was then tested in 6-DoF pose estimation of household objects in [81].

Aside from the rendering pipeline randomization, another approach in the domain randomization field for developing models robust to perturbations and adversarial attacks is based on shape-biased datasets. Recent results show that increasing the shape-bias does

improve model robustness and improves generalisation, but at the cost of lower accuracy when compared to the texture-biased alternatives [74, 82]. Though most of the work in this field focuses on object classification, a certain amount of research effort has been directed into semantic segmentation tasks as well. Recent results in deep semantic segmentation [83] show that accuracy deteriorates significantly when training with data generated using the same style transfer method as in [74].

The sim2real approach has been tested in agricultural applications as well, with procedural generation pipelines developed and applied over various crops, weeds and cultures. The synthetic datasets created using real textures on 3D models in [84] were used to compare the three approaches on the task of semantic segmentation: training on a small dataset of real data, on the synthetic data, and on the combination of those. Even though the best results were obtained on the real data, the synthetic approach proved promising as well. Similarly, a smaller real dataset augmented with synthetic images was used in the leaf segmentation problem [85], resulting in improved performance on the real test set. A synthetic dataset of 10,500 images was recently developed for the sweet peppers (*Capiscum annuum*) semantic segmentation task as well [67]. However, in the generation of these synthetic images, great care was taken to ensure a very high degree of similarity to real images in terms of background and lighting conditions, yielding a dataset tailored for a specific setup, with low generalisation ability. Additionally, the rendered dataset is only labelled for semantic segmentation, an important and relevant task, but not sufficient for the entire scope of agricultural procedures. For example, in the harvesting scenario, object detection with localization is at least as important as the pixel-wise semantic segmentation.

world as just another variation of what it has already seen, based on the variations inside the synthetic dataset. The approach was tested on object localization tasks, achieving accuracy high enough for robotic grasping in a cluttered environment. A convenient tool that stands out in this field is the NVIDIA Deep Learning Dataset Synthesizer (NDDS), a custom plug-in to the Unreal Engine 4 (UE4) [79]. This dataset was used in state-of-the-art non-realistic rendering research work, for example in generating a synthetic dataset of 100K images of cars, with randomized distractors on random backgrounds [68], and for the Falling Things dataset consisting of 61.5K images of household environment [80], which was then tested in 6-DoF pose estimation of household objects in [81].

Aside from the rendering pipeline randomization, another approach in the domain randomization field for developing models robust to perturbations and adversarial attacks is based on shape-biased datasets. Recent results show that increasing the shape-bias does improve model robustness and improves generalization, but at the cost of lower accuracy when compared to the texture-biased alternatives [74, 82]. Though most of the work in this field focuses on object classification, a certain amount of research effort has been directed into semantic segmentation tasks as well. Recent results in deep semantic segmentation [83] show that accuracy deteriorates significantly when training with data generated using the same style transfer method as in [74].

The sim2real approach has been tested in agricultural applications as well, with procedural generation pipelines developed and applied over various crops, weeds and cultures. The synthetic datasets created using real textures on 3D models in [84] were used to compare the three approaches on the task of semantic segmentation: training on a small dataset of real

data, on the synthetic data, and on the combination of those. Even though the best results were obtained on the real data, the synthetic approach proved promising as well. Similarly, a smaller real dataset augmented with synthetic images was used in the leaf segmentation problem [85], resulting in improved performance on the real test set. A synthetic dataset of 10,500 images was recently developed for the sweet peppers (*Capiscum annuum*) semantic segmentation task as well [67]. However, in the generation of these synthetic images, great care was taken to ensure a very high degree of similarity to real images in terms of background and lighting conditions, yielding a dataset tailored for a specific setup, with low generalization ability. Additionally, the rendered dataset is only labeled for semantic segmentation, an important and relevant task, but not sufficient for the entire scope of agricultural procedures. For example, in the harvesting scenario, object detection with localization is at least as important as the pixel-wise semantic segmentation.

2.6 Manipulator control in chosen agricultural procedures

The most common agricultural applications such as robotic harvesting, pruning, pollination and similar are mostly tacked with solutions based on servoing, either image-based (IBVS) or position-based (PBVS). According to the experimental validation and analysis conducted on the sweet pepper crop, the approach presented in this thesis relies on position control through trajectory planning. Another important aspect in pure position control are the safety issues arising in case of collision or similar. In the work developed and presented as a part of this thesis, the position controlled scenarios only took safety into consideration implicitly, by planning the motion such that the probability of collision is reduced, as opposed to explicitly planning collision-free trajectories.

For many other tasks in agricultural manipulation, such as soil moisture measurement and tactile exploration of the stem, physical contact is essential during task execution, and has to be considered during control framework design. The soil moisture measurements in particular rely on an IoT sensor solution that utilises a simple breakout for a straightforward method to measure water content in the soil [86]. The moisture is estimated from the measured electrical conductivity of the soil between the two exposed pads of the sensor. In the standard farming approach, these sensors are kept in the ground for continuous measurements. Unfortunately, the exposed pads are quick to corrode, causing inconsistent measurements and harming the plants. Therefore, this sensor solution is recently being replaced with more expensive versions of soil moisture sensors. However, in the envisioned scenario the existing resistance-based sensors are utilised for two main reasons. First, such an end effector adapter is cost-effective, and the deployment with the robotic arm as opposed to fixed long-term measurements reduces the corrosion issue. The second motivation to use this type of sensor is in its physical resemblance to the U-fork, patented in France 1963, also known as the grelinette [87]. Since the tool is intended for soil aeration and drainage performed by digging the dirt around the plant to gently loosen it, the control method can easily be extended to this, and other similar applications as well.

Such manipulation tasks require a more cautious approach in robot motion control. One such widely used concept is based on impedance control, inspired by the pioneering work of Hogan [88], but extended into a whole range of different impedance control strategies

through extensive research over the past decades [89]. The versatility and success of this control concept enabled application in a wide variety of tasks, ranging from industrial manipulation to robotic rehabilitation [90]. A recent example in the field of agricultural applications can be found in [91], where an impedance control strategy was applied to the problem of compliant fruit and vegetable grasping. In this thesis, the strategy is applied to other agricultural procedures as well, in particular for soil moisture measurement, and as a part of the vine plant cleaning procedure. These are easily expanded to other procedures such as soil aeration or pruning of vine cordons.

Considering the problem of soil moisture measurement, several recent robotic solutions rely on design of a custom tool for a mobile robot, such as in [92], [93] and [94], but in these cases the applied force is not controlled or even monitored. These procedures hence cannot be considered as safe, neither for the robot nor for the manipulated environment. Another example of automated irrigation control is presented in [95], where static sensor circuits are distributed across the field, and a mobile robotic platform is used for visual moisture estimation. On the other hand, the solution developed within this thesis is easily extended for other applications such as digging and aeration. A robotic interaction with the soil in the shovelling task is presented in [96], where the applied force is controlled through relying on the modification of the humanoid robot's Center of Gravity (CoG).

Viticulture encompasses a wide range of regular and seasonal activities essential in grape cultivation, and has for a long time been a widely researched field, mostly thanks to the worldwide large scale wine production and market. Though there are innumerable novel technological solutions in various components of this wide field, the focus here is on the seasonal hygiene and pruning related procedures. Recent years have seen quite a number of novel robotic solutions, particularly for the seasonal cordon pruning task. Some recent examples include AI and computer vision based pruning solutions that autonomously make decisions, plan motion and manipulation, and execute the cutting procedure [97, 98, 99]. These solutions are however only based on remote, visual sensing, and even though they do involve physical manipulation during cutting procedure, do not otherwise consider the safety in terms of contact force sensing or control. It should be mentioned that the pruning procedure is a complex problem in itself, and other related research focuses on particular details such as intelligent pruning tool design [100, 101] and pruning planning [102].

2.7 TACTILE SENSING

When designing or deploying a tactile sensor, task specific design criteria must be considered. For tasks such as object exploration and tactile servoing, a high spatial resolution is required. Manipulation with fragile objects requires high sensitivity of tactile sensors. Slip detection and stable grasp require a frequency response higher than the vibrations occurring in contact with the manipulated object. Very often, the body of the sensor also has to be flexible so it could be mounted on any actuated body. Even if the whole sensor body does not have to significantly deform, realisation of certain friction and compliance of surface is necessary for grasping tasks. Materials forming deformable sensor bodies exhibit hysteresis, which tends to inhibit fast actuator response. Tasks that require multi-modal information use multiple sensors, where sensor wiring can become an obstacle in complex manipulation.

As with almost any other device, the important criteria for designers are sensor robustness and durability [103].

Various sensing technologies have been deployed in development of tactile sensors satisfying some of these requirements, including piezoresistive, piezoelectric, capacitive, and optical based sensors, but other less frequent technologies as well, such as quantum tunnel effect based sensors, barometric measurements based sensors, multi-modal sensors, electro-resistive sensors, etc. While the former three technologies build upon earlier development results from other technology fields, optical tactile sensing is a relatively young field. To adequately imitate human perception, tactile sensors should have 1 to 2 mm of spatial resolution and 50 to 100 sensing points [1]. Such a large number of sensing points can be easily realised with optical based technology. Tactile image sensors can be classified into three typical categories: light conductive plate-based, reflective membrane-based sensors, and marker displacement-based. In this thesis, we consider optical sensors that use visual information about material deformation (as opposed to the sensing technology using reflective properties of optic waveguide materials).

Over the years, several optical sensing technologies have emerged. One prominent representative is the GelSight sensor [104], an optical based tactile sensor with a soft contact surface capable of high-resolution sensing of geometry. TacTip is another optical sensor type with a similar working principle. This bioinspired optical tactile sensor was developed by Bristol Robotics Laboratory [105] based on the work of Hristu [106] and further developed over multiple iterations [107]. As shown in schematic on Fig.2.2, the sensor surface is an opaque silicone tip filled with silicone gel. An RGB camera is fitted inside the sensor housing, directed towards the interior of the tip. The working principle is inspired by the Merkel Cell complex in human skin, filling the inside of the sensor surface with a pattern of white-tipped pins that conduct and amplify the movement of the outer tip membrane while in contact with the environment [108]. An internal light source (ring of LEDs) is directed towards the pins, in order for the camera to record the pins' movements. This optical based design, in addition to the high resolution, benefits from other advantages, such as controllable measurement area, physical isolation of the camera and the sensor surface, and the ability to use classical computer vision algorithms and tools.

Figure 2.2: Cross-section and 3D schematic of the basic TacTip architecture.

This sensing technology enables extracting both spatial information and force measurements from the recorded pin movements using proper image processing algorithms. Machine learning methods such as Bayesian inference were successfully applied to data acquired from the TacTip with perception (localization) and manipulation as objectives ([108, 109, 110, 111, 112]). Genetic programming was used for force estimation model gen-

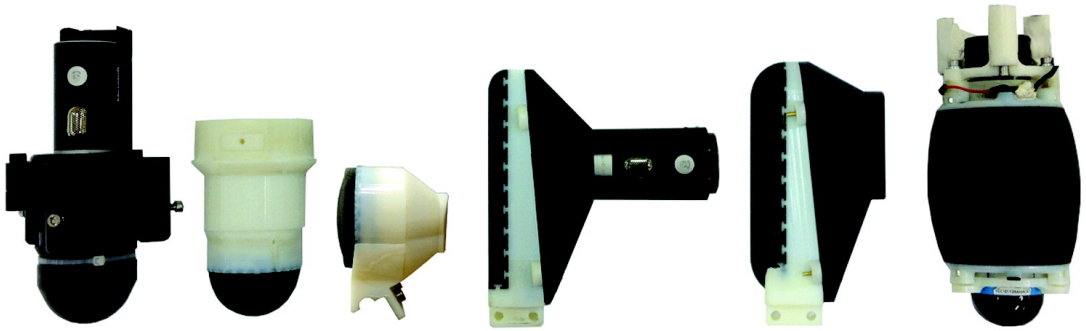


Figure 2.3: Examples of different architectures of sensors from the TacTip family of bioinspired optical based tactile sensors. From [107].

eration [113]. Analytical methods were developed for a sensor similar to TacTip based on the geometric model, and were used for touch modality and contact region identification [114]. Another analytical method was developed using geometrical projections, that enabled force estimation with an additional pressure transducer inside the sensor body [115]. Recent years have seen a surge in the deployment of deep learning algorithms, enabling prediction of both qualitative and quantitative measures. For example, a deep convolutional neural network (CNN) was trained to give robust edge perception and contour following, even in tasks beyond which the model was trained [116], while in [117] an autoencoder CNN was used to train a model predicting qualitative material characteristics, such as hard/soft and coarse. A complex approach based on CNNs along with a recurrent neural network (LSTM) layer was developed for the GelSight sensor in order to estimate the hardness of the manipulated object [118].

Many open questions remain in the field, especially in terms of model generalisation over sensor instances, since the trained models usually require retraining for each new sensor instance. The small batch production of the sensors means that the sensors are mostly assembled manually, and hence with a large variation between different instances. This issue is particularly prominent in the soft sensor field, since the soft materials building these sensors are prone to wear and tear. Furthermore, the bioinspired design enables development of different shapes of sensor housing, in an evolution depending on the deployment scenario in manufacturing, in-hand manipulation, object exploration, medical applications, and other [107].

The basic working principle is not only versatile across applications, but can be implemented in various geometries as well. Some of the already developed geometries are shown in Fig.2.3. Starting from the initial hemispherical design on the left, various design iterations produced planar and cylindrical surfaces, as well as a design suitable for exploration of structures hollow from the inside, on the far right. New designs include miniaturised versions mounted on tips of robot fingers.

3

Visual perception in agrobotics

In the research field of robotic perception for agricultural applications various factors can be identified that influence the task execution success. These factors arise in different phases, such as detection, localization, and during manipulation. One of the important external unstructured characteristics are the variable illumination conditions. Furthermore, the detected objects, i.e. natural structures such as fruit, are characterised with a high degree of variability, and require robust perception methods that can handle this variation. Leveraging the detection accuracy and computation resources, deep learning based models have in recent years emerged as a popular method with ubiquitous deployment. This is the approach exploited within this thesis as well, building upon the eye-in-hand sensory setup based on a commercial RGB-D camera. This hardware base is used for various agrotechnical procedures conducted within the SpECULARIA project, varying from regular care and hygiene procedures in cultivation such as soil moisture measurements and removal of excess flower, to seasonal activities such as harvesting. All of these rely on fusion of information obtained by 2D image based detection and 3D point cloud data, to provide the robot with a task reference in the global planning frame. The particular details of these activities are discussed in the following sections.

3.1 THEORETICAL BACKGROUND

The expert systems designed in this work rely on the sensory data acquired by an RGB-D camera that is rigidly mounted on the robot body. In the described setup the camera is mounted rigidly on the robot end effector link, in a so-called *eye-in-hand* configuration. This setup allows recording the object of interest from multiple perspectives with a single piece of hardware by simply controlling the global pose of the robot end-effector. An alternative approach would rely on a setup with multiple fixed cameras covering the workspace from various perspectives. The latter setup suffers from several drawbacks, such as higher equipment costs, and no reconfigurability, to name a few. With the *eye-in-hand* approach the scene can be actively recorded from various perspectives using only a single camera.

The sensory system has to be calibrated in order to properly use detection information in robot control. This means that a transform between the camera optical frame and the global reference frame has to be found. In our case, the global reference frame is in the robot base, and the transformation to the robot end effector frame can be found using the robot kinematic model. Then, the remaining unknown transformation is the one between

the camera optical frame, and the robot end-effector frame. A calibration procedure similar to the one described in [119] can be used to find this unknown transformation.

The calibration procedure relies on a target that is fixed in the global reference frame, and easily detected by simple computer vision algorithms. This could, for example, be an april tag, a checkerboard, or a simple contrasting blob. With the camera mounted on the robot end effector, the target is then recorded and detected in the image from various poses. A characteristic feature of the detected object can be chosen as, e.g. the central point. Using the pixel-position of this point in the image, its 3D position in the local camera reference frame can easily be extracted from the organised point cloud. This way, a set of N_{det} detections can be represented and stored as a set of data tuples 3.1, containing the position of the calibration target $\mathbf{t} = [x, y, z]^{\top}_{L_{cam}}$ in the local camera frame, and the robot joint configuration $\mathbf{q}$, in which the corresponding detection frame was recorded.

$$P = \{(\mathbf{q}_i, \mathbf{p}_i) \,|\, i = \{1, ..., N_{det}\}\} \tag{3.1}$$

The calibration procedure then aims to find the unknown transform from the camera frame to the robot end effector frame $\mathbf{T}^{\text{cam}}_{\text{ee}}$, to transform all of the recordings into the same point in the global reference frame L_0 using the complete transform $\mathbf{T}^{\text{cam}}_{0}$ obtained using eq. 3.2:

$$\mathbf{T}^{\text{cam}}_{0} = \mathbf{T}^{\text{ee}}_{0} \cdot \mathbf{T}^{\text{cam}}_{\text{ee}} | \mathbf{p}_{target} = \mathbf{T}^{\text{cam}}_{0,\text{i}} \cdot \mathbf{p}_i, \forall\, i = \{1, ..., N_{det}\} \tag{3.2}$$

where $\mathbf{T}^{\text{ee}}_{0}$ is obtained from the robot direct kinematics with known joint configuration $\mathbf{q}$. The calibration is in fact an optimization procedure, that minimises the dissipation of the target positions from the dataset P when transformed into the global coordinate frame using the complete transformation matrix $\mathbf{T}^{\text{cam}}_{0}$. Formally, the optimization problem can be formulated using eq. 3.3:

$$\mathbf{T}^{\mathbf{cam}*}_{\mathbf{ee}} = \arg\min_{\mathbf{T}} \sum_{i=1}^{\text{N}_{\text{det}}} \sum_{j=1}^{\text{N}_{\text{det}}} \| \mathbf{T}^{\text{ee}}_{0,i} \cdot \mathbf{T} \cdot \mathbf{p}_i - \mathbf{T}^{\text{ee}}_{0,j} \cdot \mathbf{T} \cdot \mathbf{p}_j \|, \tag{3.3}$$

where the dataset P from eq. 3.1 is substituted to calculate $\mathbf{T}^{ee}_{0} = f(\mathbf{q})$. The optimization problem is resolved using the Nelder-Mead optimization algorithm in the applications presented within this work, but other search methods can easily be deployed as well. Finally, with a known transform from the camera optical coordinate frame to the global robot frame, the agrotechnical procedure can be conducted in the robot workspace, based on the RGB-D camera inputs and/or feedback.

3.2 APPLICATIONS

3.2.1 *Soil moisture measurement*

Precision irrigation is emerging as an indispensable component of farming. On the one hand, plant watering is necessary for a healthy yield, and controlled amounts can improve the crop quality. On the other hand, the increasingly changing climate conditions have to be taken into account in reasonable resource consumption. Two most widespread approaches are drip-irrigation systems and sprinkler based irrigation. The most cost effective solution relies on controlling the entire system with the same control signal - either watering, or not

watering. More complex (and expensive) setups have the ability to control certain actuators individually, such as specific controllable valves for each growth unit. In these cases, a distributed sensory system is essential. Soil moisture is usually measured using simple probes whose working principle is based on measuring medium electrical conductivity. However, such metal probes are prone to corrosion, and should not be permanently placed in the soil.

The soil moisture measurement procedure is relatively fast, and represents a perfect procedure candidate for automation in agriculture. The setup is organised according to the SpECULARIA project scenario: a plant in need of treatment is positioned within the manipulator workstation. The manipulator is equipped with an RGB-D camera that outputs an RGB image of the scene, along with an organised point cloud of depth recordings in the local optical frame. Segmenting the ground pixels from the scene in order to determine the manipulation reference position for the robot manipulator in the global reference frame, can be considered a semantic segmentation problem.

As a more recent and promising approach when compared to classical object detection in images, the semantic segmentation problem has attracted a lot of research focus across a wide range of application fields, such as autonomous driving, medical systems and different robotic applications. Early solutions were based on segmentation using some intuitive features, such as pixel colour. Recently, it has found application as an initial step in the pipeline of more complex methods, such as colour masking the image during manual labelling for dataset generation, for faster labelling times. However, this approach is still widely used for certain simpler problems as well, such as the ground segmentation problem presented in this work.

The RGB image of the scene is first filtered in the HSV colour space, in order to mask everything but the ground pixels in the image. The generated mask is smoothed using basic CV filtering operations (i.e. morphological opening). This smoothed mask is applied to the organised point cloud. In an ideal case, the filtered point cloud contains only those pixels that correspond to the soil in the plant pot. In reality, there are some pixels that correspond to other parts of the scene, and some of the soil pixels are filtered out. This means that the filtered point cloud, besides most of the soil points, also contains some outliers. Using classical segmentation methods, these are easily filtered out. In our experiments, we used a basic Random Sample Consensus (RANSAC) method for fitting a plane equation to the soil samples. Upon removing the points that do not correspond to the fitted plane, we're left with the best approximation of the soil surface in the local camera coordinate frame. Using the camera-global coordinate frame transformation, optimised using eq. 3.3, an estimate of soil surface can be obtained in the robot reference frame. From here on, the robot control plans motion to the soil surface, and reach it in a perpendicular configuration for optimal manipulation of the sensing probe.

This simple approach has several serious drawbacks, among which colour filtering is the most prominent one. Even with the controlled lighting conditions in the workstation, a more robust solution than colour filtering is needed for ground detection. Namely, the colour of the soil changes depending on the moisture. Problems can also occur due to the shadows, depending on the plant canopy. One simple solution to improve the robustness of detection would be to develop a machine learning model that would determine colour

filter parameters for soil pixel segmentation. This could be, e.g., by learning the colour filter parameters from a labelled dataset of images. Another direction could be to iteratively search for soil surface directly in the 3D space, i.e. in slicing the point cloud and trying to segment the best plane, especially with an a-priori (roughly) known position of the pot with respect to the robot. This approach is shown in the left image of the Fig.3.1, where the pointcloud is sliced into 2 cm slices from the known table height upwards. When comparing these slices, the slice containing the soil is expected to contain the largest number of points, since the lower slices only contain the points lying on the pot surface. Iterative search with decreasing width of the slices can yield a point cloud conveniently subsampled for RANSAC based plane fitting. An example is shown in Fig.3.1on the right, where the point cloud scene is reduced to the slice containing the ground, and neighbouring parts of the scene containing points of close-by objects. These points can be handled in several ways, first by limiting the search space tighter around the known pot position. Then, the RANSAC algorithm in the plane fitting procedure eliminates the outliers by design.

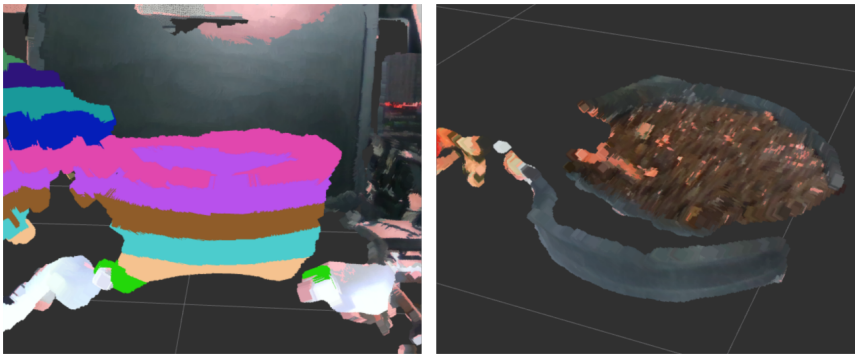


Figure 3.1: Left. The height-wise sliced point cloud for determining the soil pixels in the scene containing a potted plant at a roughly known global position. Right. Filtered point cloud of the pot slice containing soil points.

3.2.2 *Removing Excess Flowers*

Flower detection and manipulation is essential in various agrotechnical procedures. For example, many farmers rely on flower counting for timely initial yield estimation. Depending on the environmental conditions and other expert knowledge, in some cases the flowers have to be pruned for yield control. The mechanical part of this procedure is identified as another perfect automation candidate in indoor farming, using a compliant collaborative robot equipped with an adequate sensory system. Another example of a flower related activity is pollination, often conducted in indoor cultivation where there is a lack of natural pollinators. Some examples of suitable tools for these agricultural procedures are shown in Fig.3.2. In the experiments presented here, a tool primitive was used instead, i.e. a pointer whose tip was positioned as the actual agricultural tool would be.

In comparison to the problem of ground segmentation, the detection of plant parts such as flowers is significantly more complex. Therefore, a more elaborate approach is taken, based on convolutional neural networks (CNNs). More precisely, transfer learning is deployed for a MobileNet-V2 SSD network pretrained on the COCO dataset, for training a model for

Figure 3.2: Examples of possible manipulation tools for flower cutting (left), pollination (middle), or picking (right). The cutting tool on the left is inspired by [120].

flower detection. In the remainder of this section the generation of the training dataset is discussed, followed by description of a method for 3D segmentation of the manipulated plant part that ensures a higher degree of safety during procedure execution.

Figure 3.3: Examples of synthetic images of strawberry plants rendered in Blender using custom dataset generation pipeline

DATASET GENERATION.Inspired by the success of such an approach in similar problems, the first attempt in the generation of the training dataset for flower detection relied on generating a realistic 3D model of the flower through photogrammetry. Due to several reasons including the complex 3D shape of the flower, and monochrome white petals without texture photographed against texture-less backgrounds, this procedure gave poor results. A large dataset was then generated in Blender [121] using procedural generation based on manual generation of a realistic 3D strawberry plant model. Leaf models are generated by importing photographs of real strawberry leaves as planes into Blender, and morphing the planes for more realistic shapes. Other model components, namely pot, stems and flowers, are modelled from scratch, with flower texture generated using texture maps from images of real flowers. Several basic fruit models were generated, some manually, and some using open source photogrammetric tool Meshroom [122]. In procedural generation, the components were spawned pseudo-randomly around the stem as a parent. The number and arrangement of plant components are also decided by pseudo-random sampling from predefined distributions. The images are rendered using Cycles render engine, and image labels using Eevees. The labels were generated as bounding boxes for all of the separately rendered scene components. The final component of the dataset is the background, which was uniform in 5% of the images, and highly cluttered in the remaining 95% of the images.

A dataset of 100 images of different textures and environments was used as the cluttered, texture rich background. The synthetic dataset finally consisted of 1980 images of strawberry plants. Examples of rendered images are shown in Fig. 3.3.

Deep CNN models were trained for detection, relying on transfer learning for SSD with MobileNet V2 backbone. The results of an experiment proving detection precision are given in Table 3.1. The experiment was conducted using the Intel RealSense D435 depth camera mounted in the eye-in-hand configuration, and recording the same strawberry flower in 10 repetitions. The 3D position is estimated by filtering the point cloud using the bounding boxes of detections in 2D images. Ideally, the estimated position estimations would be identical in all experiment repetitions, but due to occlusions, detection and camera imprecision, and possible inaccuracy in the camera-robot calibration, a deviation at a sub-millimeter level in the estimated flower position occurs in estimations. As can be seen from the results given in Table 3.1, the position estimates over experiment repetitions are consistent with small variation. The standard deviation of estimates around the mean value is sub-millimeter, as are the average and maximum error when considering experiment repetitions. This precision is at the level of robot repeatability, and has no effect in task execution.

Table 3.1: Dissipation of the flower 3D position estimates over 10 experimental positions against the mean.

	mean error	max error	std
x [mm]	0.10	0.27	0.13
y [mm]	0.14	0.27	0.17
z [mm]	0.18	0.66	0.25

3D SEGMENTATION. The information detected from the 2D data, such as object location within the image, can be directly used in robot control in frameworks such as visual servoing. These methods are inherently not aware of the surrounding obstacles, hence neither particularly safe for objects in the working environment. This can be achieved easier using point cloud analysis and various methods developed for 3D object detection and semantic segmentation. As opposed to the 2D problems, the neural networks are not a leading approach in the state of the art of this field due to several factors, including the fact that the affordable 3D sensing hardware has emerged only recently. Furthermore, the suitable supporting hardware has not yet been developed. Such hardware would enable the development of efficient architectures and optimization methods, as the GPUs enabled for the convolutional neural networks. For these reasons, this young research direction is still favourable towards novel solutions. As an example of 3D data segmentation based on a classical approach, a method developed for the agrotechnical procedure of flower pruning is described here.

The flowers, detected in 2D images generated by the RGB-D camera mounted on the robot in certain cases have to be pruned. For example in case the plant is still young, and does not have enough resources for development of a rich fruit yield, certain flowers would be removed by the human expert, to both focus the limited existing resources on growing

fewer quality fruit, as well as on growing other supportive plant parts, that could in turn support higher yield in the next phase. When conducting this operation, the flowers of a plant can be detected by an object detection module, as described earlier based on the MobileNet SSD model trained for object detection. Merging this information with the depth information from the RGB-D camera, the flower can be localised as well. Furthermore, using 3D data enables harmless manipulation, leaving the surrounding leaves and other plant parts intact. A simple approach is developed here that reduces collision probability, without deployment of complex and computationally expensive semantic segmentation modules.

The plant in treatment is recorded by the RGB-D camera mounted on the robot end-effector. Moving the robot through a series of poses, recordings are collected from various perspectives. Depending on the plant morphology, some of the perspectives might offer a more direct or occluded approach to the manipulated flower. To achieve harmless manipulation, it is necessary to choose an optimal approach trajectory to the flower. One complete solution would be to generate a detailed 3D model of the plant by fusing point clouds from several perspectives. This step, computationally expensive itself, would then be followed by another complex step of planning a safe approach trajectory. A straight-forward and fast approximation of this process is developed and experimentally tested instead, in a procedure as follows: on the example of young strawberry plants, a flower is detected in a 2D image. The bounding box is used to filter the corresponding organised point cloud, provided by the same RGB-D camera. The flower is then modelled with an approximate bounding volume, centred around the median flower point. More precisely, for each viewpoint of the flower, only the local camera approach axis is considered, relative to the flower distance to the camera. A dead-zone threshold is defined depending on the expected flower size, for example 1 cm distance to the flower median point is considered the flower bounding volume. The proposed method then relies on counting the number of voxels (points in the point cloud) that occur between the camera and the flower central point. This is calculated for all the considered viewpoints, in which the flower is detected in the 2D image. Finally, among all the considered viewpoints, the one with the least number of points obstructing the approach to the flower is chosen as a starting point in the manipulation procedure.

The described procedure enables planning a trajectory of the robot end effector towards the manipulated flower, with a reduced probability of collision with other parts of the plant or objects in the environment. The planning problem is reduced to computation of a direct path from the initial pose. Experiments have been conducted for testing this approach, showing that the optimal approach pose can be chosen with a simple thresholding approach. In the experiments, the case in which the flower is heavily occluded in all viewpoints was not considered, i.e. an aborting decision was not implemented. Several straight-forward solutions could easily be implemented, such as maximum number of occluding points, or a margin (either relative or absolute) to the mean (median) number of occluding points for a particular set of viewpoints. More elaborate AI approaches could be developed as well, relying on hidden features and statistics in the (manually labelled) data.

Experiments have been conducted testing the entire pipeline, including plant recording from multiple viewpoints, choice of the starting point using the described method, motion planning, and finally positioning the end effector to a position ready for manipulation. The

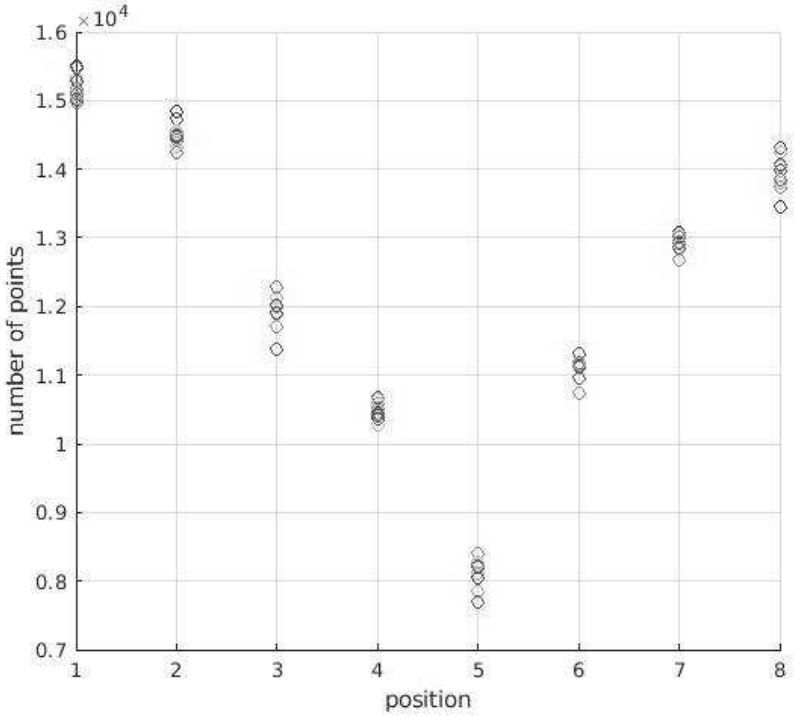


Figure 3.4: Number of obstacle points for the approach poses in which the flower has been detected in all experiment repetitions. Pose number 5 was repeatedly chosen as the pose with the minimum probability of collision.

experiments consisted of positioning the end effector on 20 different strawberry flowers and manually inspecting the positioning accuracy. The positioning was considered to be successful if the robot equipped with the appropriate tool would be capable of performing the flower thinning and flower pollination task from the reached position. Out of the 20 experiments, 90% (18/20) were successful, while 10% (2/20) of the attempts failed due to inability of the robot to reach the goal position. In other words, the flower was correctly detected, but the target pose of the end effector was outside the robot workspace. The photographs of the correct positioning are shown in Fig. 3.5.

3.2.3 *Pepper picking*

As a high-gain procedure with a widespread application and easily quantified results, harvesting has been one of the first automation candidates among agricultural procedures. Within SpECULARIA, pepper harvesting has been considered as a highly cost-effective robot procedure in indoor pepper cultivation. The solutions have been developed incrementally, relying on state of the art technologies, such as transfer learning of the pretrained MobileNet SSD architecture, fusion with organised point cloud data generated by the RGB-D camera, and finally semantic segmentation for improved estimation.

Similar to the approach in flower detection, a deep learning approach based on a MobileNet SSD was developed for the sweet pepper detection. The network was trained using transfer learning, and according to the recent research results, the training was conducted on a combination of large simulated and small real dataset. The detection results were fused

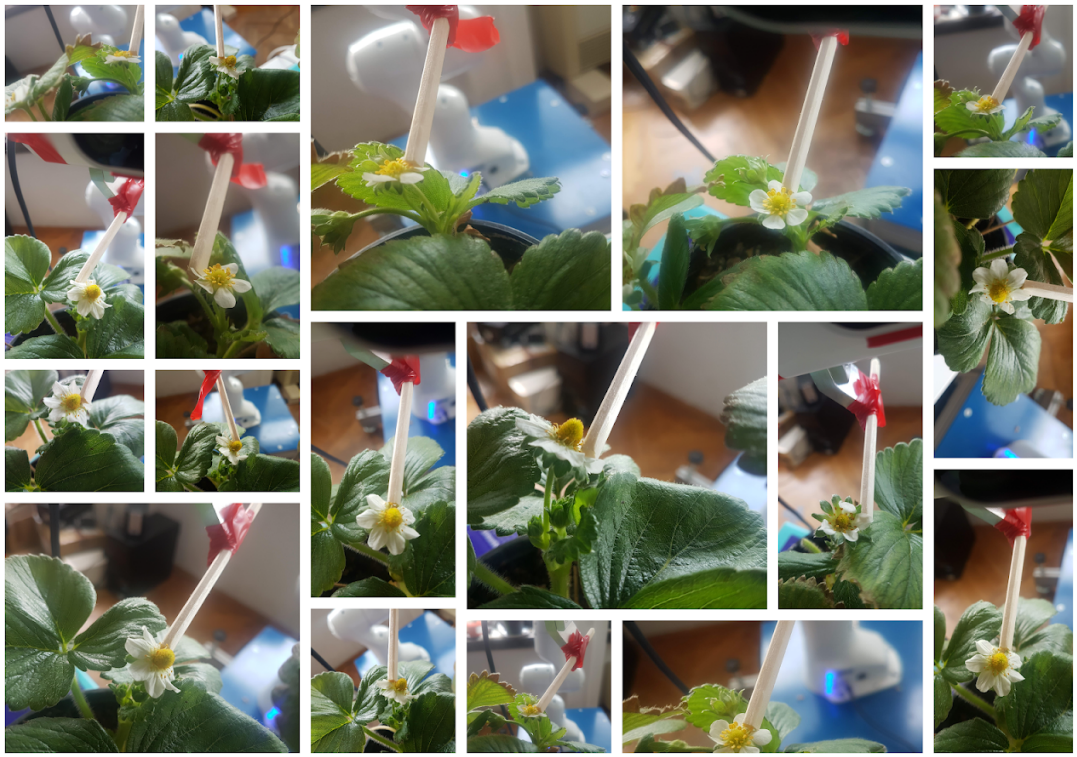


Figure 3.5: Results of the 18 successful flower manipulation experiments out of 20 trials. The example tool is positioned at the manipulation ready position, where a procedure could be executed using an actual tool.

with the depth data in the form of an organised point cloud, and a 3D geometric representation of the fruit was estimated for robot motion planning in harvesting. Then, method performance was improved with a semantic segmentation module based on a pretrained DeepLab architecture.

↬ SWEET PEPPER FRUIT DETECTION.The large simulated dataset for CNN training was created with procedural generation in Blender. The model of the plant was generated composed of various plant parts in a realistic plant shape. The parts were generated using different modelling approaches and techniques. For instance, the stem and pots were drawn and modelled completely from scratch, with standard tools available in Blender modelling software. Leaves and flowers, on the other hand, were photographed and imported in Blender as planar models, which were additionally morphed to better resemble realistic flower and leaf shapes and silhouettes, as well as to increase the diversity of shapes in the plant. More precisely, leaf and flower planes were scaled, skewed, and twisted into more realistic 3D shapes.

Finally, highly realistic 3D pepper models were generated with a more complex method based on digital photogrammetry. A series of around 200 photographs was taken using a smartphone camera of a single pepper together with a peduncle. The images were taken from different angles, capturing the entire fruit with overlap between two consecutive photographs of around 80%. Open-source software package, Meshroom [122], was used for photogrammetry. At first, the highly reflective surface of the peppers rendered initial

3D reconstructions of very poor quality. In order to prevent glossiness and add texture to the peppers, the photographs were taken after the peppers were sprayed with the matt spray that contained talk. The final Meshroom 3D model was additionally perfected in Blender, and an RGB texture was added both to the pepper and the peduncle. Even though the original pepper was red, the dataset is extended with peppers artificially coloured into the range of colours, spanning from green and yellow to orange and red. This way, the model was prevented from learning to distinguish between fruit and leaves based on colour solely. Following a similar logic, distractor objects were added to teach the model that not everything around the orange colour spectrum is a pepper.

In generating the dataset, plant models were created in a realistic shape, with stem curves resembling real live plants. To achieve this without exact empirical measurements of plant body parts, a Blender plugin sapling tree generator was used. The plugin allows control of certain key characteristics, such as branch splitting, pruning and resolution. Vertices can be chosen on the stem curve where certain plant body parts can spawn, as well as the range of allowed attitudes, with respect to the stem shape and gravity. Introducing random simple variations in morphology, scale and number of different plant parts per stem can generate a large amount of various bell pepper plant examples.

An ablation study was conducted to evaluate the validity of the generation method. The first ablation dataset evaluates introduction of distractor objects, and is hence referred to as $S2$. Here there are no distractors in the rendered scene, and only red peppers are generated. An example of an image generated in the complete dataset $S1$ and ablation dataset $S2$ can be seen in Fig.3.6on the left and centre, respectively. The second ablation dataset, $S3$ aims to justify the effort of generating morphologically organised plant models. In this dataset, the same number of plant body parts was spawned, randomly scattered within a predefined volume instead of systematically arranging them along the stem. An example from this dataset is shown in Fig.3.6on the right.

Figure 3.6: Examples of images in the synthetic dataset $S1$ (left), ablation dataset $S2$ without the distractors (centre), and the unorganised dataset $S3$ (right) for validation of plant morphology modelling.

The synthetic dataset $S1$ is used to train a detection CNN intended for deployment on real pepper plants. For model evaluation and test, a small dataset of real images was collected in the greenhouse, and manually labelled with the same labels as the synthetic dataset. This real dataset denoted R is collected with a smartphone camera and consists of two types of images. First, a set of 41 images, referred to as $R1$, contains images presenting an

unstructured greenhouse environment, as shown in the upper row of Fig.3.7. Such images may occur in applications such as yield estimation, where a robot will autonomously count the ripe fruit or flowers to predict the yield of the entire crop. Then, in the set referred to as *R*2, there are 10 images of a single plant with minimum environmental distractors, as seen in the examples in the lower row of the same figure. This set is relevant for applications in tasks such as harvesting or plant hygiene, where the handled plant is isolated from other environmental cues within the robot workspace.

Figure 3.7: Examples of real pepper images in the unstructured greenhouse environment *R*1 (upper row), and under controlled environmental conditions without distractors *R*2 (lower row).

Figure 3.8: Pepper detection results on the real images from the test set. Ground truth human annotator labels are drawn in green. Results from networks trained on datasets *S*1, *S*2, and *S*3, are shown in blue, red and yellow, respectively. Different learning results can be observed on different datasets.

As mentioned, the training procedure relies on the successful concept of transfer learning. A MobileNet-SSD network, pretrained on the COCO dataset was trained in the TensorFlow 2 environment to detect and localise sweet peppers. During the learning phase, all layer weights were unlocked for tuning. For additional data augmentation, random horizontal

Table 3.2: Ablation study validating dataset generation procedure.

	$S1$	$S2$	$S3$
R1	13.1	16.6	30.0
R2	27.1	19.7	15.9
R	17.9	14.1	27.0

flip and random image crop were used. The training was done with momentum as optimizer, in batches of 8 samples. Non max suppression is used in post-processing, with IoU score threshold set to 0.2, maximum number of detections per class set to 50 and maximum number of total detections set to 100.

First, we validate the method for procedural generation of the dataset by benchmarking the relevance of what we recognize as important dataset features. For each $S2$ and $S3$, we train a model with the same architecture, and evaluate their detection capabilities on a test set consisting of real images. This training and evaluation procedure is conducted on the pepper fruit class only. The results (AP) are presented in table 3.2. As expected, the set $S2$ that contains only red peppers and does not contain distractor objects has overall achieved the poorest performance. As can be seen in the several examples in Fig. 3.8, the red bounding boxes are missing in all the peppers that are not red. This also explains a somewhat better result when compared to $S1$ on $R1$ - the AP score was relatively high due to the large proportion of red peppers in the dataset, to which the $S2$ basically over-fitted colour-wise. The network trained on $S1$ on the other hand, in an attempt to better generalise, missed some of the red peppers. On the $R2$ set of test images, i.e. those representing a more structured environment with a single plant in the scene, the $S1$ managed to obtain a high AP score of 27.1. This result is significantly higher than for other two ablation sets, which speaks in favour of organised structure of the synthetic plant for detection problems in structured environments.

The results of the $S3$ are somewhat unexpected, seemingly surpassing the proposed approach in the overall results. This can be explained with visual similarity between the randomly generated synthetic samples in this case, and the randomness in the cluttered greenhouse environment, as well as a larger proportion of these images in the combined R dataset. In fact, this model performs surprisingly poorly on the $R2$ set, which otherwise presents an easier problem for both other models. These results suggest that the proposed method does not generalise to the whole spectrum of intended applications. On the other hand, the results imply that with minimal adaptations in the dataset generation, models can easily be tailored and optimised for the intended use-cases. Since this study relied on the pepper harvesting task, the remainder of this analysis considers results obtained with the $S1$ dataset.

To validate performance of the method against real data, the trained network performance is again evaluated with a mean Average Precision (mAP) metric. To give the results some context, it should be noted that mAP of the MobileNet-SSD network on the COCO dataset is 22.2. In this experiment, three networks of the same architecture were trained to detect both peppers and flowers. The first one was trained on a synthetic dataset developed

Table 3.3: Detection network evaluation.

Train set	synt	real	synt+real
mAP	15.9	17.4	20.8

with the described method. The second one was trained on a small dataset of manually labelled real images. Finally, the third model was first trained on the synthetic dataset, and then fine tuned on real images. The models are evaluated on the same test set of real images. The results presented in Table 3.3 imply that even though not completely realistic, the synthetic dataset is useful as a pre-training set, to improve detection performance when the real dataset is small, which is the case more often than not. The optimal deployment scenario seems to be generating a large synthetic dataset, and in addition annotating a small number of real images. The effect of training on the combined dataset instead of consequential training on synthetic and real data was not tested in this work, trusting the results from [68, 67].

The final validation of the method is conducted as a part of pepper harvesting experiments, presented in the work that can be found in the attached publication Pub.1. The detection precision is evaluated as the first step in a pipeline for estimation of 6DOF pepper pose. Furthermore, the generalisation ability of detection is tested in the evaluation of robustness of 6DOF pose estimation. The approach is finally validated in experimental repetitions of robotic harvesting of several sweet pepper varieties.

CNN FOR SEGMENTATION. As discussed previously, the field of segmentation is a relatively young research topic. Most research results rely on classical methods, while the deep learning approach is still largely in the development phase. Similar to the field of object detection, the most significant progress in deep learning based semantic segmentation was reached only with the advent of fully convolutional neural networks (FCNs) [61]. As an improvement to the early CNNs, such as AlexNet, which was developed primarily for classification tasks, the FCNs have introduced skip connections between the non-adjacent layers, in an attempt to prevent information loss due to the non-linear layers such as pooling and drop-out. Additionally, the FCN approach aims to remove the fully connected layers from the architecture. Over iterations, some of the drawbacks of the newly proposed concept have been addressed, leaving the current state of the art at the network architecture called DeepLabv3+, developed by a group of Google researchers [64].

The advantages of this approach were tested in the robotic harvesting setup. The experimental results achieved with a sweet pepper picking method based on object detection, though very promising, showed there is room for improvement, particularly in accuracy and robustness of pepper pose estimation. During experiments, it was observed that the 6DOF estimation step occasionally failed. Namely, in this phase, a cylinder primitive is fitted to the point cloud representing the target pepper. This point cloud is obtained by filtering the point cloud of the experiment scene using the object detection bounding box. The drawback of this method is that the bounding box, in addition to the detected pepper, includes other scene components as well, most often leaves. During cylinder fitting, points representing

occluding leaves in the bounding box occasionally deteriorated the cylinder fitting process, resulting in wrong estimates of pepper shape, size, and position and orientation. The initial idea would be to replace object detection with semantic segmentation, and only fit the geometric primitive to the point corresponding to the fruit.

As in other deep learning problems, the availability of training data is one of the most significant problems in semantic segmentation model development. Though elsewhere promising, synthetic dataset generation has been shown as limited in this particular application. More precisely, due to the texture bias in semantic segmentation learning, the networks only generalise well if the synthetic textures are a realistic representation of the real ones. This issue prevents the synthetic data approach from being applied to objects without visual textures, such as smooth and glossy fruit and vegetable peels. Another method was developed instead as a workaround, in which semantic segmentation is applied to those objects that are easier for the model to learn, namely the leaves. The real plant leaves of the species considered in this work are rich in texture, and therefore a good candidate to (pre)train a segmentation network. A DeepLabV3+ network architecture, pretrained on the ImageNet dataset, was trained via transfer learning using a synthetic dataset of leaves with realistic textures. The network was trained in the TensorFlow 2 environment, and all of the network weights were tuned during training. The input images were set to 640×640. The network was trained in batches of 4, up to a maximum of 200 epochs. Early stopping was used as a regularisation method, with Adam as optimizer on categorical cross-entropy. The trained pixel classes were "background" and "leaf", and their weights were adjusted to account for their unbalanced distribution. Finally, the model trained exclusively on the synthetic dataset reached accuracy of 63% on the test set, which contained only real images of sweet peppers (Fig. 3.9).

This model was then used to improve the pepper picking method, by introducing an additional filtering step in the pose estimation. Namely, the pepper subcloud, obtained by filtering the organised point cloud using the detection bounding box, is then additionally filtered using the segmentation mask obtained with the trained DeepLabv3+ model. This procedure is shown in Fig. 3.10. 6DOF pose estimation is then conducted as before, relying on geometric primitive fitting. In order to compare the behaviour of the two methods, 10 experiments were conducted both with and without segmentation, for evaluation of pose estimation precision on highly occluded peppers. In all of the repetitions, the same initial robot poses were used in the perception phase. The precision results are given in Table 3.4, providing the mean value over experiments (where applicable), as well as standard deviations of the estimated variables, for a single pepper in both cases. Without having the ground truth, the dissipation of estimates over repetitions is considered as the precision measure. While dissipation of variables in the horizontal plane is comparable in both cases, the estimation is significantly improved regarding the vertical, z axis, i.e. the main pepper growth axis. With an improved estimate of the pepper length, its centroid position benefits from a better estimation as well, resulting in more precise gripper positioning during grasping. Even better performance is expected with an improved segmentation model, since with 63% accuracy on the test set, there is quite a lot of room for improvement.

Figure 3.9: RealSense D435 RGB image of a scene with a pepper plant, and a corresponding DeepLabV3+ output semantic mask of the leaf pixels.

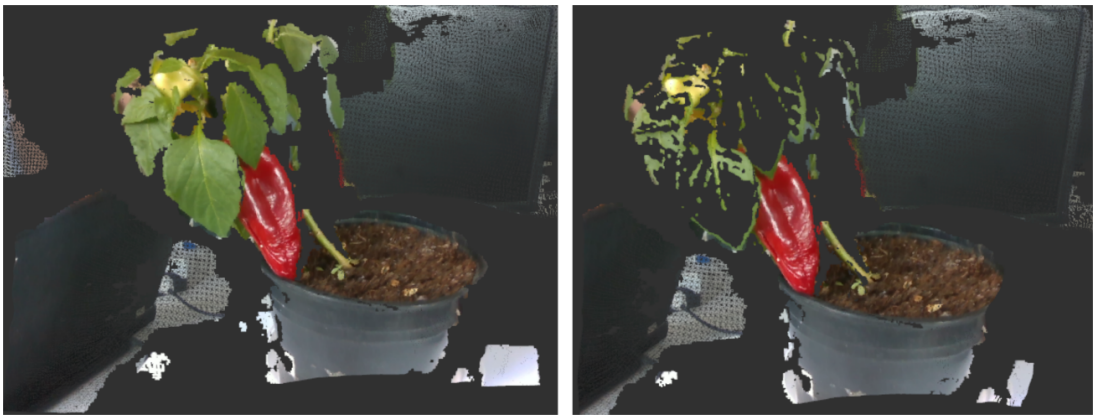


Figure 3.10: Snapshots of the complete 3D point cloud of the sweet pepper plant on the right, and a filtered 3D cloud using a leaf segmentation mask on the right.

Table 3.4: Pepper detection precision results

	Without leaf segmentation		With leaf segmentation	
	mean	std	mean	std
x [mm]		5.2		3.4
y [mm]		7.1		7.6
z [mm]		12.6		7.2
r [mm]	25.5	5.6	23.0	6.2
h [mm]	109.6	17.4	107.9	8.9

3.3 Contributions

The main research goal of the thesis was to develop an intelligent and compliant robotic system capable of autonomous execution of tasks in indoor agriculture. This was accomplished through three main contributions related to three respective enablers of such a system. The first one discussed here is the intelligent perception layer, addressed in the attached publication Pub.1.

C1: *Method for generation of semantic 3D models of plants suitable for soft robotic manipulation planning in agrotechnical procedures*

Execution of various manipulation tasks relies on initial reference inputs regarding robot positioning and force application. Starting from the classical hard coded position (in joint or Cartesian space), different approaches have developed through servoing based gradual approach, to complex planning techniques relying on modern hardware, and corresponding efficient computation and inference methods. The latter is the approach explored and exploited within this thesis, on a variety of agricultural applications. The contribution is realised through experimental validation of the following hypothesis:

H1: *A robot manipulator equipped with a sensory apparatus can be programmed to autonomously generate a 3D plant model fusing multiple measurements taken from different sensors and robot poses.*

A complex perception framework was constructed based on a robotic manipulator equipped with an RGB-D device in an eye-in-hand configuration. The perception framework relies on recording of the working environment from several perspectives, acquiring both colour and 3D data, and applying suitable inference methods for generation of a semantically rich model of the considered plant. Using detection and segmentation results from image data, 3D models can be labelled and used in target localization and manipulation planning. The model has been successfully applied for detection and segmentation of important functional parts of the plant, including fruit, flower, and leaves, as well as the relevant supporting environment such as soil. As shown in the attached publication Pub.1and in the following chapter, the contribution was realised in its full scope through successful execution of various robotic manipulation tasks based on this plant model.

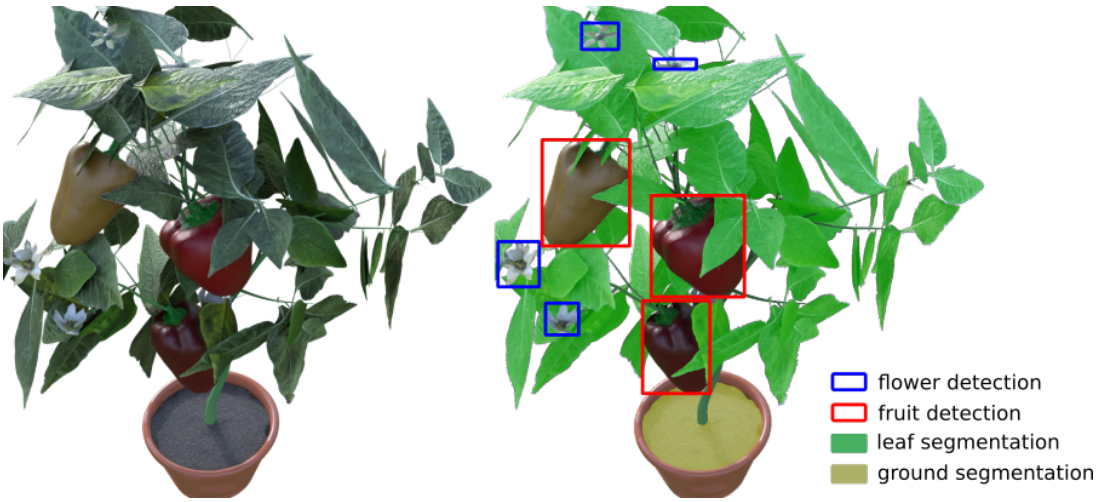


Figure 3.11: Detection of functional plant parts in 2D images used for 3D model generation.

4

Compliant robot control

In handling fragile objects such as plants, classical position based robot control is rarely sufficient for successful harmless task execution. Two fundamental approaches have emerged as control alternatives for the robot engineers, depending on the available hardware: enabling position controlled robots for compliant interaction with external force/torque (F/T) sensory setup, or development of position control frameworks for the collaborative robots. The latter platform has been exploited in the development of robotic systems for various agrotechnical procedures.

The first scenario described in the previous chapter and in the attached publication Pub.1 relies on position based motion control. These control problems dealt with safety in manipulation only implicitly, namely minimising the probability of collision during the planning phase only. The following problems, presented in this chapter, consider safety in manipulation more explicitly, relying on compliance in the closed control loop. As an extension of the position control interface, the first example is the soil moisture measurement task, based on an extended impedance controller. The second compliant control example, used in the plant hygiene procedure of mechanical stem cleaning, is described in detail in the attached publication Pub.2 on the example of the vine plant.

4.1 THEORETICAL BACKGROUND

The classical industrial manipulators were designed having in mind most usual industrial requirements, such as precise positioning and high repeatability, while the applied force is not strictly considered. This is mostly due to characteristics of the traditionally chosen automation procedures, for example pick and place, where the exact amount of exerted force is not as relevant as long as the procedure is executed at the precise position. The technological progress inspired deployment in other applications as well, including more complex tasks and environments where the amount of applied force is more relevant. This can be due to less structured working environments, more complex manipulation tasks, or handling fragile objects and sensitive materials. In such interactions, both the robot position and the contact force have to be controlled.

An opposite approach was taken in the development of robots with direct force control interface. These robots can easily be controlled to achieve a desired interaction force, but the problem arises with controlling the motion through free space. The impedance control concept enables both these robot drive types to achieve simultaneous force and position

control with a single low-level control interface. Usually, the position controlled manipulators are controlled through admittance controllers, and the force controlled robots through impedance controllers.

4.1.1 *Impedance control*

Compliant behaviour can be attained with position based robots, with additional sensors and a compliant control framework. One common approach relies on position based impedance control, often called admittance control. This control concept relies on a multiple input filter that allows simultaneous control of both position, and contact forces, through the same control interface on the joint level. Namely, the position control interface of the robot is referenced with an input calculated taking into account the force feedback signal as well, by satisfying the impedance filter dynamic equation. In this work, impedant behaviour is analysed in the Cartesian space, for manipulation and interaction of the robot end-effector with the environment. A similar concept exists at the joint level, where joint position impedant control can be implemented for each joint. In this case, local joint torques are needed, which is rarely available with classical industrial manipulators. This approach is therefore usually only possible with collaborative robots.

The Cartesian impedance controller is tuned to achieve a desired dynamic behaviour of the robot in interaction with the environment. During this interaction the environment is most commonly considered as a linear spring, and described with a linear approximation as in eq. 4.1:

$$\mathbf{F} = K_e(\mathbf{X} - \mathbf{X}_e), \tag{4.1}$$

where $\mathbf{X}$ is an $n \times 1$ measured robot position vector, and $\mathbf{X}_e \in R^{n\times 1}$ is the position vector of the environment in an unexcited state. With the linear spring approximation, the environment is described with a constant diagonal stiffness matrix $K_e \in R^{n\times n}$. $\mathbf{F} \in R^{n\times 1}$ then represents the force generated by the compressed environment along all axes of interest, in case the robot end effector penetrates into the environment. The displacement between the measured robot position and the unexcited environment position, i.e. the elasticity of the environment generates a force that acts on the robot. In other words, $\mathbf{F}$ is the measured contact force, as a result of the linear spring compression.

The impedance filter enables describing the entire robot-environment interaction system with an equivalent spring as well. By defining the target impedance behaviour of the system, the second order linear system can be designed that enables setting the dynamic relationship between the robot position and the force tracking error $\mathbf{E} = \mathbf{F}_r - \mathbf{F}$, so that it mimics a mass-spring-damper system. The $n \times 1$ force reference vector $\mathbf{F}_r$ can be freely chosen by the user. The target impedance is given with

$$\mathbf{E} = M(\ddot{\mathbf{X}}_c - \ddot{\mathbf{X}}_r) + B(\dot{\mathbf{X}}_c - \dot{\mathbf{X}}_r) + K(\mathbf{X}_c - \mathbf{X}_r), \tag{4.2}$$

where M, B and K are the $n \times n$ mass, damping and stiffness matrices of the target impedance, respectively. These matrices, most often constant and diagonal, enable setting the desired target impedance to the system by arbitrarily specifying the matrix elements. The signals $\mathbf{X}_c$ and $\mathbf{X}_r$ are Cartesian vectors representing the commanded and reference position of the robot, respectively. The reference position is the one provided by the user, or by a higher

level controller. The commanded position, on the other hand, is the reference used as input for the robot lower level Cartesian position control. The commanded position can also be referenced directly to the joint position control, using inverse kinematics or Jacobian.

The behaviour of the impedance filter enables smooth control of the robot both in contact with the environment, and during free-space motion. In this case, when there is no measured contact force and the reference is zero, there is no force tracking error $\mathbf{E} = \mathbf{o}$. The commanded position $\mathbf{X}_c$ then tracks the reference position $\mathbf{X}_r$ accurately, filtered through the dynamics specified with (4.2). On the other hand, when the robot comes into contact with the environment, the force sensors measure some contact force $\mathbf{F}$. This affects the robot position command $\mathbf{X}_c$, and, with a constant reference position $\mathbf{X}_r$, the position tracking cannot be accurate any longer. In other words, the impedance filter aims to balance between the position and force tracking errors, depending on the desired behaviour, specified through the filter parameters.

4.1.2 *Direct force control*

The most common alternative to the position controlled manipulators are the robots with torque control interfaces at joint levels. The torque control at the joint level usually relies on the internal torque feedback signal, which in turn enables design of inherently safe motion controllers with internal sensory apparatus used for collision detection. The internal joint torque measurements can also be used for estimation of external force vector $\mathbf{F}_{ext}$, under the assumption that the measured forces are the result of a single point of contact, or more precisely, of a contact with a single robot link. In the derivation, the body Jacobian matrix is used, i.e. the $6 \times n$ matrix relating infinitesimal joint displacements $d\mathbf{q}$ to infinitesimal end-effector displacements $d\mathbf{p}$ as defined by eq. 4.3:

$$d\mathbf{p} = \mathbf{J} d\mathbf{q}. \tag{4.3}$$

The relation between the joint torques and the external forces can easily be obtained using the principle of virtual work. This principle considers infinitesimal virtual displacements, which only satisfy geometric constraints, and do not have to meet other laws of motion. Neglecting effects such as gravity and moments resulting from motion of the rigid body, the virtual work $\delta\mathbf{W}$ vanishes in the equilibrium state, when considering the external forces on the end effector, and the joint torques on the joints, as given in eq. 4.4:

$$\delta\mathbf{W} = \boldsymbol{\tau}^\top \delta\mathbf{q} - \mathbf{F}^\top \delta\mathbf{p} = 0, \tag{4.4}$$

where $\boldsymbol{\tau} \in n \times 1$ are the joint torques, $\mathbf{F} \in R^{6\times1}$ are the external forces acting on the end effector, $\delta\mathbf{q} = (\delta q_1, \ldots \delta q_n)^\top$ are the virtual displacements of individual joints, $\delta\mathbf{p}$ are the virtual displacements at the end-effector. By applying the definition of body Jacobian 4.3 to the virtual displacements, and substituting the end-effector displacements in 4.4, the approximate relation of external forces and joint torques can be derived:

$$\boldsymbol{\tau} = \mathbf{J}^\top \cdot \mathbf{F}. \tag{4.5}$$

As a base of direct force control for the robot manipulators with joint torque control interface, the relation eq. 4.5 can be used by setting the reference joint torques according to the desired

external force at the end-effector. This relation can also be used in the opposite direction, for estimation of external forces acting on the end-effector, based on the measured joint torques and the current manipulator configuration. For practical applications, the geometric approximation of the principle of virtual work does not hold, and external factors should be accounted for. The external force estimation in practice attains the form given with eq. 4.6

$$\begin{aligned} \mathbf{F}_{ext} &= (\mathbf{J}^{\top})^{-1} \cdot \boldsymbol{\tau}_{ext} \\ &= (\mathbf{J}^{\top})^{-1} \cdot (\boldsymbol{\tau}_{meas} - \boldsymbol{\tau}_{gravity} - \boldsymbol{\tau}_{coriolis}), \end{aligned} \tag{4.6}$$

where the term $\boldsymbol{\tau}_{gravity}$ corresponds to the measured joint torques resulting from the effect of gravitational forces on the robot links, and the term $\boldsymbol{\tau}_{coriolis}$ similarly corresponds to the inertial moments resulting from the robot motion. Similar to the Jacobian calculation, these terms are calculated in each operating point based on the robot kinematic and dynamic model.

4.2 APPLICATIONS

4.2.1 *Soil moisture measurement*

As discussed in section 3.2.1, precision irrigation is gaining an increased focus of the research community. The measurement procedure itself is quite simple, but time consuming nonetheless in case of a large number of samples, such as the case considered in this thesis. The moisture sensor probes measuring the electrical conductivity need to be immersed into the tested medium for a short time period (several seconds) in order to provide enough measurement samples for humidity estimation. In the setup where this is done with an autonomous robot, this means that the robot manipulator is engaged in an invasive interaction with the object, namely, it has to realise a force-controlled interaction with its environment. A sensory and control framework is developed for a robot manipulator, relying on the joint position control interface, that realises the compliant manipulation task through an adaptive impedance controller framework.

The controller is developed as an extension of the classic position based impedance controller. It can be shown that the robot-environment interaction with the desired contact forces can be realised in case both precise environment position and the environment equivalent stiffness is known, by generating an adequate reference position trajectory. Namely, following the steps in [123], the optimal position reference x_r can be calculated using eq. 4.7

$$x_r = \frac{F_r}{k_e} + x_e, \tag{4.7}$$

where F_r is the desired contact force, and k_e and x_e are the environment stiffness and position, respectively. However, the environmental parameters are in practice never known precisely, and even small errors in the environment parameters k_e and x_e can result in large force errors. Furthermore, in the case of soil moisture measurements, the soil equivalent stiffness varies with the soil moisture as well, and a precise estimate is impractical to obtain a-priori. One possible solution to this issue is the estimation of these parameters. In the experimental validation of automation of this agrotechnical procedure, a position impedance

control method developed in [124] is deployed with online adaptation of the impedance filter inputs, based on the estimated environment stiffness. The experimental results demonstrate successful position and contact force tracking, with the environment stiffness unknown a-priori, and the environment position estimated as described in Chapter 3.2.1. The adaptation law is derived for the position reference, based on the adaptive parameter $\kappa(t)$ that accounts for unknown elastic properties of environment under external force,

$$x_r(t) = \kappa(t)F_r + x_e, \tag{4.8}$$

where the position reference is a function of the estimate of the initial position of the environment x_e and the force reference F_r. The exact adaptation law is given with eq. 4.9 for one spatial dimension,

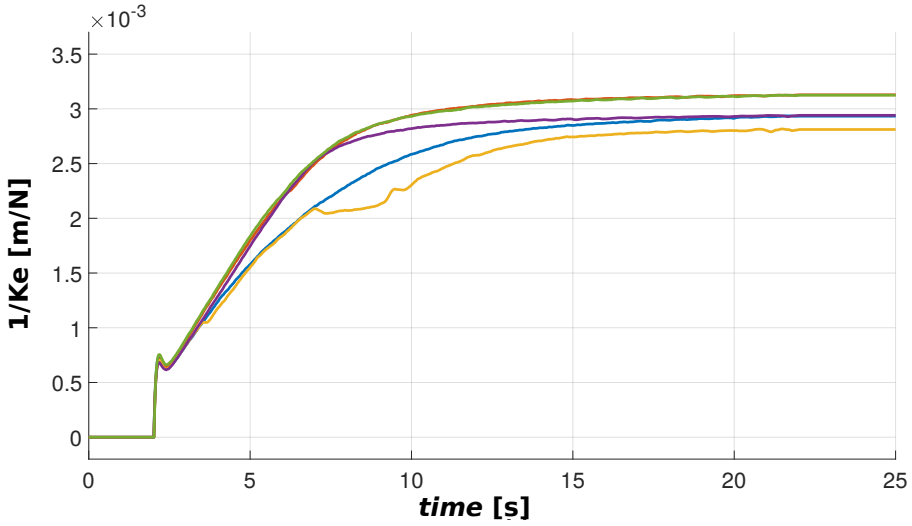
$$\begin{aligned} k\dot{\kappa}(t) + b\ddot{\kappa}(t) + m\dddot{\kappa}(t) &= -\gamma_1 q(t) + \gamma_1^* \dot{q}(t), \\ q(t) &= p_1 e(t) + p_2 \dot{e}(t), \end{aligned} \tag{4.9}$$

where k, b, and m are the impedance filter parameters, $e(t)$ is the force tracking error, and p_1 and p_2 are the free parameters tuned based on the particular application. The derivation and convergence proof for this adaptation law can be found in [124].

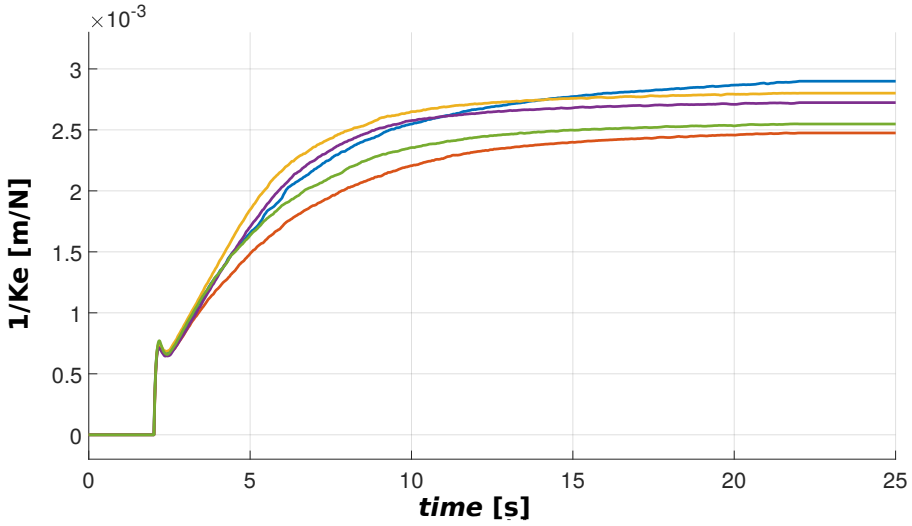
Within this thesis, this control concept is experimentally validated, within the complete soil moisture measurement framework together with the perception from chapter 3. The experimental results show successful manipulation of objects of variable and unknown stiffness, ranging from soft, wet ground, to handling collisions with rigid object such as roots or stones. The experiments were conducted in three scenarios, analysing the behaviour in case of moist soil conditions, representing the softest scenario, the dry soil conditions, representing a stiff scenario, and the extreme as a collision with a very rigid object, representing e.g. a stone in the soil that could potentially break the sensory equipment, or a part of the root system that should not be harmed. Five experimental repetitions were conducted for the first two scenarios, and three repetitions for the collision scenario. The impedance filter was the same in all of the experiments, tuned for compliant and safe behaviour on a rigid object.

The variability in the manipulated object mechanical properties was then overcome through the adaptation law that inherently models the stiffness of the manipulated object, as shown in 4.1a-4.1c. The responses show the dynamics of the environment compliance, $1/K_e$ (the inverse of mechanical stiffness). For safety reasons, the initial assumption is that the manipulated object is infinitely stiff (zero compliance), and the adaptation of the estimated stiffness gradually reaches the actual value along with the desired contact forces. The adaptation is therefore faster for a stiffer object. It should be noted that the compliance (stiffness) measure not only models the soil, but takes into account the elasticity of the robot manipulator, as well as the imprecision of the soil surface detection. Regardless, the results show that throughout the repetitions, the estimation converges to the same region of values. This not only enables precise force reference tracking, but also shows that the adaptation method is stable with respect to robot dynamics and detection imprecision, and can be used for feature rich modelling of the plant.

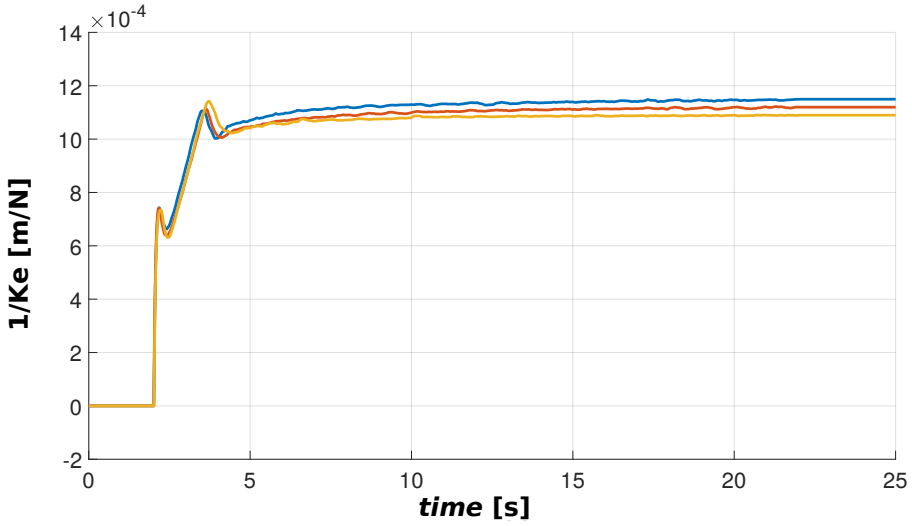
The adaptation law enabled precise force tracking as shown in the graphs in figures 4.2a-4.2c. The measured contact forces are provided by the Franka Panda dynamics estimation



(a)



(b)



(c)

Figure 4.1: Adaptation dynamics during experiments with moist soil4.1a, dry soil4.1b, and in case of collision with a rigid object4.1c. The estimation also compensates manipulator dynamics and surface detection pipeline. Adaptation tuned for fastest convergence on stiffest objects.

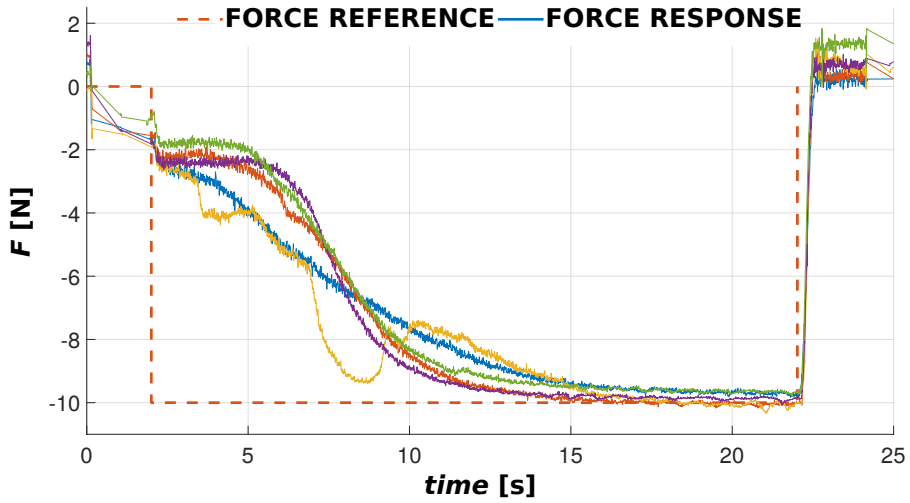
model, which is prone to error due to imprecision in the model. This can be observed particularly in the variable baseline offset in the measurements at the beginning and at the end of each experiment. The expected measured force here is zero because the robot is not in contact with the environment, and there are no external forces acting on the end-effector. However, the estimated forces are non-zero and vary depending on the robot pose and velocity. This could easily be overcome by using an external, more precise force/torque sensor, instead of the cobot estimates.

4.2.2 *Plant stem exploration*

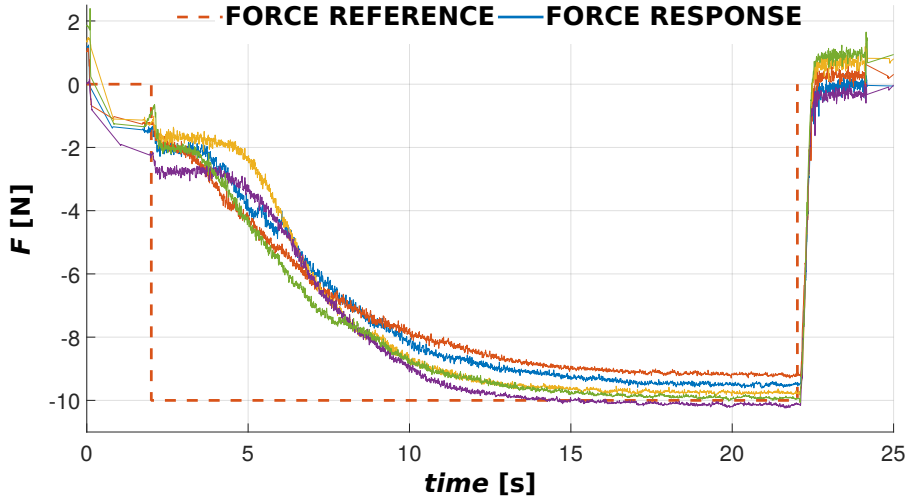
Various agricultural procedures require sensing and control of the applied force during procedure execution. One example application, still mostly conducted by human workers, is the stem cleaning procedure, in which unwanted sprouts, buds and old rind are removed. This cleaning procedure is periodically conducted on various crops and cultures, either by precise pinching, or by scraping larger portions of the plant stem. Within this thesis, a method is developed for vine plant treatment. Relying on the capabilities of a collaborative Franka Emika Panda robot, a robot motion controller is developed based on the concepts of impedance filtering. The impedant motion controller incorporates an exploration-driving signal, that enables compliant exploration of the plant stem and generation of a 3D spatial model suitable for manipulation in the cleaning procedure.

The method is built upon the assumption that the sprouting position is known through another perception module, such as those discussed in Chapter 3. This module provides a starting position for the robot arm, positioning the robot tool with a loose grip around the plant stem. The exploration procedure yields a point cloud of recorded points, i.e. of the traversed path of the robot tool. Unsupervised learning is deployed on this data, and a crude 3D model of the plant is generated, precise enough for execution of manipulation trajectory with a tight grip on the stem for cleaning.

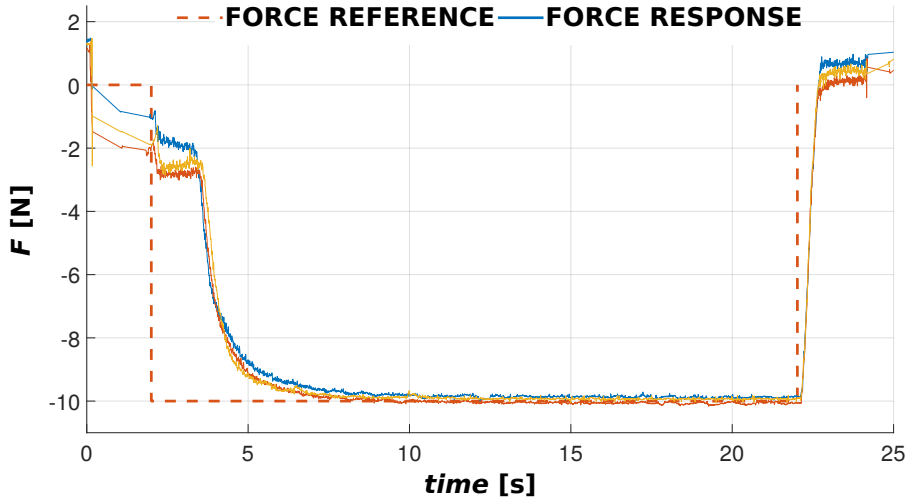
There are several important enablers for this compliant exploration approach described in the attached publication Pub.2. First, the compliant robot manipulator is controlled through the low level joint torque interface, based on eq. 4.5. The reference force is provided by the impedance filter, developed from its basic form as in eq. 4.2. The inputs to the filter are provided from multiple sources: robot provides the measured position and contact force of the end effector, the desired force is kept at zero during the exploration procedure, and the exploration driving signal is generated with the exploration module that incorporates task information. This system is sufficient for control of the end effector position. However, several factors including the force estimation method and the lever-effect of the robot tool deteriorate the sensing and control of the robot tool orientation. A solution to this was developed in the form of a local curvature estimation model. In all, the impedance filter generates a reference force torque signal, that models the net effect of compensating for unwanted contacts (desired measured forces are zero), and incorporates a virtual force driving the exploration along the plant stem. The reconstruction of the vine plant model is presented in the attached publication Pub.2.



(a)



(b)



(c)

Figure 4.2: Force tracking of the robot end effector in experiments on the softest object, moist soil 4.2a, on the stiffer object, namely dry soil4.2b, and in case of collision with a rigid object 4.2c. The robot motion is safe for the manipulated objects regardless of their stiffness.

5

Tactile sensing

Arguably the most important step in enabling robots to reach human dexterity is reproducing the capabilities of the human tactile sensing [125]. Tactile feedback is essential for complex precise manipulation tasks as well as for safe human–robot interaction [107]. Different technologies are employed in trying to develop a sensor that can adequately balance the quality and usability of acquired sensory information with a preferably compliant mechanical structure of the sensor body that satisfies safety requirements of close human-robot interaction. The quality of information depends on the particular task, but can in general be qualitatively described as the ability to achieve the desired performance, e.g. object recognition and exploration, grasp stability estimation and slip detection, force control, or tactile servoing [103]. Typically, tactile sensing refers to (mainly) indirect measurements of force/torque and/or displacement.

An optical based family of tactile sensors TacTip is a technology in the focus of the contribution related to tactile sensing within this thesis. One of the reasons behind choosing this particular technology is in the information rich format of the raw sensor data, namely the RGB camera raw image output. Though lower complexity technologies, for example capacitive contact sensors, could provide a higher measurement rate, such technologies could hardly ever reach this sensing resolution. The contemporary cameras are however still being developed with their frame rates increasing, and are already at a frame rate high enough for closed loop robot control. It was repeatedly shown that the high dimensionality of the raw sensor output is best tackled using black box based approaches, such as machine or deep learning. In case of CNN inference, the forward pass is currently the computational bottleneck in terms of measurement rate, until embedded hardware reaches the capabilities of the powerful state of the art GPUs.

5.1 Theoretical background

Deep learning based approach is optimal for the described system characteristics, because such black box approaches can be trained to generalise over the system inconsistencies, and the modern approaches relying on transfer learning enable fast deployment for new sensors. Within this thesis, fast deployment was tackled with dimensionality reduction, again through deep learning. For this family of sensors, the camera output image can be considered raw output. This raw output is inherently of high dimensionality, i.e. a large number of correlated features. The CNN approach is inherently designed to tackle this

kind of spatial information. A step further is taken in exploiting these capabilities, through the concept of autoencoder networks. These neural networks are designed with an input, encoder part, and an output, decoder part, which are together trained to reconstruct the input of the network at the output side. The catch is in the architecture of the deep layers, i.e. in the reduced number of features in the central part of the network, as shown by schematic in Fig. 5.1. By training the network to reconstruct the high-dimensional output from the reduced number of features, these features, sometimes called *latent space representation*, are extracted, such that they carry (most of) the information contained within the raw input.

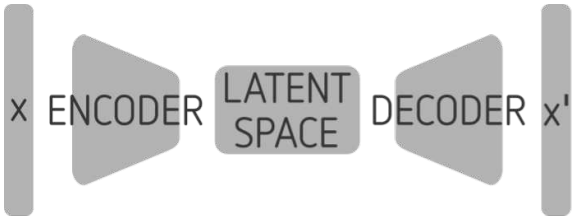


Figure 5.1: Autoencoder network architecture schematic.

The approach proposed within this thesis and discussed in the attached publication Pub.3 relies on the features extracted using the trained encoder, which essentially becomes a part of the sensor, meaning that the raw sensor output always goes through the encoder block, in any inference pipeline. The output of the encoder is ideally a small set of orthogonal features. Then, any black box can come in place of the decoder, trained on the image features for estimation of a measured variable of interest, such as position, force, or certain characteristics of the manipulated object. The same set of features can ideally be applied to a variety of use cases, on different instances of the same sensor with the encoder trained on one instance only, and the same features used for training inference models on another sensor instance.

The encoder as feature extractor can be applied to other sensor shapes as well, such as planar, cylindrical, or other more complex surface geometries, with similar signal transduction principles. One such example could be a miniaturised version of the TacTip, similar to the design presented in [126]. This miniature sensor is designed and developed with the idea of mounting it on a robot end effector as a soft finger with tactile sensing capabilities. With the same working principle as in the original spherical TacTip, this design allows inference of similar variables.

5.2 APPLICATIONS

The presented concept is not restricted to the family of small, bioinspired tactile sensors. On the contrary, as explained in the work in the attached publication Pub.4, a similar approach can be applied to fundamentally different applications, such as robotic plastering with a deformable plastering tool. The soft plastering tool, a flexible knife, applies and spreads a coat of plaster to the plastered surface by application of force. The application of the plaster mass depends on the applied force, and the shape of the plastering tool, i.e. the angle between the tip of the plastering knife and the plastered surface. In order to estimate and control these variables during the plastering procedure, the plastering knife surface

is enriched with visual cues inspired by the TacTip design, as shown in the schematic in Fig.5.2. A camera is mounted on the robot end-effector, recording the plastering knife. The knife is colored into an easily distinguished colour, and a pattern of dots is printed on top to provide the camera with richer information. The results show that a CNN trained through transfer learning can successfully be applied in this control problem, for a single variable control. Since the knife tool can be distorted into complex 3D shapes, the CNN approach enables inference of more complex variables, such as 3D deformations and 6DOF force and torque estimation, provided an appropriate training dataset.

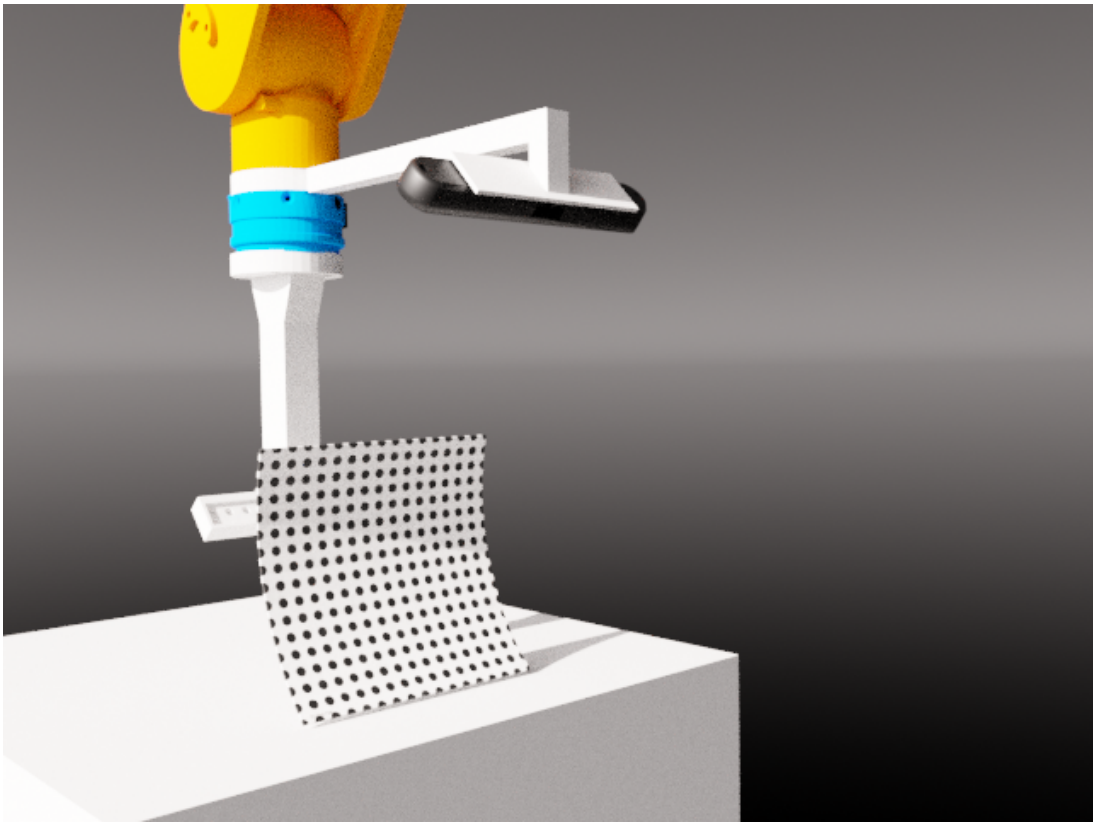


Figure 5.2: The plastering knife tool enriched with visual cues, becoming a sensory end effector as well.

5.3 Contributions

Availability of contact feedback is a fundamental prerequisite for intelligent and compliant robot manipulation. In its most basic form in compliant control frameworks, this is usually in the form of contact force and torque measurements. More recent approaches rely on richer tactile information, acquired by different types of sensors. The third and final contribution of this thesis considers this sensing modality in relation to safe execution of delicate manipulation tasks.

C3: *Tactile sensor adapted for compliant manipulation of fragile objects based on artificial intelligence.*

Tactile sensing technology considered in this thesis is inspired by the TacTip family of sensors, where contact information is encoded in the pattern displacement information. As

an advancement of existing solutions, processing of raw camera output is proposed using the deep learning approach. More precisely, instead of a classical approach the work developed within this thesis proposes using a black box based dimensionality reduction method, using convolutional neural networks. The contribution is realised through experimental validation of the hypothesis:

H1: *It is possible to obtain contact information with a novel optical tactile sensor through machine learning models.*

The experimental validation of the hypothesis on the example of feature extraction using a convolutional encoder and transfer learning in different tactile sensing applications resulted in improvement in the state of the art of optical based tactile sensing. The variability in sensor shapes and application tasks can be considered a demonstration of versatility of the proposed improvement, and its applicability to other, novel tactile sensing shapes to be developed in the future.

6

Conclusion

Review of state of the art solutions and challenges in the field of agricultural robotization and automation clearly shows that precise object detection and positioning is essential for various agricultural procedures. The most common approach in processing visual feedback is based on CNNs, thanks to the widespread use and development of these models over various application fields, and the possibility to reuse and adjust existing results for fast re-deployment. This is usually done through the concept of transfer learning. Namely, over the last decade, several prominent deep CNN architectures have been developed with an immense amount of parameters that managed to learn complex detection problems, however at the cost of expensive hardware resources, long training procedures, and for large dataset generation. Transfer learning is then deployed to reuse the training results, i.e. network weights tuned on general object detection or classification tasks. In transfer learning, most of the network weights are kept at their pretrained values. The last several layers are removed, and new layers are added and trained for particular task. In detection problems, the size of the last layer of the network depends on the number of classes to be detected. For example, the same network could be trained to simultaneously detect different plant parts such as fruit, leaves, flowers, etc. The additional fine tuning can either be conducted on the whole network, or only for a subset of parameters, and a smaller number of iterations is usually sufficient to learn new problems as well. As the need for the large datasets is one of the biggest obstacles in the development of the deep learning based perception methods, the results obtained as a part of this thesis partially rely on synthetic data generation, which alleviates the cost of the time consuming data annotation. This is especially relevant for the semantic and instance segmentation tasks, where manual labelling is significantly more time consuming and tedious than for the object detection task.

Among several established CNN architectures, the Single Shot Detector (SSD) approach is chosen for these applications. The SSD networks are capable of fast simultaneous detection of multiple objects in a single shot, especially with light-weight backbones such as the MobileNet architecture, which was designed for fast computation on embedded and mobile devices (hence the name). Within SpECULARIA, instead of deployment on embedded devices, this architecture was chosen mainly thanks to the possibility of real time deployment in closed control loops. Transfer learning was used in all applications, with TensorFlow implementation of MobileNet SSD network pretrained on the COCO dataset. The perception results presented in this thesis are not an end in itself, but instead only an

intermediate step in execution of different agricultural procedures. The problems tackled within this thesis range from harvesting and pollination, to applications relying on physical contact such as soil moisture measurement and generation of the plant stem model through tactile exploration. The latter type of applications rely on measurement and control of the contact forces occurring during manipulation.

Contact force control is an open research question, tackled in several ways depending on task requirements and available hardware. One of the popular approaches relies on an impedance framework that enables tracking of both positional and force references. Thanks to the versatility of the method, it is easily adapted to both classical position controlled robot manipulators and to collaborative robots. The results obtained within this work can be considered as one of the first steps for introduction of affordable and safe robots into the food production industry. Not only are cobots safe for human co-workers and generally less expensive than the industrial manipulators, but also more convenient for non-experts, e.g. through programming by demonstration. The considered manipulation examples were constrained to plant exploration and stem cleaning, and soil moisture measurements. The most straight-forward extensions within the part of the future work are harvesting and pruning tasks, through development of compliant motion algorithms that exert a necessary contact force where needed, and guard the other plant parts from harm elsewhere. The soil moisture measurement can easily be extended with soil aeration procedure, even using the same or similar tools. All of these examples however consider imitation of human workers by the robot manipulators, relying on estimation of contact forces using rigid measurement equipment. In order to fully replace human labour in these strenuous and repetitive, but very delicate tasks, the robots should be equipped with a sensory setup with capabilities more similar to the natural ones in humans.

In an attempt to enable delicate interaction with fragile environments, an entire field of robotics research has emerged in recent years, developing nature inspired robotic structures. The interesting field, rich in scientific outputs of soft actuators, could find various applications in the agricultural domain, but due to the complexity of this field, this was outside the scope of this thesis. Instead, soft sensing components are considered only. Nature inspired design of such components yields high-dimensional raw output data, and in theory enables imitation of certain aspects of human tactile sensing capabilities to a very high degree. For example, modern technological solutions enable high resolution sensing of spatial information (displacements), which is related to high sensitivity to contact forces. Other sensing modalities, such as thermal, can in some cases easily be introduced, but are out of scope of this thesis as well. The high sensitivity comes at a cost of high dimensional sensor data. In application scenarios where different contact information is simultaneously needed for robot control, using separate complex models for each type of measurement is costly both in terms of communication (transporting large raw data), and computation (measurement estimation for the big data). A relevant example can be found in simultaneous estimation of both the contact force, and local spatial features of the manipulated object. As an alternative to the separate estimation models per estimated contact variable is using a single point dimensionality reduction. This approach relies on disseminating the lower dimensionality data to the smaller and simpler (faster) separate estimation models. Such an approach is proposed within this thesis, relying on the autoencoder neural network structure. This

technique is well known and broadly applied across a variety of other inference fields, and has proved promising in tactile sensing as well. Not only does it reduce online runtime in estimation, but it also allows faster deployment of new sensor instances. Namely, it was experimentally shown that this approach is capable of generalising over the sensor instances built inherently with specific differences. The feature extraction generalisation also promises fast development of other sensor shapes and modalities, and finally their introduction into the robot control framework. As a part of the future work, the developed perception models will be deployed in exploration and manipulation tasks. The contact force measurements will be used for local object exploration, resulting in an object point cloud map coding local curvature and stiffness within the occupancy grid. This in turn will allow a compliant (stiffness-aware) approach in manipulator trajectory planning. Another line of work will be directed towards manipulator control algorithm design that takes advantage of simultaneous availability of various tactile information.

7

List of publications

Pub1 M. Polić, J. Tabak and M. Orsag. Pepper To Fall: A Perception Method For Sweet Pepper Robotic Harvesting. Intelligent Service Robotics, 35(3-4):130-140, 2021, IF: 2.246 (Q3).

Pub2 M.Polic, M.Car, F.Petric, and M.Orsag Compliant Plant Exploration for Agricultural Procedures With a Collaborative Robot. IEEE Robotics and Automation Letters, 6(2), 2768-2774, 2021, IF: 3.741 (Q2).

Pub3 M.Polic, I.Krajacic, N.Lepora, and M.Orsag Convolutional autoencoder for feature extraction in tactile sensing. IEEE Robotics and Automation Letters, 4(4), 3671-3678, 2019, IF: 3.6, (Q1).

Pub4 M.Polic, B.Maric, and M.Orsag Soft robotics approach to autonomous plastering. *IEEE International Conference on Automation Science and Engineering (CASE)*. Lyon, France, 482–487, 2021.

8

Author's contribution to publications

The results presented in this thesis are based on the research carried out in the Laboratory for robotics and intelligent control systems (LARICS), lead by Professors Zdenko Kovacic and Stjepan Bogdan, at the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia during the period of 2016 - 2021 as a part of a research project SpECULARIA - Structured Ecological CULtivation with Autonomous Robots In Agriculture, which was funded by the Croatian Science Foundation under the project Specularia UIP-2017-05-4042.

A part of the thesis also includes the research carried out at the Bristol Robotics Laboratory within the Tactile Robotics group lead by Professor Nathan Lepora. The collaboration was financially supported by the British Scholarship Trust.

The thesis includes four publications written in collaboration with co-authors of the published papers. The author's contribution to each paper consists of the method design, software implementation, testing in simulations and real world experiments, result analysis and written presentation.

Pub1 In the paper entitled *Pepper To Fall: A Perception Method For Sweet Pepper Robotic Harvesting* the author proposed a method for sweet pepper detection, localization and harvesting under controlled indoor cultivation conditions. The detection relies on a deep CNN, and is used as an intermediary step in the localisation pipeline. The localization combines the detection in 2D images with 3D data provided by the RGB-D camera. The author developed the detection and localization pipeline, and implemented the robot control software. The author conducted real world experiments together with other co-authors.

Pub2 In the paper entitled *Compliant Plant Exploration for Agricultural Procedures With a Collaborative Robot* the author proposed a compliant manipulator control method that utilises cobot capabilities in performing compliant and safe exploration motion. The exploration is defined in task space based on the tool design, and transformed into the robot joint space through a selection matrix and an impedance filter. The impedance filter aims to minimise contact forces during exploration, and drives the motion along unexplored parts of the plant stem. The author developed the compliant exploration framework, implemented the robot controllers, and performed exploration and manipulation experiments on a real vine plant stem.

Pub3 In the paper entitled *Convolutional autoencoder for feature extraction in tactile sensing*

the author proposed a feature extraction method for optical based tactile sensors. The method is used for input dimensionality reduction, and relies on a convolutional encoder structure. The author performed the dataset collection and encoder training on the training datasets. Along with other co-authors, the author conducted experimental validation on several use-cases, as well as benchmarking experiments against other methods.

Pub4 In the paper entitled *Soft robotics approach to autonomous plastering* the authors propose a novel tactile sensing inspired approach to robotic plastering. The plastering tool deformation is modelled using a CNN model, and used as an input for a compliant manipulator control framework. The author trained and implemented the CNN solution. The model was experimentally tested in a series of experiments, validating the framework in a pose control task, and its performance on a surface with disturbances.

BIBLIOGRAPHY

[1] Kalinowska, D., Wielgat, P., Kolerski, T., Zima, P., "Model of nutrient and pesticide outflow with surface water to puck bay (southern baltic sea)", Water, Vol. 12, No. 3, 2020, str. 809.

[2] Jepson, P. C., Murray, K., Bach, O., Bonilla, M. A., Neumeister, L., "Selection of pesticides to reduce human and environmental health risks: a global guideline and minimum pesticides list", The Lancet Planetary Health, Vol. 4, No. 2, 2020, str. e56–e63.

[3] Kondo, N., Monta, M., Noguchi, N., Agricultural robots: mechanisms and practice. Apollo Books, 2011.

[4] Kootstra, G., Wang, X., Blok, P. M., Hemming, J., Van Henten, E., "Selective harvesting robotics: current research, trends, and future directions", Current Robotics Reports, 2021, str. 1–10.

[5] van Herck, L., Kurtser, P., Wittemans, L., Edan, Y., "Crop design for improved robotic harvesting: A case study of sweet pepper harvesting", Biosystems Engineering, Vol. 192, 2020, str. 294–308.

[6] Vougioukas, S. G., "Agricultural robotics", Annual review of control, robotics, and autonomous systems, Vol. 2, 2019, str. 365–392.

[7] Champ, J., Mora-Fallas, A., Goëau, H., Mata-Montero, E., Bonnet, P., Joly, A., "Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots", Applications in plant sciences, Vol. 8, No. 7, 2020, str. e11373.

[8] Sivakumar, A. N., Modi, S., Gasparino, M. V., Ellis, C., Velasquez, A. E. B., Chowdhary, G., Gupta, S., "Learned visual navigation for under-canopy agricultural robots", arXiv preprint arXiv:2107.02792, 2021.

[9] Gu, Y., Li, Z., Zhang, Z., Li, J., Chen, L., "Path tracking control of field information-collecting robot based on improved convolutional neural network algorithm", Sensors, Vol. 20, No. 3, 2020, str. 797.

[10] Zhang, J., Chambers, A., Maeta, S., Bergerman, M., Singh, S., "3d perception for accurate row following: Methodology and results", in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013, str. 5306–5313.

[11] Durand-Petiteville, A., Le Flecher, E., Cadenat, V., Sentenac, T., Vougioukas, S., "Tree detection with low-cost three-dimensional sensors for autonomous navigation in orchards", IEEE Robotics and Automation Letters, Vol. 3, No. 4, 2018, str. 3876–3883.

[12] Sparrow, R., Howard, M., "Robots in agriculture: prospects, impacts, ethics, and policy", Precision Agriculture, Vol. 22, No. 3, 2021, str. 818–833.

[13] Bogue, R., "Fruit picking robots: has their time come?", Industrial Robot: the international journal of robotics research and application, 2020.

[14] Fountas, S., Mylonas, N., Malounas, I., Rodias, E., Hellmann Santos, C., Pekkeriet, E., "Agricultural robotics for field operations", Sensors, Vol. 20, No. 9, 2020, str. 2672.

[15] Fountas, S., Espejo-García, B., Kasimati, A., Mylonas, N., Darra, N., "The future of digital agriculture: technologies and opportunities", IT professional, Vol. 22, No. 1, 2020, str. 24–28.

[16] Zhang, B., Xie, Y., Zhou, J., Wang, K., Zhang, Z., "State-of-the-art robotic grippers, grasping and control strategies, as well as their applications in agricultural robots: A review", Computers and Electronics in Agriculture, Vol. 177, 2020, str. 105694.

[17] Yaguchi, H., Nagahama, K., Hasegawa, T., Inaba, M., "Development of an autonomous tomato harvesting robot with rotational plucking gripper", in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016, str. 652–657.

[18] Zhang, B., Zhou, J., Meng, Y., Zhang, N., Gu, B., Yan, Z., Idris, S. I., "Comparative study of mechanical damage caused by a two-finger tomato gripper with different robotic grasping patterns for harvesting robots", Biosystems engineering, Vol. 171, 2018, str. 245–257.

[19] Navas, E., Fernández, R., Sepúlveda, D., Armada, M., Gonzalez-de Santos, P., "Soft grippers for automatic crop harvesting: A review", Sensors, Vol. 21, No. 8, 2021, str. 2689.

[20] Liu, C.-H., Chiu, C.-H., Chen, T.-L., Pai, T.-Y., Chen, Y., Hsu, M.-C., "A soft robotic gripper module with 3d printed compliant fingers for grasping fruits", in 2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, 2018, str. 736–741.

[21] Silwal, A., Davidson, J. R., Karkee, M., Mo, C., Zhang, Q., Lewis, K., "Design, integration, and field evaluation of a robotic apple harvester", Journal of Field Robotics, Vol. 34, No. 6, 2017, str. 1140–1159.

[22] Hohimer, C. J., Wang, H., Bhusal, S., Miller, J., Mo, C., Karkee, M., "Design and field evaluation of a robotic apple harvesting system with a 3d-printed soft-robotic end-effector", Transactions of the ASABE, Vol. 62, No. 2, 2019, str. 405–414.

[23] Chen, S., Pang, Y., Yuan, H., Tan, X., Cao, C., "Smart soft actuators and grippers enabled by self-powered tribo-skins", Advanced Materials Technologies, Vol. 5, No. 4, 2020, str. 1901075.

[24] Arad, B., Kurtser, P., Barnea, E., Harel, B., Edan, Y., Ben-Shahar, O., "Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. the case study of sweet pepper robotic harvesting", Sensors, Vol. 19, No. 6, 2019, str. 1390.

[25] Zemmour, E., Kurtser, P., Edan, Y., "Automatic parameter tuning for adaptive thresholding in fruit detection", Sensors, Vol. 19, No. 9, 2019, str. 2130.

[26] Li, J., Huang, W., Zhao, C., "Machine vision technology for detecting the external defects of fruits—a review", The Imaging Science Journal, Vol. 63, No. 5, 2015, str. 241–251.

[27] Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., "Sensors and systems for fruit detection and localization: A review", Computers and Electronics in Agriculture, Vol. 116, 2015, str. 8–19.

[28] Zhao, Y., Gong, L., Huang, Y., Liu, C., "A review of key techniques of vision-based control for harvesting robot", Computers and Electronics in Agriculture, Vol. 127, 2016, str. 311–323.

[29] Zhang, J., Karkee, M., Zhang, Q., Zhang, X., Yaqoob, M., Fu, L., Wang, S., "Multi-class object detection using faster r-cnn and estimation of shaking locations for automated shake-and-catch apple harvesting", Computers and Electronics in Agriculture, Vol. 173, 2020, str. 105384.

[30] Fu, L., Gao, F., Wu, J., Li, R., Karkee, M., Zhang, Q., "Application of consumer rgb-d cameras for fruit detection and localization in field: A critical review", Computers and Electronics in Agriculture, Vol. 177, 2020, str. 105687.

[31] Nguyen, T. T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J. G., Saeys, W., "Detection of red and bicoloured apples on tree with an rgb-d camera", Biosystems Engineering, Vol. 146, 2016, str. 33–44.

[32] Perez, R. M., Cheein, F. A., Rosell-Polo, J. R., "Flexible system of multiple rgb-d sensors for measuring and classifying fruits in agri-food industry", Computers and Electronics in Agriculture, Vol. 139, 2017, str. 231–242.

[33] Andujar, D., Ribeiro, A., Fernández-Quintanilla, C., Dorado, J., "Using depth cameras to extract structural parameters to assess the growth state and yield of cauliflower crops", Computers and Electronics in Agriculture, Vol. 122, 2016, str. 67–73.

[34] Lin, G., Tang, Y., Zou, X., Xiong, J., Fang, Y., "Color-, depth-, and shape-based 3d fruit detection", Precision Agriculture, Vol. 21, No. 1, 2020, str. 1–17.

[35] Tu, S., Pang, J., Liu, H., Zhuang, N., Chen, Y., Zheng, C., Wan, H., Xue, Y., "Passion fruit detection and counting based on multiple scale faster r-cnn using rgb-d images", Precision Agriculture, Vol. 21, No. 5, 2020, str. 1072–1091.

[36] Wang, Z., Walsh, K. B., Verma, B., "On-tree mango fruit size estimation using rgb-d images", Sensors, Vol. 17, No. 12, 2017, str. 2738.

[37] Milella, A., Marani, R., Petitti, A., Reina, G., "In-field high throughput grapevine phenotyping with a consumer-grade depth camera", Computers and electronics in agriculture, Vol. 156, 2019, str. 293–306.

[38] Kang, H., Chen, C., "Fruit detection, segmentation and 3d visualisation of environments in apple orchards", Computers and Electronics in Agriculture, Vol. 171, 2020, str. 105302.

[39] Lehnert, C., English, A., McCool, C., Tow, A. W., Perez, T., "Autonomous sweet pepper harvesting for protected cropping systems", IEEE Robotics and Automation Letters, Vol. 2, No. 2, 2017, str. 872–879.

[40] Wang, J., He, J., Han, Y., Ouyang, C., Li, D., "An adaptive thresholding algorithm of field leaf image", Computers and electronics in agriculture, Vol. 96, 2013, str. 23–39.

[41] Jiang, J.-A., Chang, H.-Y., Wu, K.-H., Ouyang, C.-S., Yang, M.-M., Yang, E.-C., Chen, T.-W., Lin, T.-T., "An adaptive image segmentation algorithm for x-ray quarantine inspection of selected fruits", Computers and electronics in agriculture, Vol. 60, No. 2, 2008, str. 190–200.

[42] Ostovar, A., Ringdahl, O., Hellström, T., "Adaptive image thresholding of yellow peppers for a harvesting robot", Robotics, Vol. 7, No. 1, 2018, str. 11.

[43] Arad, B., Balendonck, J., Barth, R., Ben-Shahar, O., Edan, Y., Hellström, T., Hemming, J., Kurtser, P., Ringdahl, O., Tielen, T. *et al.*, "Development of a sweet pepper harvesting robot", Journal of Field Robotics, Vol. 37, No. 6, 2020, str. 1027–1039.

[44] Sa, I., Lehnert, C., English, A., McCool, C., Dayoub, F., Upcroft, B., Perez, T., "Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-d information", IEEE Robotics and Automation Letters, Vol. 2, No. 2, 2017, str. 765–772.

[45] Tian, Y., Duan, H., Luo, R., Zhang, Y., Jia, W., Lian, J., Zheng, Y., Ruan, C., Li, C., "Fast recognition and location of target fruit based on depth information", IEEE Access, Vol. 7, 2019, str. 170 553–170 563.

[46] Tao, Y., Zhou, J., "Automatic apple recognition based on the fusion of color and 3d feature for robotic fruit picking", Computers and electronics in agriculture, Vol. 142, 2017, str. 388–396.

[47] Liu, J., Yuan, Y., Zhou, Y., Zhu, X., Syed, T. N., "Experiments and analysis of close-shot identification of on-branch citrus fruit with realsense", Sensors, Vol. 18, No. 5, 2018, str. 1510.

[48] Barth, R., Hemming, J., van Henten, E. J., "Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation", Biosystems Engineering, Vol. 146, 2016, str. 71–84.

[49] Lehnert, C., Sa, I., McCool, C., Upcroft, B., Perez, T., "Sweet pepper pose detection and grasping for automated crop harvesting", in 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016, str. 2428–2434.

[50] Kamilaris, A., Prenafeta-Boldú, F. X., "Deep learning in agriculture: A survey", Computers and electronics in agriculture, Vol. 147, 2018, str. 70–90.

[51] Ren, S., He, K., Girshick, R., Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems, Vol. 28, 2015, str. 91–99.

[52] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., "You only look once: Unified, real-time object detection", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 779–788.

[53] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., "Ssd: Single shot multibox detector", in European conference on computer vision. Springer, 2016, str. 21–37.

[54] Wan, S., Goudos, S., "Faster r-cnn for multi-class fruit detection using a robotic vision system", Computer Networks, Vol. 168, 2020, str. 107036.

[55] Gené-Mola, J., Vilaplana, V., Rosell-Polo, J. R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., "Multi-modal deep learning for fuji apple detection using rgb-d cameras and their radiometric capabilities", Computers and electronics in agriculture, Vol. 162, 2019, str. 689–698.

[56] Liu, Z., Wu, J., Fu, L., Majeed, Y., Feng, Y., Li, R., Cui, Y., "Improved kiwifruit detection using pre-trained vgg16 with rgb and nir information fusion", IEEE Access, Vol. 8, 2019, str. 2327–2336.

[57] Tu, S., Xue, Y., Zheng, C., Qi, Y., Wan, H., Mao, L., "Detection of passion fruits and maturity classification using red-green-blue depth images", Biosystems Engineering, Vol. 175, 2018, str. 156–167.

[58] Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., "Deepfruits: A fruit detection system using deep neural networks", sensors, Vol. 16, No. 8, 2016, str. 1222.

[59] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications", arXiv preprint arXiv:1704.04861, 2017.

[60] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., "Mobilenetv2: Inverted residuals and linear bottlenecks", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, str. 4510–4520.

[61] Long, J., Shelhamer, E., Darrell, T., "Fully convolutional networks for semantic segmentation", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[62] Ronneberger, O., Fischer, P., Brox, T., "U-net: Convolutional networks for biomedical image segmentation", in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, str. 234–241.

[63] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PP, 06 2016.

[64] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., "Encoder-decoder with atrous separable convolution for semantic image segmentation", in Proceedings of the European Conference on Computer Vision (ECCV), September 2018.

[65] Hinterstoisser, S., Pauly, O., Heibel, H., Martina, M., Bokeloh, M., "An annotation saved is an annotation earned: Using fully synthetic training for object detection", in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, str. 0–0.

[66] Khan, S., Phan, B., Salay, R., Czarnecki, K., "Procsy: Procedural synthetic dataset generation towards influence factor studies of semantic segmentation networks.", in CVPR Workshops, 2019, str. 88–96.

[67] Barth, R., IJsselmuiden, J., Hemming, J., Van Henten, E. J., "Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset", Computers and electronics in agriculture, Vol. 144, 2018, str. 284–296.

[68] Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., To, T., Cameracci, E., Boochoon, S., Birchfield, S., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, str. 969–977.

[69] Doersch, C., Zisserman, A., "Sim2real transfer learning for 3d human pose estimation: motion to the rescue", in Advances in Neural Information Processing Systems, Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., (ur.), Vol. 32. Curran Associates, Inc., 2019.

[70] Pashevich, A., Strudel, R., Kalevatykh, I., Laptev, I., Schmid, C., "Learning to augment synthetic images for sim2real policy transfer", in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019, str. 2651-2657.

[71] Baker, N., Lu, H., Erlikhman, G., Kellman, P. J., "Local features and global shape information in object classification by deep convolutional neural networks", Vision Research, Vol. 172, 2020, str. 46-61.

[72] Brendel, W., Bethge, M., "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet", International Conference on Learning Representations, 2019.

[73] Baker, N., Lu, H., Erlikhman, G., Kellman, P., "Deep convolutional networks do not classify based on global object shape", PLOS Computational Biology, Vol. 14, 12 2018, str. e1006613.

[74] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., Brendel, W., "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness", ArXiv, Vol. abs/1811.12231, 2019.

[75] Malhotra, G., Bowers, J., "The contrasting roles of shape in human vision and convolutional neural networks", in Proceedings of the 41st Annual Conference of the Cognitive Science Society, Goel, A., Seifert, C., Freksa, C., (ur.), Jul. 2019.

[76] Kim, M., Byun, H., "Learning texture invariant representation for domain adaptation of semantic segmentation", 06 2020, str. 12 972-12 981.

[77] Gao, L., Zhang, L., Zhang, Q., "Addressing domain gap via content invariant representation for semantic segmentation", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 9, May 2021, str. 7528-7536, dostupno na: https://ojs.aaai.org/index.php/AAAI/article/view/16922

[78] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., "Domain randomization for transferring deep neural networks from simulation to the real world", in 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017, str. 23–30.

[79] To, T., Tremblay, J., McKay, D., Yamaguchi, Y., Leung, K., Balanon, A., Cheng, J., Hodge, W., Birchfield, S., "NDDS: NVIDIA deep learning dataset synthesizer", https://github.com/NVIDIA/Dataset_Synthesizer. 2018.

[80] Tremblay, J., To, T., Birchfield, S., "Falling things: A synthetic dataset for 3d object detection and pose estimation", in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, str. 2119-21 193, dostupno na: https://doi.ieeecomputersociety.org/10.1109/CVPRW.2018.00275

[81] Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S., "Deep object pose estimation for semantic robotic grasping of household objects", in CoRL, 2018.

[82] Brochu, F., "Increasing shape bias in imagenet-trained networks using transfer learning and domain-adversarial methods", CoRR, Vol. abs/1907.12892, 2019, dostupno na: http://arxiv.org/abs/1907.12892

[83] Islam, M. A., Kowal, M., Esser, P., Jia, S., Ommer, B., Derpanis, K. G., Bruce, N. D. B., "Shape or texture: Understanding discriminative features in cnns", CoRR, Vol. abs/2101.11604, 2021, dostupno na: https://arxiv.org/abs/2101.11604

[84] Di Cicco, M., Potena, C., Grisetti, G., Pretto, A., "Automatic model based dataset generation for fast and accurate crop and weeds detection", in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, str. 5188–5195.

[85] Ward, D., Moghadam, P., Hudson, N., "Deep leaf segmentation using synthetic data", CoRR, Vol. abs/1807.10931, 2018, dostupno na: http://arxiv.org/abs/1807.10931

[86] Saleh, M., Elhajj, I. H., Asmar, D., Bashour, I., Kidess, S., "Experimental evaluation of low-cost resistive soil moisture sensors", in 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). IEEE, 2016, str. 179–184.

[87] Farhat, K., "Digging fork device for at least one digging fork", WO/2014/154983.

[88] Hogan, N., "Impedance control: An approach to manipulation", in 1984 American Control Conference, 1984, str. 304-313.

[89] Al-Shuka, H., Leonhardt, S., Zhu, W.-H., Song, R., Ding, C., Li, Y., "Active impedance control of bioinspired motion robotic manipulators: An overview", Applied Bionics and Biomechanics, 06 2018.

[90] Song, P., Yu, Y., Zhang, X., "A tutorial survey and comparison of impedance control on robotic manipulation", Robotica, Vol. 37, 01 2019, str. 1-36.

[91] Wang, X., Xiao, Y., Bi, S., Fan, X., Rao, H., "Design of test platform for robot flexible grasping and grasping force tracking impedance control", Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering, Vol. 31, 01 2015, str. 58-63.

[92] Azar, A. T., Ammar, H. H., de Brito Silva, G., Razali, M. S. A. B., "Optimal proportional integral derivative (pid) controller design for smart irrigation mobile robot with soil moisture sensor", in International Conference on Advanced Machine Learning Technologies and Applications. Springer, 2019, str. 349–359.

[93] Ünal, İ., Kabaş, Ö., Sözer, S., "Real-time electrical resistivity measurement and mapping platform of the soils with an autonomous robot for precision farming applications", Sensors, Vol. 20, No. 1, 2020, str. 251.

[94] Cao, P. M., Hall, E. L., Zhang, E., "Soil sampling sensor system on a mobile robot", in Intelligent Robots and Computer Vision XXI: Algorithms, Techniques, and Active Vision, Vol. 5267. International Society for Optics and Photonics, 2003, str. 304–310.

[95] Senthil, P., Akila, I., "Automated robotic moisture monitoring in agricultural fields", in 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE, 2018, str. 375–380.

[96] Komatsu, S., Kakiuchi, Y., Nozawa, S., Kojio, Y., Sugai, F., Okada, K., Inaba, M., "Tool force adaptation in soil-digging task for humanoid robot", in 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids). IEEE, 2017, str. 378–383.

[97] Botterill, T., Paulin, S., Green, R., Williams, S., Lin, J., Saxton, V., Mills, S., Chen, X., Corbett-Davies, S., "A robot system for pruning grape vines", Journal of Field Robotics, Vol. 34, No. 6, 2017, str. 1100–1122.

[98] He, L., Schupp, J., "Sensing and automation in pruning of apple trees: A review", Agronomy, Vol. 8, No. 10, 2018, str. 211.

[99] Majeed, Y., Karkee, M., Zhang, Q., "Estimating the trajectories of vine cordons in full foliage canopies for automated green shoot thinning in vineyards", Computers and Electronics in Agriculture, Vol. 176, 2020, str. 105671.

[100] Zahid, A., He, L., Zeng, L., Choi, D., Schupp, J., Heinemann, P., "Development of a robotic end-effector for apple tree pruning", Transactions of the ASABE, Vol. 63, No. 4, 2020, str. 847–856.

[101] You, A., Sukkar, F., Fitch, R., Karkee, M., Davidson, J. R., "An efficient planning and control framework for pruning fruit trees", in 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, str. 3930–3936.

[102] Kolmanič, S., Strnad, D., Kohek, Š., Benes, B., Hirst, P., Žalik, B., "An algorithm for automatic dormant tree pruning", Applied Soft Computing, Vol. 99, 2021, str. 106931.

[103] Kappassov, Z., Corrales, J.-A., Perdereau, V., "Tactile sensing in dexterous robot hands", Robotics and Autonomous Systems, Vol. 74, 2015, str. 195–220.

[104] Yuan, W., Dong, S., Adelson, E. H., "Gelsight: High-resolution robot tactile sensors for estimating geometry and force", Sensors, Vol. 17, No. 12, 2017, str. 2762.

[105] Winstone, B., Griffiths, G., Melhuish, C., Pipe, T., Rossiter, J., "Tactip—tactile fingertip device, challenges in reduction of size to ready for robot hand integration", in Robotics and Biomimetics (ROBIO), 2012 IEEE International Conference on. IEEE, 2012, str. 160–166.

[106] Hristu, D., Ferrier, N., Brockett, R. W., "The performance of a deformable-membrane tactile sensor: basic results on geometrically-defined tasks", in Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on, Vol. 1. IEEE, 2000, str. 508–513.

[107] Ward-Cherrier, B., Pestell, N., Cramphorn, L., Winstone, B., Giannaccini, M. E., Rossiter, J., Lepora, N. F., "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies", Soft robotics, 2018.

[108] Pestell, N., Lloyd, J., Rossiter, J., Lepora, N., "Dual-modal tactile perception and exploration", IEEE Robotics and Automation Letters, 2018.

[109] Lepora, N. F., "Biomimetic active touch with fingertips and whiskers", IEEE transactions on haptics, Vol. 9, No. 2, 2016, str. 170–183.

[110] Ward-Cherrier, B., Rojas, N., Lepora, N. F., "Model-free precise in-hand manipulation with a 3d-printed tactile gripper", IEEE Robotics and Automation Letters, Vol. 2, No. 4, 2017, str. 2056–2063.

[111] Ward-Cherrier, B., Cramphorn, L., Lepora, N. F., "Tactile manipulation with a tacthumb integrated on the open-hand m2 gripper", IEEE Robotics and Automation Letters, Vol. 1, No. 1, 2016, str. 169–175.

[112] Lepora, N. F., Aquilina, K., Cramphorn, L., "Exploratory tactile servoing with active touch", IEEE Robotics and Automation Letters, Vol. 2, No. 2, 2017, str. 1156–1163.

[113] de Boer, G., Wang, H., Ghajari, M., Alazmani, A., Hewson, R., Culmer, P., "Force and topography reconstruction using gp and mor for the tactip soft sensor system", in Conference Towards Autonomous Robotic Systems. Springer, 2016, str. 65–74.

[114] Ito, Y., Kim, Y., Obinata, G., "Contact region estimation based on a vision-based tactile sensor using a deformable touchpad", Sensors, Vol. 14, No. 4, 2014, str. 5805–5822.

[115] Ito, Y., Kim, Y., Obinata, G., "Multi-axis force measurement based on vision-based fluid-type hemispherical tactile sensor", in Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. IEEE, 2013, str. 4729–4734.

[116] Lepora, N. F., Church, A., De Kerckhove, C., Hadsell, R., Lloyd, J., "From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor", IEEE Robotics and Automation Letters, Vol. 4, No. 2, 2019, str. 2101–2107.

[117] Takahashi, K., Tan, J., "Deep visuo-tactile learning: Estimation of tactile properties from images", in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, str. 8951–8957.

[118] Yuan, W., Zhu, C., Owens, A., Srinivasan, M. A., Adelson, E. H., "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor", in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, str. 951–958.

[119] Maric, B., Polic, M., Tabak, T., Orsag, M., "Unsupervised optimization approach to in situ calibration of collaborative human-robot interaction tools", in 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). IEEE, 2020, str. 255–262.

[120] Kocer, B. B., Ho, B., Zhu, X., Zheng, P., Farinha, A., Xiao, F., Stephens, B., Wiesemüller, F., Orr, L., Kovac, M., "Forest drones for environmental sensing and nature conservation", in 2021 Aerial Robotic Systems Physically Interacting with the Environment (AIRPHARO). IEEE, 2021, str. 1–8.

[121] Hess, R., Blender Foundations: The Essential Guide to Learning Blender 2.6. Focal Press, 2010.

[122] AliceVision, "Meshroom: A 3D reconstruction software.", dostupno na: https://github.com/alicevision/meshroom 2018.

[123] Seraji, H., Colbaugh, R., "Force tracking in impedance control", in [1993] Proceedings IEEE International Conference on Robotics and Automation. IEEE Comput. Soc. Press, str. 499–506, dostupno na: http://ieeexplore.ieee.org/document/291908/

[124] Marković, L., Car, M., Orsag, M., Bogdan, S., "Adaptive stiffness estimation impedance control for achieving sustained contact in aerial manipulation", in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, str. 117–123.

[125] Lepora, N. F., "Soft biomimetic optical tactile sensing with the tactip: A review", arXiv preprint arXiv:2105.14455, 2021.

[126] Lepora, N. F., Stinchcombe, A., Ford, C., Brown, A., Lloyd, J., Catalano, M. G., Bianchi, M., Ward-Cherrier, B., "Towards integrated tactile sensorimotor control in anthropomorphic soft robotic hands", arXiv preprint arXiv:2102.03318, 2021.

PUBLICATIONS

.1 PUBLICATION 1 - PEPPER TO FALL: A PERCEPTION METHOD FOR SWEET PEPPER ROBOTIC HARVESTING

M. Polić, J. Tabak and M. Orsag. Pepper To Fall: A Perception Method For Sweet Pepper Robotic Harvesting. Intelligent Service Robotics, online, 2021, IF: 2.246 (Q3).

Intelligent Service Robotics
https://doi.org/10.1007/s11370-021-00401-7

ORIGINAL RESEARCH PAPER

Pepper to fall: a perception method for sweet pepper robotic harvesting

Marsela Polic[1] · **Jelena Tabak**[1] · **Matko Orsag**[1]

Received: 15 July 2021 / Accepted: 6 November 2021

Abstract
In this paper we propose a robotic system for picking peppers in a structured robotic greenhouse environment. A commercially available robotic manipulator is equipped with an RGB-D camera used to detect a correct pose to grasp peppers. The detection algorithm uses the state-of-the-art pretrained CNN architecture. The system was trained using transfer learning on a synthetic dataset made with a 3D modeling software, Blender. Point cloud data are used to detect the pepper's 6DOF pose through geometric model fitting, which is used to plan the manipulator motion. On top of that, a state machine is derived to control the system workflow. We report the results of a series of experiments conducted to test the precision and the robustness of detection, as well as the success rate of the harvesting procedure.

Keywords Robotic harvesting · Transfer learning · RGB-D · Convolutional neural networks · Depth camera · Sim2real

1 Introduction

One of the most prominent contributors to the cost of food is the input manual labor, especially in organic agriculture with a decreased use of pesticides and GMO crops. One of the goals within the SpECULARIA project [1] is to reduce human labor input in small indoor farms by replacing it with a heterogeneous team of robots. In the presented work, this team of robots is used in structured greenhouse cultivation of sweet peppers. An aerial vehicle plays a surveillance role, detecting and deciding which plants need to be treated. The plants are grown in container units so that they can be transported around the greenhouse by an unmanned ground vehicle (UGV). The UGV transports the plants to the workstation, where a robotic manipulator treats the plants under controlled conditions. The collaborative manipulator is in the focus of the presented work, in a pepper harvesting procedure example.

Marsela Polic
marsela.polic@fer.hr

Jelena Tabak
jelena.tabak@fer.hr

Matko Orsag
matko.orsag@fer.hr

[1] Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia

As a labor-intensive activity performed several times a season, harvesting is a perfect candidate for automation that has been in the focus of scientific research for decades [2]. Although crop-specific solutions exist, harvesting robots are not yet a standard due to the limited capabilities of existing systems. The main challenges of harvesting with robots fall into several categories. First, a *perception* method is needed for plant detection, localization and grasping. Furthermore, within the *motion planning* problem, a robot control strategy is needed for precise positioning of the end effector during target approach [3]. Finally, *dexterous manipulation* mostly concerns grasping the fruit without harming it. Depending on the type of fruit and the environment, different problems occur in each of the domains.

In this paper, we propose a commercial manipulator-based pepper harvesting system. We claim the following contributions: (i) RGB-D and deep learning-based estimation method of the sweet pepper 6DOF pose for grasping, (ii) control framework for a structured greenhouse harvesting in a form of a state machine, and (iii) experimental validation of a sim2real and transfer learning-based perception method by closed-loop motion control and grasping using a Franka Emika collaborative robotic arm with Intel RealSense D435 camera. The remainder of this paper discusses the claimed contributions in detail. First, we review related work that addresses the three main challenges in robotic harvesting. The following section describes the system design, including

Published online: 01 December 2021

Fig. 1 Collaborative robot Franka Panda performing harvesting procedure on a sweet pepper (Capsicum annuum) plant in a controlled indoor farming setup

hardware and software solutions, as well as the underlying perception, control, and planning methods. Finally, experimental results are presented and discussed.

2 Related work

Most recent work in the field of perception relies on consumer RGB-D cameras, thanks to rapid development and price reduction. A comprehensive overview of existing technologies and perception methods is given in [4]. According to this overview, most of the existing work is based on the Kinect V2 [5–8], a consumer Time of Flight (ToF) camera, while previous research was mostly based on its predecessor Kinect v1 [9–11]. On the other hand, Intel RealSense has only recently come into research focus, with previous work limited to some applications using the R200 [12], F200 [13], and D435 [14] cameras in grapevine, pepper, and apple orchards, respectively. In this work, we test the capabilities of the Intel RealSense D435 in a robotic application for harvesting peppers.

Unsolved perception challenges include uncontrolled and changing light conditions, which are usually addressed with case-specific solutions. A novel and promising approach is proposed by authors in [15] with the acquisition method involving controlled illumination acquisition protocol called Flash-No-Flash (FNF). In this method, two images are acquired consequentially for the same setup, with and without strong artificial illumination, and the detection is based on the difference between these images, effectively achieving control over the ambient light. However, in the presented work, as envisioned by the SpECULARIA project, the robotic manipulator is static at its designated workstation in the greenhouse (Fig. 1). As mentioned above, the plants are transported to the manipulator in their containers by a UGV. The advantages of such a setup are that the lighting conditions at the manipulation site can be predefined and controlled, avoiding environmental variations.

Another problem in perception, apart from lighting conditions, is generalization over variation in plant physiognomy. Most early detection systems relied on color filtering for ripe fruit detection and were very poor at generalization due to sensitivity to fine-tuned image processing methods, as shown in [3,9,16]. With the advent of CNN-based models, the detection generalization performance has been immensely improved [4]. Recent work combines 2D image detection with depth information to determine the exact grasp position. Examples include surface normal estimation to determine optimal grasp positions [13] and peduncle model fitting [17]. In [16], a geometric model fitting approach is used where a superellipsoid is fitted to the pepper model. For simplicity and flexibility, we propose here instead a 6DOF pose estimation for peppers based on a cylinder fitting method.

The use of deep learning perception methods is associated with the cost of large training dataset generation [18,19]. In an attempt to reduce the cost of the tedious labeling procedure, researchers have recently turned to the generation of synthetic datasets, with applications in agriculture for various crops and cultures [20–22], including a synthetic dataset for the semantic segmentation tasks of *C. annuum* [23]. However, this dataset is only labeled for semantic segmentation, an important and relevant task, but insufficient for the harvesting application where object detection with localization is more important than pixel-wise semantic segmentation. For this purpose, a user-defined synthetic dataset is generated for both object detection and semantic segmentation tasks. Using transfer learning, an SSD network [24] with MobileNet V2 backbone [25] is trained for sweet pepper detection. This sim2real model is validated in the laboratory condition in harvesting experiments.

3 Method

3.1 System setup

The hardware setup consists of a Franka Panda collaborative robot, with an Intel RealSense D345 RGB-D camera mounted on Panda hand. The robot-camera system is calibrated in an autonomous optimization procedure as described in [26], with a target fixed in the global coordinate frame,

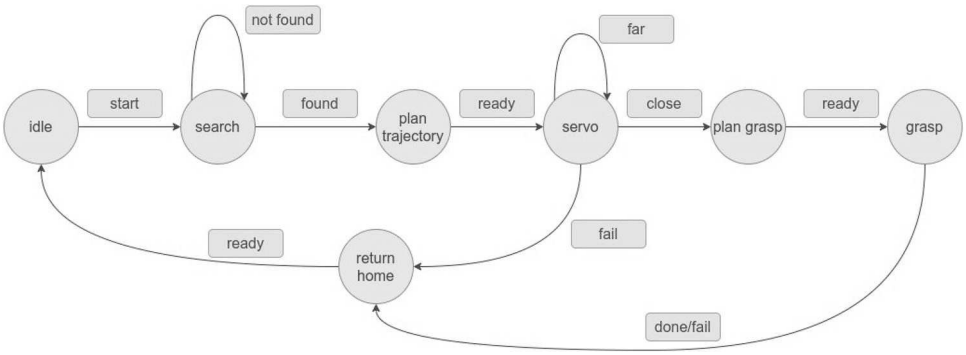


Fig. 2 Pepper harvesting state machine for Franka Emika collaborative robot

and camera-end effector transformation as the variable in the minimization problem for target position dissipation. The fixed target is an AprilTag, whose central point is fixed (and unknown) in the global coordinate frame. By recording the position of this point in the local RGB-D camera frame, and the corresponding end effector global poses, a minimization problem can produce a transform from the end-effector frame to the camera frame, such that all the recorded local April tag positions are transformed into the same point in the global coordinate frame. For a dataset of April tag positions $\{p_i | i \in 1, ..., N_s\}$ in the local camera frame, recorded with the end effector at global positions $\{T^{ee}_{O,i} | i \in \{1, ..., N_s\}\}$, this is achieved by solving the minimization problem 1 using the Nelder–Mead simplex. The resulting position dissipation of the point cloud data is at approximately 5 mm.

$$\mathbf{T}^* = \arg\min_{T} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \left\| T^{ee}_{0,i} \cdot T \cdot p_i - T^{ee}_{0,j} \cdot T \cdot p_j \right\| \quad (1)$$

It is worth noting that in actual implementation, the search space is not defined in the 3D transformation matrix $T \in \mathbb{R}^{4\times4}$ space, but in position/quaternion $t \in \mathbb{R}^{7\times1}$ vector space instead, where the first three variables correspond to translation component of the transformation matrix T, and the latter four correspond to the four components of rotation quaternion. The simplex method does not take care of quaternion normalization, and this step should be taken care of before transforming the quaternion into the rotation matrix. This can additionally lead to getting stuck in local optima. In practice, restarting the optimization at a new random state can easily resolve this problem.

Since Franka cobot is controlled in real time, a real-time patched Linux kernel is used on the control computer. On the other hand, the MobileNet SSD network is implemented in TensorFlow, which enables fast inference on a GPU using NVidia's CUDA Toolkit. However, NVidia drivers do not officially support real-time kernels. Therefore, a remote master ROS setup was used, with one computer controlling the robot motion in real time, and the other running the detection network on the NVidia GPU.

The robot is controlled through ROS, using MoveIt package and the underlying solvers for inverse kinematics and motion planning. The detection network is run on a NVidia GeForce MX330 GPU with TensorFlow 2 with CUDA. Point cloud manipulations including filtering and geometric model fitting are implemented using the open sourced Point Cloud Library.

3.2 State machine

The autonomous pepper harvesting mission is implemented in the form of a state machine, as shown in Fig. 2. The robot is controlled in a position trajectory mode, with the position reference provided by the object detection module.

In the **Search** state, the robot records the plant from several predefined Cartesian poses. In each pose, RGB-D camera image and point cloud are saved for detecting and choosing a target pepper, as described in *6DOF pose detection* Sect. 3.4. A pepper should be detected from at least two frames to be chosen as harvesting target. A time-out signal is defined for the **Search** state, with a number of sequentially visited search poses without a recognized target.

In order to reduce the probability of collision while approaching the pepper, the motion is planned through two waypoints. First, the gripper is referenced to a Cartesian pose at the plant's bounding cylinder in **Plan to cylinder** state. This position is chosen to be the closest to the target pepper position, and the orientation is chosen as the final grasp orientation. This way, the probability of collision with other plant parts is reduced, since the robot approaches the pepper from outside the plants bounding volume. It should be noted that the entire plant is not within the robot workspace, and the bounding cylinder positions are only searched within a 120 degree range, as shown in Fig. 3. For peppers detected in other parts of the plant, the heterogeneous robotic team collaboration has to be deployed to adjust the plant orientation relative to the robot manipulator. The details of this implementation go beyond the scope of this work.

With the approach pose chosen in the described way, the final waypoint for grasping is realized by moving along the

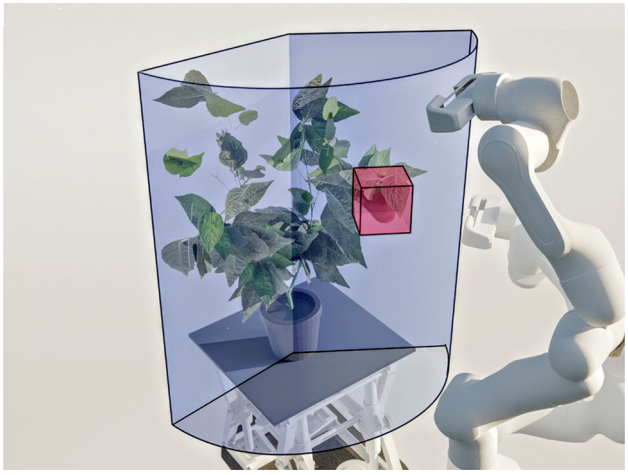


Fig. 3 Target pepper for harvesting is detected in 2D image, and marked with a bounding box (here shown as a red 3D box). A bounding cylinder is defined for the plant, on which an approach position is planned as a point closest to the detected pepper, to reduce collision probability. The approach position, shown in a semi-transparent robot pose, is only searched within the central 120 degrees of the bounding cylinder, as a simplified form of the robot workspace. The orientation is determined by the pepper orientation

grippers local approach axis only towards the **Plan to pepper** state. Finally, when the gripper reaches the pepper, it is able to grasps it. A standard gripper can be adapted with an additional peduncle cutting tool to physically remove the fruit from the plant.

It should be noted that a common approach in the state-of-the-art robot harvesting solutions involves a servoing state. We have tested the system performance with this additional servoing state, and obtained comparable results in terms of precision, with no information gain, and an increased task execution time. In fact, due to the sensing technology limitations, the pepper position estimation was deteriorated in servo control. Namely, the minimum depth sensing range of the depth camera is $20cm$. In practice, the depth camera estimated poorly at close distance up to twice the nominal minimum distance. This of course causes poor pepper position estimation, and leads to failed grasp planning. Detailed analysis and discussion is given in the results section.

3.3 2D detection

A common approach in fruit detection is Single Shot Detector (SSD) deep learning model, based on a convolutional neural network (CNN) structure. The SSD networks are capable of fast simultaneous detection of multiple objects in a single shot. In this work, MobileNet architecture was used as a base of SSD, with additional convolutional layers added to the end of the base network. As is becoming standard in deep learning, transfer learning technique was deployed on a TensorFlow implementation of MobileNet SSD network pretrained on the COCO dataset.

The training dataset for pepper detection was generated synthetically using Blender [27]. A realistic pepper 3D model was generated using open source photogrammetric tool Meshroom [28] on a set of 200 images obtained by a smart phone camera. This model was augmented into various peppers using morphological and color transformations. Variations in morphology and texture of the original bell pepper were controlled such that the pepper shape still resembles the bell cultivar. The plant stem and pots were modeled from scratch. Leaves and flowers were photographed and imported as plane models. The planes were slightly modified for a more realistic shape. They were also additionally morphed and twisted for augmentation. A realistic shape of the plant was obtained using a standard curve generator plugin, the sapling tree generator, available in Blender [27]. With additional variation in the scene background and lighting conditions, a procedural generation of automatically labeled object detection and semantic segmentation dataset yielded 540 images. After training the network on the synthetic dataset, additional fine tuning was conducted on a small dataset of real, manually labeled pepper images. This dataset of around 40 images was collected using real pepper plants mostly in controlled laboratory environment, and a small number of samples in the cluttered greenhouse scene.

MobileNet SSD detects objects in 2D RGB images, producing 2D bounding boxes denoting position of detected peppers in the image. An example of a successful multiple object detection is shown in Fig. 4. The images are obtained with the Intel RealSense D435 RGB-D camera, at a rate of approximately 15Hz. The depth channel of the RGB-D camera allows for a 3D object detection and localization. The bounding boxes of detected objects are applied to the organized point cloud generated by the D345. The centroid of the resulting filtered point cloud is considered to be the pepper position in the camera coordinate frame. The 3D detection pipeline finally produces the global position of the pepper through transformation of pepper position into the robot base reference frame. This information is used while approaching the pepper.

3.4 6DOF pose detection

Since the pepper centroid position is not sufficient for a successful grasp, and hence harvesting, a more precise 6DOF pose of the pepper is determined by geometric model fitting. Object detection is conducted in the **Search** state. In each search pose, the detected 3D objects are stored, and afterward matched based on their 3D global positions using known transformation of the camera frame to the global robot reference frame. In the matching procedure, a position threshold is allowed due to the described position ambiguity. The

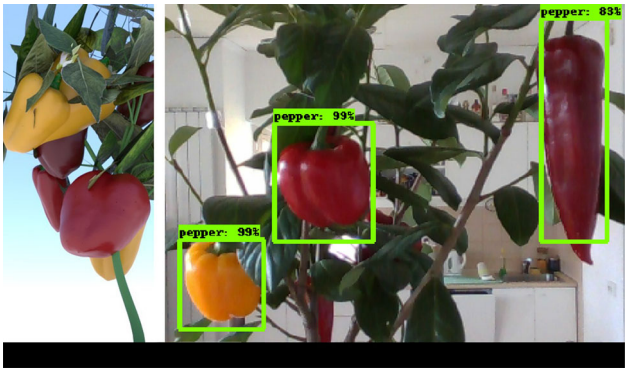


Fig. 4 **a**) Example of a synthetic pepper image generated in Blender software. **b** Example of detected sweet peppers in an Intel RealSense 2D image, with the bounding boxes used in point cloud filtering. The image shows one real example for all of the pepper varieties used in this work: a yellow bell pepper on the left, a red bell pepper in the middle, and a red horn variety pepper on the right

closest object from among those detected in at least two different frames is chosen as target. The cylinder model is fitted on the filtered point cloud, and grasping pose is chosen at the top of the detected pepper (with an offset corresponding to the width of the gripper plates), perpendicular to the pepper orientation.

For each detection of the same pepper, subcloud containing the pepper is extracted from the corresponding point cloud using the bounding box from image detection. The remaining points are filtered based on their relative distance to the camera, i.e., only those within 5cm of the pepper centroid, along the local camera z axis, are kept. The resulting subclouds containing the same pepper from several perspectives are transformed into the global reference frame, and iterative closest point algorithm is used to incrementally register a series of point clouds two by two. The result is shown in Fig. 5 a). Along with the pepper, some leaves voxels are visible on the right. Then, an infinite length cylinder is fitted on the 3D point cloud, in an optimization procedure minimizing the distance of the model points to the cylinder surface. The peppers are expected to grow vertically. Conducting experiments with a variety of pepper sorts, we have observed that certain varieties (horn-shaped ones) tend to increase the fitted cylinder radius due to their curved tip. For this reason, the cylinder was fitted on the upper 3/4 of the filtered point cloud (z axis in the global reference frame). This step is shown in subfigure b) of Fig. 5. The next step is filtering the pepper cloud in order to remove the points outside the fitted cylinder, usually corresponding to objects such as leaves or stems. The pepper length, i.e., the cylinder height, is determined from the projection of the remaining pepper points on the central cylinder axis, as shown in subplot c). Finally, the pepper's 6DOF pose can be estimated, with position defined with the cylinder center, and orientation by the cylinder axis orientation. Additionally, the cylinder radius and height can be used to estimate the pepper size.

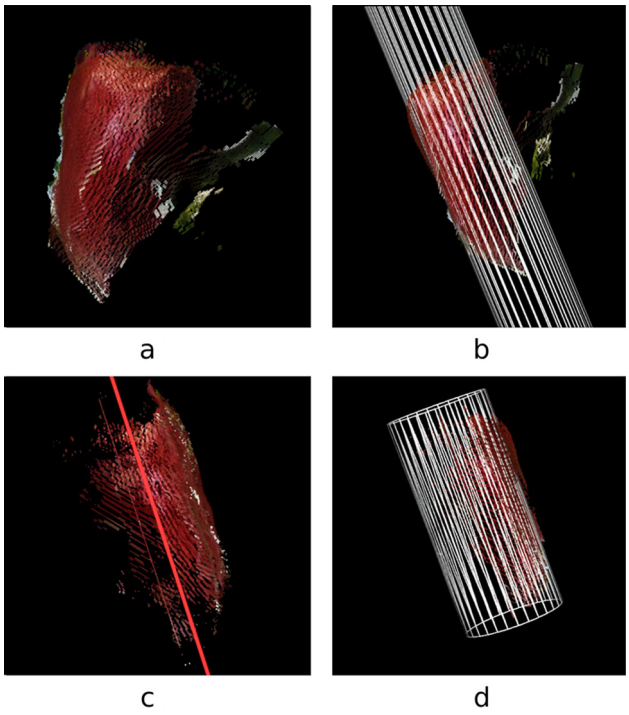


Fig. 5 6DOF pose estimation using cylinder fitting. Image **a** shows the subcloud obtained by bounding box and relative depth filtering. Image **b** shows the fitted cylinder model. The points outside of the cylinder model are filtered out, e.g., points corresponding to leaves in this example. Image **c** shows the cylinder axis used for projection of pepper points in the pepper length, i.e., cylinder height estimation. Image **d** shows the resulting 6DOF pose and shape estimation of the pepper with a cylinder geometric model

3.5 Grasping

The pepper detection method is evaluated through the precision in grasping position, avoiding removing the fruit for these experiments, a default Franka Emika hand was used. A minor modification was introduced to the fingers that enabled a more gentle contact. 3D printed plastic plates were attached thereby adding some length to the fingers, as well as a larger contact surface. The compliance of the gripper is only available through the black box grasping interface of the gripper controller. For better performance, a custom flexible gripper with additional peduncle cutting adapter will be designed and deployed.

4 Experimental validation

The described method is evaluated in several stages. First, the precision of the object detection method is statistically

evaluated. Then, the robustness of detection is evaluated with respect to perturbation in search poses. Finally, a statistical evaluation of the grasping precision is conducted. All of the evaluation experiments are conducted on three types of sweet peppers: red bell peppers, yellow bell peppers, and red horn peppers. It is worth repeating that the original pepper upon which the synthetic dataset was generated was a red bell pepper cultivar. The yellow bell peppers occurred in the dataset through augmentation, by applying color transformations in plant generation. The horn cultivar was not seen in the training dataset, but was, however, successfully detected by the network, most probably thanks to the smooth texture and red color. The examples of the three pepper types are shown in Fig. 4.

4.1 Detection precision

The detection system consists of a collaborative robot equipped with a consumer RGB-D camera. The detection is based on fusion of the object detection CNN in 2D images, and 3D data from the RGB-D point cloud. The cumulative precision and repeatability of the detection system is statistically modeled in a series of detection experiments for various types of peppers. This precision measure encompasses contributions of the following inaccuracies: RGB-D camera precision, camera-robot transform calibration, 2D object detection, 3D geometric modeling, and including obstacles such as leaves occluding pepper parts. For each of the pepper varieties, detection was conducted in 10 experiment repetitions. Each pepper was detected from the same three search positions of the robot. The expected output is to obtain 10 identical pose estimations. In reality, the cumulative effect of component-wise imprecision is detected in the deviation of the pepper pose estimation. The results presented in Table 1 show the mean (where applicable) and standard deviation of the estimated variables, for three pepper varieties: red bell pepper (P1), yellow bell pepper (P2), and red horn pepper (P3). Without the ground truth, the dissipation of estimates is taken as the precision measure. Since the ground truth is not available, we use the mean of the position estimates over experiment repetitions as the correct pepper position, and provide the dissipation of estimates with respect to these values.

The deviation of the pepper position in three axis is sufficiently small when the relation of sizes of the used gripper and the peppers is taken into account. Regarding the geometric model, the dissipation of fitted cylinder model parameters is again at a similar level for the radius estimation, and somewhat more significant for the pepper height estimation. In control, the height of the pepper is used when determining the grasping position with respect to the pepper center, so that the grasp is realized below the pepper top. The size of the gripper plates, at around 3 cm in diameter, is able to compensate for this estimation error, except in rare extreme cases when the grasping fails, as shown later in the grasp pose precision analysis.

Table 1 Pepper detection precision results

	P1		P2		P3	
	Mean	SD	Mean	SD	Mean	SD
x (mm)		8.3		3.8		3.7
y (mm)		4.6		1.5		1.6
z (mm)		5.4		2.3		3.7
r (mm)	28.0	5.7	30.7	1.9	18.0	2.5
h (mm)	74.7	10.7	67.5	5.4	123.8	8.3

4.2 Detection robustness

The second set of experiments analyzes the robustness of the detection with respect to perturbation in search poses. Instead of using the same search poses, in each of the ten repetitions, three random robot poses were used for camera image and point cloud acquisition. Similar to the previous analysis, mean values and standard deviations are calculated for variables of interest, namely pepper position and size, and given in Table 2, for examples from three pepper varieties: red bell pepper (P1), yellow bell pepper (P2), and red horn pepper (P3).

It is interesting to observe the reduction in dissipation for the red bell pepper when compared to the precision results in Table 1, as well as to the other pepper varieties. This can be explained by the fact that this pepper variety was the most represented pepper in the object detection datasets (both synthetic and real), and its detection is therefore the most precise. This is in contrast with the results from Table 1, but can be explained with the relatively poor choice of the fixed search poses for this experiment. In this case, the randomness of the search state managed to improve overall detection by statistically filtering out the local minima. For the other two pepper varieties, the dissipation results are worsened, implying that the detection pipeline is not trained for these varieties as well as for the original one. However, the position error is still low enough to compensate with the gripper size. The estimates of the pepper sizes are comparable to the previously discussed results, except for the horn pepper variety. This can again be explained by the fact that this pepper variety is under-represented in detection dataset. In fact, peppers of such shape (width-height ratio) are not seen by the network during training, which influences the detection network performance. Additionally, the elongated pepper shape tends to get visually obstructed by other plant parts, especially leaves. This severely impairs the correct pepper geometric modeling from certain viewpoints. However, the occlusion of lower

Table 2 Pepper detection robustness analysis

	P1		P2		P3	
	Mean	SD	Mean	SD	Mean	SD
x (mm)		2.9		6.4		3.7
y (mm)		1.9		2.6		4.2
z (mm)		2.1		4.9		12.7
r (mm)	29.4	2.8	34.0	5.1	18.0	3.9
h (mm)	78.8	15.5	66.5	6.7	89.2	23.2

parts of the pepper leading to smaller estimation of pepper size at the same time influences the correlated position of pepper centroid. The correlation causes these errors to cancel out, and reduces the overall harvesting error probability, as seen in the results in the following subsection.

4.3 Grasp pose precision

Precision of robot gripper positioning during grasping is analyzed as a validation of the harvesting procedure. Different peppers from the three pepper varieties are detected, and grasps were attempted in a series of experiments. The success was determined by manual inspection of the grasping results. As shown in Table 3, the grasping procedure overall was successful in 75% (31/41) of the attempts. Out of the (10) failed attempts, 20% (2) correspond to the situation where the pepper is either out of the robot reach, or a feasible grasp position is not found by the IK. The remaining 80% (8) failed attempts are due to errors in detection, either in position, orientation, size, or a combination of those.

When considering the success across the pepper varieties, the grasping success rate for the main, red bell pepper variety, is at the average of 75% (16/21). The yellow bell pepper had a grasping success of 67% (10/15). For the horn variety, only 5 attempts were made, and all were successful, showing the relatively large dissipation in pepper height and position estimation to be easily compensated by the gripper physical properties (plates size). The photographic results of the experiments are documented at [1]. The videos of the pepper grasping and harvesting results can also be found there, along with the simulation scenario grasping experiments with 90% of the experiments (18/20) successful, and the remaining failed in the perception stage due to leaves and stem occluding the target pepper.

4.4 Servo information gain

To compare the proposed approach to existing state-of-the-art solutions, as well as to test whether servoing improves the performance, we replaced the **Plan to cylinder** state from the described state machine with the **Servo** state. The servo

Table 3 Pepper grasping results. **S**uccess (S), **F**ailed **D**etection (FD), **F**ailed **P**lanning (FP)

	P1		P2		P3		Overall	
	Abs	%	Abs	%	Abs	%	Abs	%
S	16	76.2	10	66.7	5	100	31	75.6
FD	3	14.3	5	33.3	0	0	8	19.5
FP	2	9.5	0	0	0	0	2	4.9

approach is often deployed in applications where variations in the working environment are expected during operation. Since our scenario relies on a strictly controlled indoor operation where external interference is not expected, the servoing approach is not used to account for variation, but only as means of improving the detection during the approach. During servoing, we approach the pepper in increments of 1cm based on the latest valid detection (within a predefined distance to the current estimate). **Servo** state is terminated once the distance between the gripper and the target object is below 20cm. Lower threshold could not improve the performance, as the minimum distance from the depth camera to the scene, for which Intel RealSense D435 provides depth data at the 640x480 resolution, is 17.5cm.

Global object positions extracted from valid detections in **Servo** state are shown in Fig. 6. The results show there is no significant improvement in object position estimation, with maximum difference between two detections at around 5mm, which is an error easily compensated with the gripper size. Moreover, multiple outliers can be observed towards the end of servoing, unsurprising as the depth readings deteriorate with the distance from the camera to the object approaching the 17.5cm threshold. In other words, not only is the information gain in servoing state insignificant, but as a result of poor detection at low distances, noise is introduced into the estimated object position. Finally, without an improvement in precision, servoing state even increased task execution time significantly. The proposed state machine is hence designed without the **Servo** state.

5 Discussion

In this work, the functionality of a sweet pepper harvesting system based on a commercial manipulator is presented and verified under experimental conditions. A sweet pepper detection method is developed and evaluated on several sweet pepper varieties. The precision and robustness results imply that additional improvements can be made in the detection pipeline. The network was trained on a dataset that significantly favors one pepper variety, namely the red bell peppers. Another variety, yellow bell peppers, was generated in the synthetic dataset through augmentation, but at a much

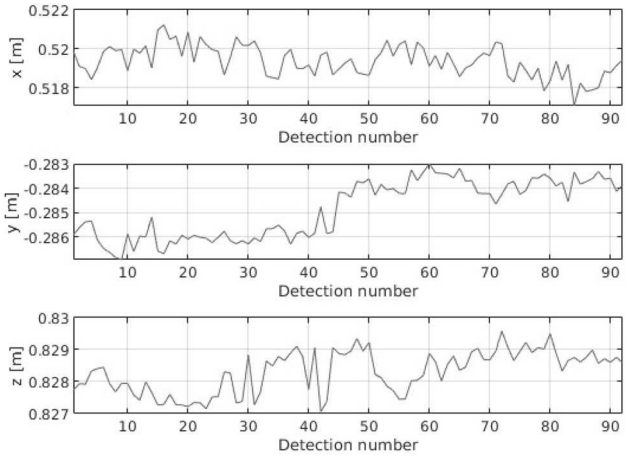


Fig. 6 3D global position of the target pepper during **Servo** state. The variation or improvement of the position estimation at a millimeter level can be neglected, and the figure thereby justifies replacing the servo motion with direct pepper approach, since the position estimation is not significantly improved over time

smaller scale, and the deterioration in detection performance can be observed. The third variety, the horn peppers, was not included in the training dataset at all, and the detection performance on this variety decreases even more. With additional work on the synthetic dataset, namely fairer distribution of generated samples over expected targets, improvement of detection pipeline can be expected.

The statistics of the grasping experiments reflect the robustness of the developed system as a whole, and prove the approach is promising. A most prominent issue was detected in the pepper 6DOF estimation procedure, where obstacles and occlusions deteriorated the geometrical modeling performance. Here, semantic segmentation, deployed in combination with the presented object detection, is expected to greatly improve the overall system performance. Namely, in case the geometric model was only fitted to the points corresponding to the chosen target pepper, disregarding the occluding leaves' or stems' voxels, a more precise and robust estimation would be expected. A simpler alternative to this approach would be additional color filtering of the pixels within the detected bounding box. A more complex solution could rely on instance segmentation. As part of the future work, we plan to implement and compare the mentioned approaches.

Regarding the motion planning in the agricultural procedure, contrary to a lot of existing work, we conclude that the servoing approach does not benefit our method. However, a compliant motion algorithm considering potential collisions in runtime could benefit from using the servoing approach, e.g., thorough introduction of dynamic potential fields. Despite being quite simplified, the collision avoidance technique proposed in the planning stage of the state machine was shown highly effective, without additionally restricting the manipulator motion. Again, improvements could be introduced using semantic or instance information during the perception phase. Semantic segmentation could also improve the geometric primitive fitting by only taking into consideration the points corresponding to target object. In the case of peppers, this could improve modeling in case leaves and other plant parts occlude a target harvested pepper. Finally, generalization ability of the method is a valuable characteristic, considered as a part of the future work through extension to other crops and cultures.

Author contributions All authors contributed to the study conception and design. Sensing method conception was performed by Marsela Polic and implementation by Jelena Tabak and Marsela Polic. Motion control was implemented and experiments conducted by Marsela Polic and Jelena Tabak. Analysis was performed by all authors. Most of the first draft of the manuscript was written by Marsela Polic. All authors commented and contributed on this manuscript. All authors read and approved the final manuscript.

Funding This work has been supported by Croatian Science Foundation under the project Specularia UIP-2017-05-4042 [1].

Availability of data and material The data presented in this work are publically available on the project web site.

Declarations

Conflicts of interests The authors have no relevant financial or non-financial interests to disclose.

Code availability The code developed and used in this work is publically available within the Laboratory GitHub repository.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Orsag M et al (2021) Specularia. https://sites.google.com/view/specularia-pepper-picking, Accessed: 2021-03-31
2. Bac CW, van Henten EJ, Hemming J, Edan Y (2014) Harvesting robots for high-value crops: state-of-the-art review and challenges ahead. J Field Robot 31(6):888–911
3. Barth R, Hemming J, van Henten EJ (2016) Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation. Biosys Eng 146:71–84
4. Fu L, Gao F, Wu J, Li R, Karkee M, Zhang Q (2020) Application of consumer rgb-d cameras for fruit detection and localization in field: a critical review. Comput Electron Agric 177:105687
5. Lin G, Tang Y, Zou X, Xiong J, Fang Y (2020) Color-, depth-, and shape-based 3d fruit detection. Precision Agric 21(1):1–17
6. Tu S, Pang J, Liu H, Zhuang N, Chen Y, Zheng C, Wan H, Xue Y (2020) Passion fruit detection and counting based on multiple scale faster r-cnn using rgb-d images. Precision Agric 21(5):1072–1091
7. Wang Z, Walsh KB, Verma B (2017) On-tree mango fruit size estimation using rgb-d images. Sensors 17(12):2738

8. Zhang J, Karkee M, Zhang Q, Zhang X, Yaqoob M, Fu L, Wang S (2020) Multi-class object detection using faster r-cnn and estimation of shaking locations for automated shake-and-catch apple harvesting. Comput Electron Agric 173:105384
9. Nguyen TT, Vandevoorde K, Wouters N, Kayacan E, De Baerdemaeker JG, Saeys W (2016) Detection of red and bicoloured apples on tree with an rgb-d camera. Biosys Eng 146:33–44
10. Perez RM, Cheein FA, Rosell-Polo JR (2017) Flexible system of multiple rgb-d sensors for measuring and classifying fruits in agri-food industry. Comput Electron Agric 139:231–242
11. Andujar D, Ribeiro A, Fernández-Quintanilla C, Dorado J (2016) Using depth cameras to extract structural parameters to assess the growth state and yield of cauliflower crops. Comput Electron Agric 122:67–73
12. Milella A, Marani R, Petitti A, Reina G (2019) In-field high throughput grapevine phenotyping with a consumer-grade depth camera. Comput Electron Agric 156:293–306
13. Lehnert C, English A, McCool C, Tow AW, Perez T (2017) Autonomous sweet pepper harvesting for protected cropping systems. IEEE Robot Autom Lett 2(2):872–879
14. Kang H, Chen C (2020) Fruit detection, segmentation and 3d visualisation of environments in apple orchards. Comput Electron Agric 171:105302
15. Arad B, Kurtser P, Barnea E, Harel B, Edan Y, Ben-Shahar O (2019) Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. the case study of sweet pepper robotic harvesting. Sensors,19(6): 1390
16. Lehnert C, Sa I, McCool C, Upcroft B, Perez T (2016) Sweet pepper pose detection and grasping for automated crop harvesting. In: 2016 IEEE international conference on robotics and automation (ICRA), pp 2428–2434, IEEE
17. Sa I, Lehnert C, English A, McCool C, Dayoub F, Upcroft B, Perez T (2017) Peduncle detection of sweet pepper for autonomous crop harvesting-combined color and 3-d information. IEEE Robot Autom Lett 2(2):765–772
18. Hinterstoisser S, Pauly O, Heibel H, Martina M, Bokeloh M (2019) An annotation saved is an annotation earned: using fully synthetic training for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision workshops
19. Khan S, Phan B, Salay R, Czarnecki K (2019) Procsy: Procedural synthetic dataset generation towards influence factor studies of semantic segmentation networks. In: CVPR workshops, pp 88–96
20. Di Cicco M, Potena C, Grisetti G, Pretto A (2017) Automatic model based dataset generation for fast and accurate crop and weeds detection. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 5188–5195, IEEE
21. Olatunji J, Redding G, Rowe C, East A (2020) Reconstruction of kiwifruit fruit geometry using a cgan trained on a synthetic dataset. Comput Electron Agric 177:105699
22. Zhang K, Wu Q, Chen Y (2021) Detecting soybean leaf disease from synthetic image using multi-feature fusion faster r-cnn. Comput Electron Agric 183:106064
23. Barth R, Isselmuiden J, Hemming J, Van Henten EJ (2018) Data synthesis methods for semantic segmentation in agriculture: a capsicum annuum dataset. Comput Electron Agric 144:284–296
24. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision, pp 21–37, Springer
25. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
26. Maric B, Polic M, Tabak T, Orsag M (2020) Unsupervised optimization approach to in situ calibration of collaborative human-robot interaction tools. In: 2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI), pp 255–262, IEEE
27. Hess R (2010) Blender foundations: the essential guide to learning blender 2.6. Focal Press
28. AliceVision, Meshroom: A 3D reconstruction software., 2018

.2 PUBLICATION 2 - COMPLIANT PLANT EXPLORATION FOR AGRICULTURAL PROCEDURES WITH A COLLABORATIVE ROBOT

M.Polic, M.Car, F.Petric, and M.Orsag Compliant Plant Exploration for Agricultural Procedures With a Collaborative Robot. *IEEE Robotics and Automation Letters*, 6(2), 2768-2774, 2021, IF: 3.74 (Q1).

Compliant Plant Exploration for Agricultural Procedures With a Collaborative Robot

Marsela Polic, Marko Car, Frano Petric, and Matko Orsag

***Abstract*—This letter presents a compliant exploration framework based on a collaborative robot Franka Panda that builds a 3D plant stem model. The model is built for agricultural plant treatment procedures without external sensors, as contact forces are estimated from joint torques and robot's dynamic and kinematic model. By devising an impedance-based exploratory control algorithm capable of following an unknown shape, while being provided with only a general direction in which to explore, we eliminate the need for a precise position controller. Our approach is validated through experiments with several mock-up plant stems, showing that the proposed framework is capable of building a satisfactory 3D model of a plant. The method is evaluated against the ground truth model, and compared to the state of the art approach based on an industrial manipulator with external sensors.**

***Index Terms*—Compliance and impedance control, force and tactile sensing, robotics and automation in agriculture and forestry.**

Fig. 1. Collaborative robot Franka Panda performing the exploration procedure on a Vitis vinifera plant stem using a rigid end-effector gripper.

I. Introduction

WHILE being on the forefront of sustainable agriculture with minimal environmental impact, organic agriculture faces economic challenges. With decreased use of pesticides and GMO cultures, farmers need to put in more time to generate smaller yield than traditional agriculture, making organic food less affordable due to higher cost of production. We believe some of these costs can be offset by using a heterogeneous team of robots in a small indoor farm, which we are deploying through the SpECULARIA project [1]. The crucial component of this heterogeneous team of robots is a robotic manipulator capable of treating plants. When interacting with plants or other fragile everyday objects in unstructured environments, apart from the mechanically soft components, the robot has to sense and control the force it applies to the environment. Due to their inherent safety and compliance, collaborative robots present an obvious choice for the job. For this experiment we chose an affordable collaborative robot, shown treating a vine stem in Fig. 1.

In the envisioned greenhouse, the compliant manipulator, designed following the soft robotics paradigm, should be able to autonomously conduct plant treatment procedures like pinching, topping, pruning or removing excess flowers. Treating a plant is inherently a delicate procedure that requires either precise positioning of the robot's end effector, or a sense of touch. In robotics, spatial models of objects are usually acquired using vision or lidar based technologies, which fail to provide an adequate model if the object, in this case the stem, is occluded (e.g. leaves or flowers). Furthermore, when dealing with complex objects such as plants, it is often hard to program the exact scanning procedures that can predict the plant shape a-priori while respecting the robot's workspace limits. In this letter we explore the possibility of using a robotic manipulator to clean vine stems from scrubs and old rind. The procedure is depicted in Fig. 2, and can be applied to other cultures, fruits and olives, when it is necessary to clean the lower stems of the plant in order to allow the rest of the plant to thrive. Prior knowledge of the sprouting position of the plant stem within the growth containers is used as an input for compliant exploration. Once the robot achieves a loose grip around the plant, it begins to explore the stem. The exploration yields a crude 3D model of the plant, precise enough to allow the arm to grip tight around the plant, and follow a 3D model based trajectory planned for cleaning the stem. Capturing important physical characteristics of the plant stem enables execution of manipulation procedure, which revolves around the concept of exerting sufficient force only where necessary, and leaving the rest of the plant stem unharmed.

To perform effective exploration of a plant, we propose an impedance based framework that exploits joint torque interface of a Franka Panda collaborative robot to estimate external forces without the need for additional external sensory apparatus, either

Manuscript received October 15, 2020; accepted February 6, 2021. Date of publication February 25, 2021; date of current version March 17, 2021. This letter was recommended for publication by Associate Editor G. Palli and Editor C. Gosselin upon evaluation of the reviewers' comments. This work was supported by the Faculty of Electrical and Computer Engineering, University of Zagreb and in part by the Croatian Science Foundation under the project Specularia UIP-2017-05-4042 [1]. *(Corresponding author: Marsela Polic.)*

The authors are with the Faculty of Electrical and Computer Engineering, University of Zagreb, Zagreb 10000, Croatia (e-mail: marsela.polic@fer.hr; marko.car@fer.hr; frano.petric@fer.hr; matko.orsag@fer.hr).

Digital Object Identifier 10.1109/LRA.2021.3062301

Fig. 2. Cleaning vine stems from scrubs and old rind starts with localizing and perching the plant proceeding to plant exploration moving upwards. Once we reach a certain threshold, the exploration stops, the grip is tightened and the cleaning procedure starts from the top.

force/torque, visual or tactile sensors. Even without external sensors and relying on low-precision force estimation by the robot, the framework is able to generate a 3D model of the plant sufficiently precise for the chosen agricultural procedure, as shown through experiments with real plants in a structured greenhouse. Our approach also goes beyond the state of the art by relaxing requirements on the position control of a robot, since it is only provided with the general direction of a desired movement, and not the exact position reference, while maintaining the ability to trace the contour of the plant even in presence of noise. Additionally, we generate an optimisation based online local plant model that enables end effector orientation control. We validate the exploration framework with specifically designed end effector on real plants. One of the validation criteria is the ability to remove the laterals without harming the rest of the plant, using exploration results as the only input for end effector position control.

Following this introductory section, we position our research with respect to state of the art in Section II. In Section III we propose our tactile exploration framework which is evaluated through experiments shown in Section IV. We conclude the letter with final remarks and guidelines for future work in Section V.

II. Related Work

The plant manipulation task described in this work, as most of the work in the research of active touch, heavily relies on human dexterity imitation. A set of specialized exploratory procedures has been described in [2]. Based on human exploration strategies, a various set of problems has been more or less successfully solved [3]. Within this work, the experimental setup, procedures, and control methods have been task-specifically designed. One of the fundamental task-specific design decisions was a robotic gripper in the enclosure form [2]. This form is chosen as a simplification of the envisioned two fingered gripper, that would be able to enclose the plant parts in grasping procedures.

Over the previous decades, a lot of research effort has been put into designing optimal exploration strategies. In many applications, surface exploration is considered [4]–[6]. A similar approach relies on edge (contour) following [7]–[9]. Different methods are deployed for the next-best move decision, e.g. based on Gaussian Processes [10], decision trees [9], and learning [7], [11]. In the presented use-case, the enclosure gripper inherently constrains exploration into movement along a single axis, thereby reducing the degrees of freedom for search direction optimization. This constraint is modeled within the exploration step extending the constrained global motion planning proposed in [12] to local planning (servoing).

The surface and edge following approaches usually require controllers with both position and force tracking capabilities. [13]–[15]. Here, an impedance filter is deployed, since a strict control of position is not required, and the force tracking problem is relaxed to a force minimisation problem. This is achieved thanks to both the gripper characteristics, and the joint torque interfaced cobot, which allow us to command a zero desired contact force along with the direction of spatial exploration.

Furthermore, the range of tactile technologies has been rapidly increasing [16]. Depending on the capabilities and sensing modalities of the deployed tactile technology, specific methods have been proposed [17]–[20]. The availability of multi-finger hands and other more complex end effectors has opened new opportunities for human dexterity formalization [21], [22]. Within this research, a simple passive gripper is deployed without tactile feedback. Thanks to the built-in capabilities of the cobot, the exploration is driven by force measurements provided by joint torque linearized transformation.

While this work is focused on the exploration of a plant, it is driven by the goal of the SpECULARIA project, namely being able to conduct various plant treatment procedures autonomously, such as cleaning and pinching the laterals. Since such delicate procedure requires precise description of a plant, optimal model representation is also considered. Depending on the intended use of the acquired object information, different formats and model generation methods are proposed [21], [23]. In the work presented in this letter, a local curvature estimation procedure is deployed online, similar to [24]. This local estimate is used within the exploration procedure. The global model is generated offline in a form of a point cloud, based on the recorded robot positions, then additionally processed to generate a set of 3D points that best describe the plant stem. This format allows for efficient trajectory planning for plant treatment procedures, since calculated 3D points can be directly used as end effector waypoints.

III. Method

The procedure described in the letter, and others such as pinching or topping, require application of external force only to the relevant plant parts. For such confined application of external force in an autonomous manner, an adequate 3D model of the plant is needed.

Within this project, the plants are grown in containers of a structured greenhouse, and their global position, i.e. the sprouting position, is known through an auxiliary system. The sprouting position is used as an input to the exploration method. The proposed method explores the main plant stem compliantly

through the underlying impedance filter. This basic approach takes into consideration plants of low ramification, i.e. plants with a main stem and soft lateral branches or leaves, such as the wine and currant plants. Further development of the method, and corresponding end effector design, can allow generalization to other plant anatomies, e.g. with improvements towards detection and handling of branching, as explained further in text.

A. Robot Manipulator

The task involving manipulation of sensitive objects was approached using a Franka Panda collaborative robot manipulator (cobot). The lightweight cobots such as Franka Panda can usually be controlled through both joint position and torque control interfaces. This work relies on the possibility of estimating the external force vector $\mathbf{F}_{ext}$ with the built-in joint torque interface. Namely, an approximate relation of the external forces $\mathbf{F}_{ext}$ and measured joint torques $\vec{\tau}_{meas}$ 1 is derived using the body Jacobian matrix, the energy conservation law, and model-based estimations of torques resulting from inertial forces ($\vec{\tau}_{coriolis}$, and $\vec{\tau}_{gravity}$). In the case of Franka Panda, the identified robot model is provided as a black box to the users.

$$\begin{aligned}\mathbf{F}_{ext} &= (\mathbf{J}^T)^{-1} \cdot \vec{\tau}_{ext} \\ &= (\mathbf{J}^T)^{-1} \cdot (\vec{\tau}_{meas} - \vec{\tau}_{gravity} - \vec{\tau}_{coriolis}) \end{aligned} \quad (1)$$

In this work, we explore the possibilities of replacing the external sensory apparatus such as force-torque sensors with the torque based force estimation in selected control scenarios. This represents a trade-off between high precision of industrial manipulation and sensory apparatus, and the reduced deployment price for the small scale production systems in organic agriculture based on an inherently safe collaborative robotic system. The latter approach comes with a drawback of less reliable force estimates, e.g. in joint configurations where the inversion of Jacobain fails 1. These are however rare in normal operation, and were not encountered in this work.

An important assumption of this method is the localized external force application point at the robot end effector. This can result in significant estimation error in case a contact with another robot link occurs and is unaccounted for, similar to the problem of industrial position based manipulator collision. One typical approach is restriction of access into the robot work space. Another, more expensive approach would include additional sensory equipment such as robot skin, or visual surveillance system. This is however in contrast with our objective, which is to sustainably reduce the cost of organic agriculture. A more practical solution could also be implemented based on torque distribution analysis, which could to an extent enable contact detection, based on expected joint torque values and dynamics, but this interesting research topic is beyond the scope of this letter.

In the framework presented here, the robot is controlled through the joint torque interface by means of a force feedback closed-loop PI controller. As shown in Fig. 3, the force reference provided by the controller is mapped into the desired joint torques, using the inverse of 1.

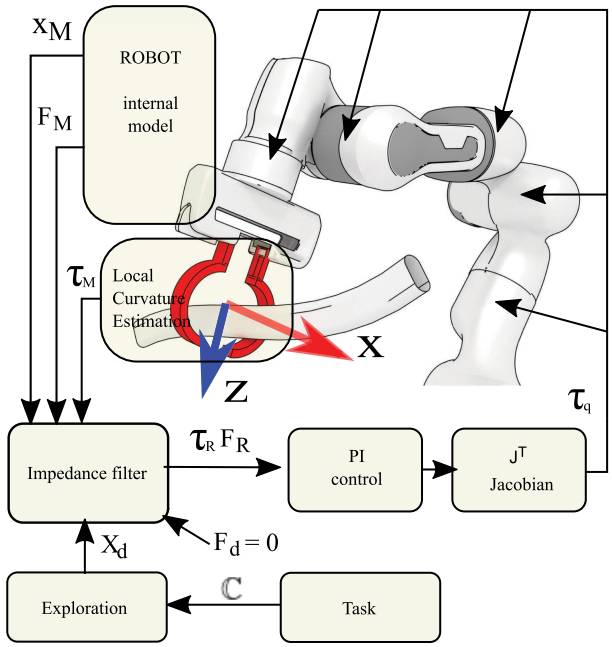


Fig. 3. Block diagram of a closed loop control of a cobot for tactile exploration based on force and torque estimation from the internal kinematic and dynamic model of a cobot. The goal of the control algorithm is to maintain the orientation of the end-effector so that the exploration is possible along x-axis,but restricted in other two axes.

B. Exploration

In the structured greenhouse workspace, an estimate on the plant sprouting position is available through the predefined container positions. Based on this initial information, the manipulator can be positioned to an exploration starting position near the sprouting point. Then, the compliant exploration procedure is deployed, based on two basic paradigms: impedance control, and tactile servoing. The exploration is constrained along the plant stem. These constraints are introduced into servoing, a *local* motion planning method, by extending the concepts from the constrained *global* motion planning [12].

The robot motion planning is defined within the robot coordinate system. On the other hand, the plant exploration task is defined in the global frame, by plant position and its growth direction (roughly the global z axis). In addition, the robot end effector design with fingers enclosure allows movement along the end-effector **local** x axis. These restrictions of exploration motion are introduced by means of a selection matrix $\mathbb{C}$. The exploration motion, defined in the task frame as "along plant stem", can be transformed into the allowed world frame motion $\Delta\vec{x}_{global}$ by 2. Rigid transformation between the task frame and end effector frame propagates the task restriction (plant growth) to the allowed end effector space (movement along gripper enclosure): $\Delta\vec{x}_{ee} = \mathbb{C}\mathbf{T}^{ee}_{task}\Delta x_{task}$. Due to the design of the parallel (enclosure) gripper, movement along other axes would result in non-zero contact forces, as depicted in Fig. 3. In this particular use-case, the transformation matrix $\mathbf{T}^{ee}_{task}$ between task and end-effector frames consists only of rotation component, which describes the relation between the local end-effector

exploration x axis, and the plant growth direction at the current plant exploration point in the global frame (roughly global z axis). There is no translation in the transform because the end-effector position $\vec{x}_{global}$ in the world frame coincides with the plant contact point position. The robot end-effector is however not controlled in the end-effector frame, but in the global frame, and an additional transform of constrained reference has to be calculated into the global frame with (2).

$$\Delta\vec{x}_{global} = \mathbf{T}^{ee}_{global}\Delta\vec{x}_{ee} = \mathbf{T}^{global}_{ee}\mathbb{C}\mathbf{T}^{ee}_{task}\Delta\vec{x}_{task} \tag{2}$$

By applying an impedance filter to the (constrained) exploration motion reference, the exploration procedure attains compliance. The impedance filter (3) generates a reference force $\mathbf{F}_r$, depending on the relation of the desired contact force $\mathbf{F}_d$ and position $\vec{x}_d$ with the measured force $\mathbf{F}_m$ and position $\vec{x}_m$.

$$\begin{aligned}\mathbf{F}_r =& m\ddot{\vec{x}}_d - \frac{m}{M}(\mathbf{D}\left(\dot{\vec{x}}_m - \dot{\vec{x}}_d\right) + \mathbf{K}(\vec{x}_m - \vec{x}_d)\Big) \\ &+ \left(\frac{m}{M} - 1\right)(\mathbf{F}_m - \mathbf{F}_d) - \mathbf{F}_d\end{aligned} \tag{3}$$

The filter is designed so that the position error acts as a spring with mass M, damping $\mathbf{D}$ and spring constant $\mathbf{K}$, which are design parameters that can be selected according to the use case. The robot end-effector is modeled as an inertial body m whose motion is a result of both external (contact) forces $\mathbf{F}_m$ and the forces produced by joint torques $\mathbf{F}_r$. The filter takes into account both contact force minimisation, and exploration objectives. In the servo control, the desired motion is defined in the global frame in incremental movements $\Delta\vec{x}_d$ relative to the current robot position (4). The other filter inputs define the desired zero contact force $\mathbf{F}_d = 0$, and constant velocity ($\dot{\vec{x}}_d = 0$, $\ddot{\vec{x}}_d = 0$). The compliant exploration controller is then described with (5). The designed exploration controller thus generates a virtual force field, driving the robot end effector towards the unexplored areas of interest with constant velocity, simultaneously reacting to undesired contact forces and moments.

$$\Delta\vec{x}_d = \vec{x}_d - \vec{x}_m = \mathbf{T}^{global}_{ee}\mathbb{C}\mathbf{T}^{ee}_{task}\Delta\vec{x}_{task} \tag{4}$$

$$\mathbf{F}_r = \left(\frac{m}{M} - 1\right)\mathbf{F}_m - \frac{m}{M}\left(\mathbf{D}\dot{\vec{x}}_m - \mathbf{K}\Delta\vec{x}_d\right) \tag{5}$$

Due to the fact that the external torque estimation based on (1) is not precise enough, the equation (3) only applies to the forces, thus controlling the end effector position. The referent torques are computed based on the local curvature estimation, effectively controlling the end effector orientation.

C. Local Curvature Estimation

The control of external torques through the impedance filter is impaired due to task and setup characteristics: firstly, the design of the exploration gripper for plant stem grasping results in high transmission factor from the contact to the measurement point, i.e. in low and unreliable external torque measurements. Secondly, the sensitive torque estimation is based on an approximate dynamic robot model, heavily dependent on the robot pose. Therefore a local plant stem curvature is predicted and used in torque control instead.

The curvature estimation is based on a simple polynomial fitting least squares optimization. The plant is locally modelled as a 3D polynomial curve. As the robot moves along the plant, 3 rd order polynomials ($p_i, \forall i \in \{x, y, z\}$, with parameters a_i, b_i, c_i, d_i) are fitted for each of the three global axes based on a window of recorded gripper positions (6a). The polynomial derivative in the current gripper position is used as prediction of the local plant curvature. This derivative can simply be obtained using 6b, if the end-point is chosen to correspond to parametric time $t = 0$ (and the start point to $t = -1$).

$$p_i(t) = a_i \cdot t^3 + b_i \cdot t^2 + c_i \cdot t + d_i \tag{6a}$$

$$\dot{p}_i(t = 0) = c_i \tag{6b}$$

The polynomial fitting is conducted on buffer points collected at a predefined minimum euclidean distance. The buffer size and minimum distance determine the amount of the plant we model, and are chosen taking into account the species characteristics. The plant curvature estimation is obtained as a normalized direction vector $\vec{o}_r$ in the world reference frame, 7, using optimization result and 6b:

$$\vec{o}_r = \frac{1}{\|-\dot{p}_x, -\dot{p}_y, -\dot{p}_z\|}[-\dot{p}_x, -\dot{p}_y, -\dot{p}_z]^\top \tag{7}$$

This estimate is used as a feedback of the exploration PI controller that controls the external moments, given with 8. An additional stochastic perturbation $\vec{\tau}_{rand}$ is introduced to compensate for the possible error in curvature prediction. The stochastic perturbation is generated from a uniform distribution, but with a gentle preference to y axis rotation, since this direction opposes robot end effector exploratory movements. The amplitude of the perturbation is inversely proportional to the end effector velocity $\dot{\vec{x}}_m$, so that the effect of the perturbation is only present when faulty predicted exploration direction results in zero exploration velocity. In other words, similar to human blind exploration process, the robot tries to wiggle itself out once it gets stuck.

Putting this all together yields an augmented PI control law, written in the form of a vector equation:

$$\begin{aligned}\vec{\tau}_r =& \mathbf{K}_p(\vec{o}_r - \vec{o}_m) + \mathbf{K}_i\sum(\vec{o}_r - \vec{o}_m)\Delta t \\ &+ k_r\vec{\tau}_{rand}(\dot{\vec{x}}_m),\end{aligned} \tag{8}$$

with $\mathbf{K}_p$ and $\mathbf{K}_i$ denoting 3×3 diagonal matrices of respective proportional and derivative gains of each axis in Cartesian space. k_r is a scalar gain, that regulates the amount of random wiggle motion.

D. 3D Object Model

The objective of the described exploration procedure is generation of a 3D model of a plant. A local plant model is generated online using simple polynomial fitting, and used only as a local approximation. In practice, the plant stem can be of arbitrary complexity, and poorly fitted to even high order polynomials, depending on the plant species. Instead, a global spatial model in form of a point cloud is chosen. The point cloud is generated

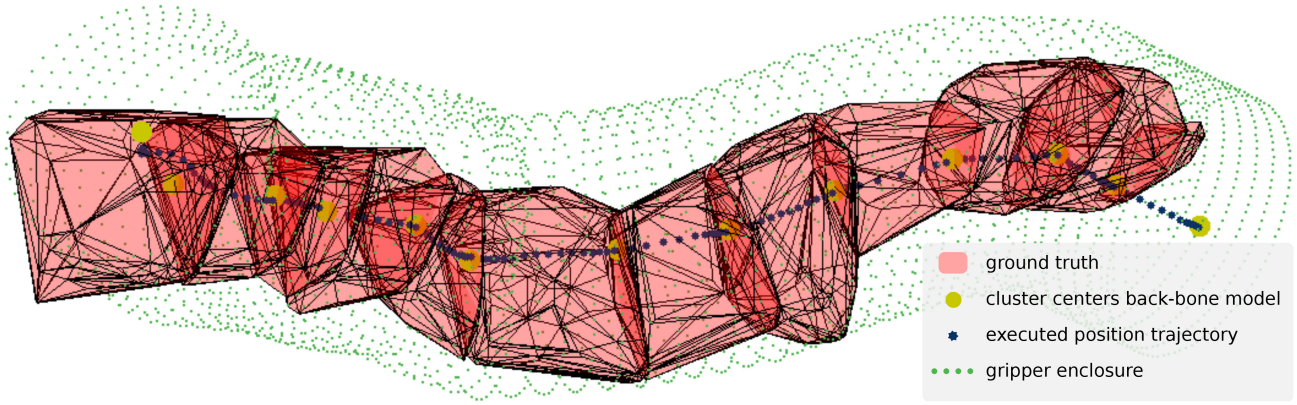


Fig. 4. Comparison of 3D models. The clustering based back-bone model generated through exploration (yellow) is used as trajectory planning input. The executed position trajectory (blue) deviates slightly from the starting point, until it smoothly reaches the planned trajectory. The ground truth model of the plant stem is obtained fusing multiple views of a depth camera into a single model. Ground truth shows convex hulls of this model's segments, split at the cluster centres (red). The image also shows the gripper enclosure during the procedure execution (green).

using the end-effector positions saved during exploration. This step can be conducted both offline and online.

The data-rich point cloud is processed using an unsupervised learning clustering method. By replacing a large amount of data with a single representative, in this case a cluster centre, we enable online decision making and control in plant treatment procedures. In particular, we use Agglomerative clustering method [25] with complete linkage criterion. Since the agglomerative clustering only labels the data, the cluster centres have to be generated otherwise. Here, we use Central feature algorithm that chooses the cluster center as the cluster member with the shortest summed distance to all other members.

E. Exploration Gripper

A simple custom gripper design is proposed for plant stem exploration and modelling. As shown in Fig. 3, it is designed as a two finger parallel gripper, with fingers curved in such a way that they form an enclosure [2], thus only allowing exploration along its local x axis. The design is a simplification of a two fingered gripper, with the fingers rigidly attached, for exploration method validation. In the following design iteration, the fingers of a similar actuated gripper would form an enclosure during exploration. Then, during the actual plant manipulation procedure, a firmer grasp of the fingers would apply tangential force along the stem, enabling removal of lateral branches and leaves.

Another important practical improvement when using such an actuated gripper would be in case the plant body contains significant side branches in addition to the minor laterals that are being cleaned. In this case, the stem could be released by the fingers at the bifurcation point. The robot would then plan a motion that evades the branch, and re-grasp the plant further along the stem. This would be achieved using the described local curvature model, fused with information from the available part of the global model. This is however beyond the scope of this work, since here the compliant exploration method is validated on simpler mock-up and real plant stems.

The gripper was manufactured using PLA plastic. Due to the fact that the size of the explored models varied significantly, and the fingers are not actuated to form enclosures of different sizes, several models were designed and manufactured, and later deployed in exploration of objects of corresponding average size.

IV. Experimental Results

The validation of the method is conducted in several ways in a laboratory environment. Firstly, a comparison is made with the manually generated ground truth, similar to the existing tactile exploration work. The precision of the resulting model is further experimentally validated, in an experiment where the robot follows the plant shape in a position control scheme, simulating the stem cleaning procedure. The method reliance on the inherent cobot capabilities is compared to the classical industrial approach with a commercial force/torque (f/t) sensor.

A. Plant Modeling

The agglomerative clustering applied on the recorded positions during exploration produced a set of trajectory waypoints that represent the 3D model of the plant stem. The number of waypoints is chosen such that each cluster center represents approximately 2-3 cm of the plant stem length. A trajectory optimized through these waypoints is used in manipulation procedure as position reference for the robot. Similar to other work in the field of tactile exploration and 3D object model generation where directly comparable ground truth is not available ([26]–[28]), the obtained model (set of representative points) is compared to the manually obtained ground truth 3D model.

Here, the ground truth model is generated through manual fusion of multiple depth images from a commercial depth camera (Intel RealSense D345). The comparison is shown in Fig. 4. The ground truth model, transparent red in the figure, is split into segments at the cluster centres, shown with yellow dots. For clarity, a convex hull of each segment is shown instead of a raw point cloud. A planned trajectory is visible within the model, as well as the cylindrical hull drawn by the robot end effector

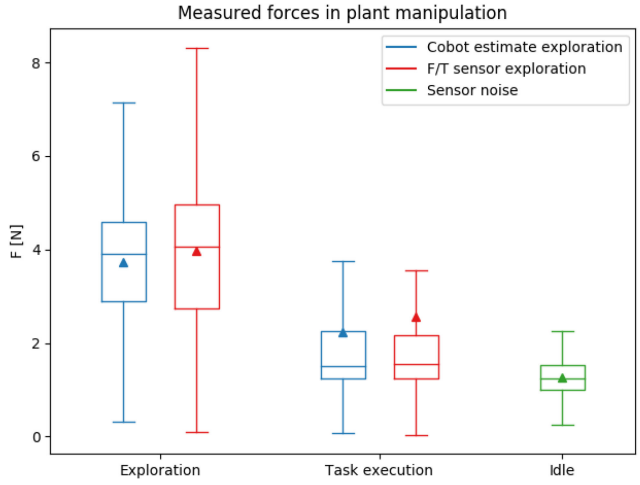


Fig. 5. Distribution of contact forces in various parts of the plant manipulation procedure. Graph compares results for 5 experiment runs based on cobot force estimates, and 5 runs based on commercial f/t sensor measurements. Data is recorded with f/t sensor in both cases. Contact forces are reduced almost to the level of noise during task execution using the generated model.

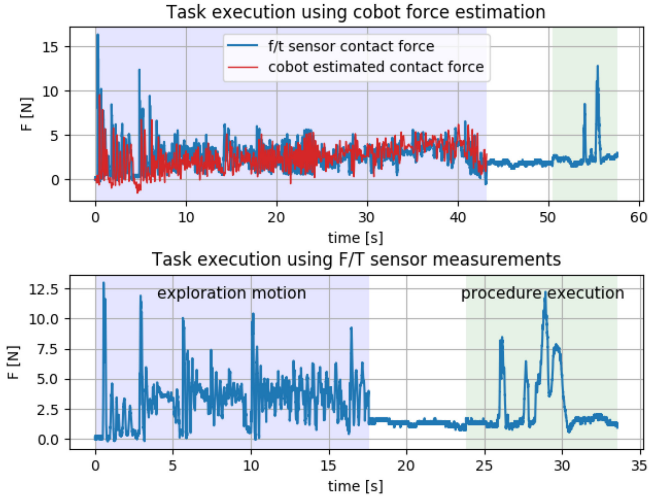


Fig. 6. Measured contact forces over time in task execution. Two random runs are chosen, one based on cobot force estimates (red), and the other on f/t sensor measurements (blue). Both graphs show the f/t sensor measurements, during exploration and procedure execution on the same vine plant. Behaviour is comparable, with the first case taking longer overall.

while traversing the stem along the trajectory (green). The figure shows the overlap of the ground truth model and the generated model, with the entire planned trajectory contained within the ground truth model, and the entire ground truth model within the end effector cylindrical hull.

B. Evaluation

Method evaluation was conducted with a commercial force/torque sensor (Optoforce HEX 70) mounted on the robot gripper. These experiments served two purposes: firstly, to verify the compliance through precise contact force measurements, and secondly, to compare the cobot-based exploration to a classic approach using a position controlled robot with external sensory apparatus. Two sets of experiments were conducted, one with cobot force estimates as feedback signal, and the other using f/t sensor instead. The results showing the distribution of realized contact forces are shown in the box plot in figure 5. The figure shows that the realized forces are within a reasonably safe range for plant manipulation, on average smaller than 5 N. They unsurprisingly show that the model generation results in a smoother and more precise motion, with less contact and lower measured contact forces. In fact, most of the measured contact force amplitude can be ascribed to the sensor noise, when comparing the task execution distributions with that in idle state.

Fig. 6 additionally justifies the proposed approach, by showing that the cobot force estimates can serve as an adequate replacement for precise but expensive external f/t sensor in practical production problems. Aside from the discussed issue regarding the unknown origin of measured joint torques, anyway present even in case a classical industrial manipulator is used, the other drawback in terms of task execution can be observed. The lower sensitivity in the estimates deteriorates the method performance when compared to more reliable f/t sensor measurements, resulting in a slower exploration, due to the milder controller reactions.

The precision of the obtained model was validated experimentally, proving the cobot capable to conduct a manipulation task in a safe manner. Stem cleaning procedure was performed by executing a trajectory through model waypoints. The cobot was controlled through a collaborative position control interface. The motion is allowed through free space, with reactive stopping behaviour in case a hazardous contact is detected. Here we report that in the trials performed, the cobot halt during trajectory execution was never observed. The video recordings on both the mock-up pipe model, and a real vine plant can be found on [29]. Since the robot managed to complete the given position trajectory, both with the pipe model, and on a real plant, we can conclude that the model is precise enough to ensure harmless manipulation in direct vicinity of a fragile object based on the built 3D model.

V. Conclusion

An impedance based framework for compliant plant exploration is presented, along with experimental results validating the potential of the method in agricultural procedures. Relying on the collaborative manipulator capabilities, adequate 3D models of explored objects were constructed without using external sensory apparatus. The method manages to overcome the limitations of force estimation using cobot capabilities, with result comparable to the more expensive case where external sensors are used. These successful trials are among the first steps for introduction of affordable and safe robots into the food production industry. Not only are cobots safe for human co-workers and generally less expensive than the industrial manipulators, but also more convenient for non-experts, e.g. through programming by demonstration.

The positive results of this study allow for other future research directions. Taking advantage of the tested enclosure form, we plan to test the method using more elaborate grippers, starting with actuation of the gripper fingers. This will enable a more elaborate plant model generation that would take into account local plant deformation under the influence of force. It will also enable implementation of a compliant control method for grasping in plant treatment procedures. Introduction of more elaborate tactile sensing will also be tested in practice, retaining the sensing capabilities along with a soft bodied mechanical compliance with rubber based fingertips. Under-actuated multi-DOF enclosure fingers will also be considered in the presented task. Future work should also take into account the robustness of a low cost approach, since in certain joint configurations the Jacobian inverse will fail to provide force estimate.

REFERENCES

[1] M. Orsag *et al.* "Specularia." 2019. Accessed: Sep. 8, 2019. [Online]. Available: http://specularia.fer.hr

[2] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cogn. Psychol.*, vol. 19, pp. 342–368, Jul. 1987.

[3] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[4] S. Ahmad and C. N. Lee, "Shape recovery from robot contour-tracking with force feedback," *Adv. Robot.*, vol. 5, no. 3, pp. 257–273, 1990.

[5] Z. Doulgeri and Y. Karayiannidis, "Force/position regulation for a robot in compliant contact using adaptive surface slope identification," *IEEE Trans. Autom. Control*, vol. 53, no. 9, pp. 2116–2122, Oct. 2008.

[6] C. Rosales, F. Spinelli, M. Gabiccini, C. Zito, and J. L. Wyatt, "GPAtlas-RRT: A local tactile exploration planner for recovering the shape of novel objects," *Int. J. Humanoid Robot.*, vol. 15, Feb. 2018, Art. no. 1850014.

[7] N. F. Lepora, K. Aquilina, and L. Cramphorn, "Exploratory tactile servoing with active touch," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 1156–1163, Apr. 2017.

[8] U. Martinez-Hernandez, T. J. Dodd, L. Natale, G. Metta, T. J. Prescott, and N. F. Lepora, "Active contour following to explore object shape with robot touch," in *Proc. IEEE World Haptics Conf.*, Apr. 2013, pp. 341–346.

[9] K. S. Roberts, "Robot active touch exploration: Constraints and strategies," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1989, pp. 980–985.

[10] M. Regoli, N. Jamali, G. Metta, and L. Natale, "Controlled tactile exploration and haptic object recognition," in *Proc. IEEE 18th Int. Conf. Adv. Robot.*, 2017, pp. 47–54.

[11] X. Yan, A. Knott, and S. Mills, "A model for learning representations of 3d objects through tactile exploration: Effects of object asymmetries and landmarks," in *Proc. Australas. Joint Conf. Artif. Intell.*, 2018, pp. 271–283.

[12] M. Stilman, "Global manipulation planning in robot joint space with task constraints," *IEEE Trans. Robot.*, vol. 26, no. 3, pp. 576–584, Jun. 2010.

[13] S. Jung, T. C. Hsia, and R. G. Bonitz, "Force tracking impedance control of robot manipulators under unknown environment," *IEEE Trans. Control Syst. Technol.*, vol. 12, no. 3, pp. 474–483, May 2004.

[14] R. E. Goldman, A. Bajo, and N. Simaan, "Algorithms for autonomous exploration and estimation in compliant environments," *Robotica*, vol. 31, pp. 71–87, Mar. 2012.

[15] T. Yoshikawa and A. Sudou, "Dynamic hybrid position/force control of robot manipulators-on-line estimation of unknown constraint," *IEEE Trans. Robot. Automat.*, vol. 9, no. 2, pp. 220–226, Apr. 1993.

[16] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robot. Auton. Syst.*, vol. 74, pp. 195–220, 2015.

[17] N. Sommer, M. Li, and A. Billard, "Bimanual compliant tactile exploration for grasping unknown objects," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 6400–6407.

[18] A. M. Okamura and M. Curkosky, "Feature-guided exploration with a robotic finger," in Proc. ICRA. *IEEE Int. Conf. Robot. Automat. (Cat. No 01CH37164)*, 2001, vol. 1, pp. 589–596.

[19] M. Schultz and V. Dürr, "Object localisation with a highly compliant tactile sensory probe via distributed strain sensors," in *Proc. Conf. Biomimetic Biohybrid Syst.*, 2018, pp. 428–438.

[20] H. Liu *et al.*, "Finger contact sensing and the application in dexterous hand manipulation," *Auton. Robots*, vol. 39, no. 1, pp. 25–41, 2015.

[21] A. Bierbaum, I. Gubarev, and R. Dillmann, "Robust shape recovery for sparse contact location and normal data from haptic exploration," in Proc.*IEEE*/RSJ Int. Conf. Intell. Robots Syst., Sep. 2008, pp. 3200–3205.

[22] A. Bierbaum, M. Rambow, T. Asfour, and R. Dillmann, "A potential field approach to dexterous tactile exploration of unknown objects," in *Proc. Humanoids 8th IEEE-RAS Int. Conf. Humanoid Robots*, 2008, pp. 360–366.

[23] S. Ottenhaus, M. Miller, D. Schiebener, N. Vahrenkamp, and T. Asfour, "Local implicit surface estimation for haptic exploration," in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots (Humanoids)*, 2016, pp. 850–856.

[24] A. Fedele, A. Fioretti, C. Manes, and G. Ulivi, "On-line processing of position and force measures for contour identification and robot control," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1993, pp. 369–374.

[25] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classification*, vol. 1, no. 1, pp. 7–24, 1984.

[26] S. Ottenhaus, P. Weiner, L. Kaul, A. Tulbure, and T. Asfour, "Exploration and reconstruction of unknown objects using a novel normal and contact sensor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1614–1620.

[27] Z. Yi *et al.*, "Active tactile object exploration with gaussian processes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4925–4930.

[28] S. Caccamo, Y. Bekiroglu, C. H. Ek, and D. Kragic, "Active exploration using gaussian random fields and gaussian process implicit surfaces," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 582–589.

[29] Laboratory for robotics and intelligent control systems youtube channel, "Compliant tactile exploration for 3D object reconstruction," Accessed: 2021-03-08. [Online]. Available: https://youtu.be/dmIiGECj9Yc

.3 PUBLICATION 3 - CONVOLUTIONAL AUTOENCODER FOR FEATURE EXTRACTION IN TACTILE SENSING

M.Polic, I.Krajacic, N.Lepora, and M.Orsag Convolutional autoencoder for feature extraction in tactile sensing. *IEEE Robotics and Automation Letters*, 4(4), 3671-3678, 2019, IF: 3.6 (Q1).

Convolutional Autoencoder for Feature Extraction in Tactile Sensing

Marsela Polic, Ivona Krajacic, Nathan Lepora, and Matko Orsag

***Abstract*—A common approach in the field of tactile robotics is the development of a new perception algorithm for each new application of existing hardware solutions. In this letter, we present a method of dimensionality reduction of an optical-based tactile sensor image output using a convolutional neural network encoder structure. Instead of using various complex perception algorithms, and/or manually choosing task-specific data features, this unsupervised feature extraction method allows simultaneous online deployment of multiple simple perception algorithms on a common set of black-box features. The method is validated on a set of benchmarking use cases. Contact object shape, edge position, orientation, and indentation depth are estimated using shallow neural networks and machine learning models. Furthermore, a contact force estimator is trained, affirming that the extracted features contain sufficient information on both spatial and mechanical characteristics of the manipulated object.**

***Index Terms*—Force and tactile sensing, deep learning in robotics and automation, soft sensors and actuators, perception for grasping and manipulation.**

I. Introduction

ONE of our long term goals in project SPECULARIA [1] is to develop robotic manipulators capable of treating plants. When interacting with plants or other fragile everyday objects in unstructured environments, aside from the mechanically soft components, the robot has to be equipped with the ability to sense and control the force it applies to the environment. Typical manipulation problems include object recognition and exploration, grasp stability estimation and slip detection, force control, and tactile servoing [2]. For such interaction problems, as well as for complex precise manipulation tasks, tactile feedback is essential [3]. Different soft sensing technologies are being explored, in attempt to develop compliant mechanical structures that simultaneously provide tactile feedback, and satisfy safety requirements of close human-robot interaction.

Manuscript received February 24, 2019; accepted June 21, 2019. Date of publication July 10, 2019; date of current version July 26, 2019. This letter was recommended for publication by Associate Editor K. H. Petersen and Editor K.-J. Cho upon evaluation of the reviewers' comments. This work was supported by the Croatian Science Foundation under the Project Specularia UIP-2017-05-4042. *(Corresponding author: Marsela Polic.)*

M. Polic, I. Krajacic, and M. Orsag are with the Laboratory for Robotics and Intelligent Control Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb 10000, Croatia (e-mail: marsela.polic@fer.hr; ivona.krajacic@fer.hr; matko.orsag@fer.hr).

N. Lepora is with the Department of Engineering Mathematics and Bristol Robotics Laboratory, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: n.lepora@bristol.ac.uk).

Digital Object Identifier 10.1109/LRA.2019.2927950

Fig. 1. Experimental setup used in dataset acquisition. The TacTip sensor is mounted on a Franka Emika 7DoF arm, and contact is realized with rigid objects of various shapes and sizes. Optoforce sensor was used to obtain force measurement dataset.

One of the contributions to this body of knowledge in this work is extending the capabilities of a family of tactile sensing technologies with an unsupervised feature extraction method based on convolutional neural network (CNN) autoencoder (CAE). This online method aims to map the perception problem into a lower dimensionality feature space, eliminating the costly computation algorithms for manipulating high dimensional raw sensor outputs. This in turn enables real-time feedback acquisition for high-level control in applications such as force control.

The method is experimentally validated on a bioinspired optical tactile sensor, TacTip, developed by Bristol Robotics Laboratory (BRL) [4], and comparison to existing methods is given where previous work exists. The method also allowed the calibration of the TacTip for force measurement applications (Fig. 1).

In our work, the unsupervised autoencoder training is conducted once, on a single classification dataset. The resulting encoder model is then applied in different perception problems (both classification and regression), thus demonstrating its versatility and generalizing ability. The proposed method will allow us to build and develop complex shapes of sensor membranes for specific tasks rapidly.

This letter is organized as follows: in Section II we give a state-of-the-art overview. Section III presents the feature extraction method, developed and tested on the chosen sensing technology. The validation results using different benchmarking algorithms are presented in Section IV, and robustness analysis in Section V. Conclusion and future work are given at the end of the letter.

II. Related Work

A trend of research approaches treating tactile information as images can be observed in both optical based sensors and tactile sensing arrays [5]. As one of the most common approaches in image processing, CNNs have in the domain of tactile sensing almost entirely been deployed only as perception models, even though widely used for feature extraction in image applications. CNN has been trained for GelSight sensor output images for contact force estimation [6], and in a more complex neural network model for object hardness estimation [7]. Recent results on object edge detection have been reported by the BRL group [8]. An approach similar to this has used a CNN-MLP autoencoder for feature extraction from an optical based sensor [9]. There however, the sensor outputs close-up photographs of an object in question, without other tactile cues that could provide generalised information on the mechanics of the contact. Additionally, the method is only manually validated on a single use-case, without providing robustness with respect to variations in hardware or perception problem.

From the classical dimensionality reduction methods, principal component analysis (PCA), though deprived of physical meaning of the data, has often been used in different tactile sensing technologies as a preprocessing tool for other inference algorithms [10]–[12]. In [13] a decomposition of a covariance matrix of a tactile sensor array output was used to reduce the feature set dimension. Out of the other common feature extraction methods, we mention unsupervised approaches of [14] where a sparse coding of tactile features was used, and a denoising autoencoder applied to distributed tactile sensors in [15].

The domain of tactile perception, which deals with providing useful, task-oriented information for planning and control, builds upon different sensing technologies and raw data acquisition. The perception problem was tackled by the BRL group using machine learning approaches (Bayesian inference in [16]–[20], SVM in [21]) with localization, manipulation, and slip detection as control objectives, and recently, CNNs in [8] for object edge detection. Similar vision-based sensors were geometrically described, and analytic approximations were found for touch modality and contact region identification [22], distribution of force vectors and surface traction fields [23], and contact object surface normals [24]. No generalised methods providing versatility in data manipulation have been developed.

III. Materials and Methods

A. Sensing Technology

A bioinspired optical tactile sensor was developed by Bristol Robotics Laboratory [4], and further developed over multiple iterations [3]. As shown in schematic on Fig. 2, the TacTip sensor consists of a hemispherical (black) silicone tip filled with a silicone gel, and a camera directed towards the interior of the tip. Inspired by the human skin, the inside of the hemisphere contains a pattern of white-tipped pins that conduct and amplify the movement of the outer tip membrane while in contact with the environment [16]. An internal light source (ring of LEDs) is directed towards the pins, in order for the camera to record the pins' movements.

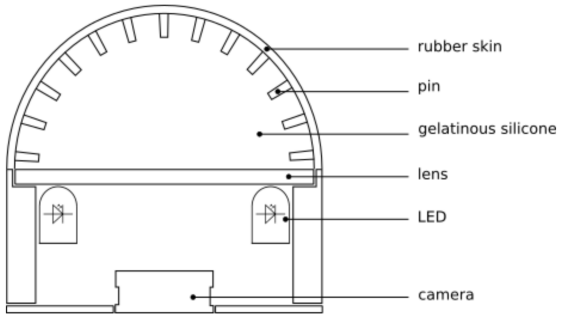


Fig. 2. Cross section of TacTip showing the hemispherical membrane with pins (papillae) on the inner side, a light source, and a camera recording pin movements.

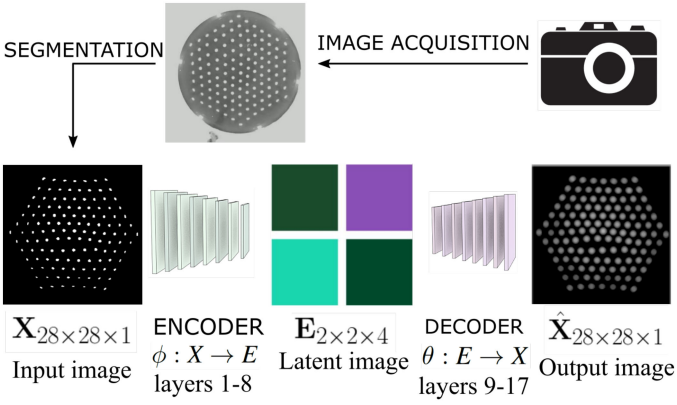


Fig. 3. Examples of raw (upper left), and preprocessed images (lower left), and corresponding encoder output representation (right). Middle image serves purely as the means to visualize the encoder and its purpose in the pipeline. It shows the encoder data as a 2 × 2 pixel ×4 channel CMYK image (2D representation of 16 values).

Due to the partial transparency of the black sensor tip, environmental cues such as external light sources around the sensor setup can impact the image characteristics. The raw images acquired by the sensor camera (Fig. 3 upper) are therefore initially preprocessed, in order to compensate for the inconsistency of lighting conditions across different experimental settings (Fig. 3 lower left), i.e. to filter out any information not related to pin positions. The result of the preprocessing is a binary image of the pins, resized from 300 × 300 pixels into 28 × 28 using linear interpolation.

Even though this single tactile sensing technology is tested here, the generality of the described approach allows for deployment on different sensor membrane shapes, or even other sensing technologies, as long as the tactile data can be represented in

2+D image-like format. For example, output from a capacitive sensor array could be stored in the form of an image, containing information on activation of neighbouring modules, and the feature extraction, as well as measurements, could be conducted as described here.

B. Feature Extraction - Convolutional Autoencoder

As in most tactile sensing technologies (e.g. robot skin), a feature extraction step is beneficial before applying any inference method to the sensor outputs. In [16]–[21], the feature extraction step consists of a Python OpenCV contour detection algorithm that enables pin tracking. The chosen features are therefore the 254 pin coordinates (x and y coordinates of the 127 pins).

Here, an **unsupervised** feature selection is conducted on sensor images reduced to 28×28 pixels. The manual choice of adequate representation variables is bypassed using the denoising autoencoder neural network architecture. Simply put, autoencoder network tries to approximate the identity function, so as to output $\mathbf{y} = \hat{\mathbf{x}}$ that is similar to input $\mathbf{x}$. The traditional autoencoder neural network architecture consists of an encoder and a decoder part, connected at the narrowest point of the network. The two can be defined as mapping:

$$\left.\begin{array}{l} \phi : X \to E \\ \theta : E \to X \end{array}\right\} \ni \underset{\phi,\theta}{\operatorname{argmin}} \|X - (\theta \circ \phi)X\| \qquad (1)$$

The *encoder* part (ϕ) of the network reduces the number of features from one hidden layer to the next, optimally resulting in a set of (pseudo-)orthogonal features (E) with minimal information loss, so that the decoder part (θ) is able to reconstruct the input. The backpropagation algorithm adjusts the autoencoder neuron weights in the direction of the loss function negative gradient, minimising the difference between the input and reconstructed output image. Once the autoencoder is trained, only the encoder part is deployed in applications, providing a smaller set of features, while the decoder part is only used as a means to train the encoder part, and is disregarded after training.

An autoencoder architecture can be built using any type of neurons. As an alternative to computer vision techniques, convolutional kernels were used as the basic building blocks of the autoencoder neural network, due to their ability to efficiently handle 2D data. The choice of this format was inspired by the 2D format of the sensor output (camera image) for speed and simplicity, but with the primary reason being the ability to store and emphasize the local spatial relations. Using CNNs bypasses computer vision algorithms that can fail in blob detection under large deformations. Another advantage of using CNNs is the ability to preserve and utilise the spatial information inherently encoded in the image. When using linearised representations of 2D data, such as an array of $(x, y) \in \mathbb{R}^{n\times 2}$ positions of n pins, some spatial information can be lost depending on the pin ordering within the array.

C. Encoder Design

The CAE architecture (Table I) was chosen through a cross-validation of a set of architectures. Number of encoder features is determined by a combination of data input size, number of hidden layers, number of kernels, and kernel sizes. These values were hyper-optimized so that encoders with $\{8, 16, 32, 64\}$ features were trained, some in several different architectures (e.g. less hidden layers with more kernels per layer). The cross-validation procedure resulted in an autoencoder with $2 \times 2 \times 4$ data dimension in the narrowest layer, i.e. an encoder with 16 features. A visualisation of obtained features is shown in Fig. 3, where the 4 channels are regarded as CMYK channels of a 2×2 pixel image. The CAE training was conducted offline using the open source neural network library Keras [25].

TABLE I
AUTOENCODER ARCHITECTURE. LAYERS 1–8 REPRESENT ENCODER, ENCODING $28 \times 28 \times 1$ SPATIAL INFORMATION INTO $2 \times 2 \times 4$ INFORMATION MATRIX, FURTHER REGARDED AS A 16×1 VECTOR IN PERCEPTION ALGORITHMS. LAYERS 9–17, REPRESENTING DECODER, ARE NOT RELEVANT FOR THE DEPLOYMENT SCENARIO, BUT ARE USED IN TRAINING THE ENCODER PART DURING OFFLINE FEATURE EXTRACTION LEARNING

	layer	kernel size	width	activation	output size
1	Conv2D	3×3	16	relu	$28 \times 28 \times 16$
2	max-pool	2×2			$14 \times 14 \times 16$
3	Conv2D	3×3	8	relu	$14 \times 14 \times 8$
4	max-pool	2×2			$7 \times 7 \times 8$
5	Conv2D	3×3	8	relu	$7 \times 7 \times 8$
6	max-pool	2×2			$4 \times 4 \times 8$
7	Conv2D	3×3	4	relu	$4 \times 4 \times 4$
8	max-pool	2×2			$2 \times 2 \times 4$
9	Conv2D	3×3	4	relu	$2 \times 2 \times 4$
10	up-sample	2×2			$4 \times 4 \times 4$
11	Conv2D	3×3	8	relu	$4 \times 4 \times 8$
12	up-sample	2×2			$8 \times 8 \times 8$
13	Conv2D	3×3	8	relu	$8 \times 8 \times 8$
14	up-sample	2×2			$16 \times 16 \times 8$
15	Conv2D	3×3	16	relu	$14 \times 14 \times 16$
16	up-sample	2×2			$28 \times 28 \times 16$
17	Conv2D	3×3	1	sigmoid	$28 \times 28 \times 1$

The dataset used for autoencoder training is described in IV. From the *shape-z* dataset, 720 images were used for training, and 480 for cross-validation. The network was trained on batches of 256 input images, for a maximum of 10000 epochs, with *Adadelta* as an optimizer, and binary cross-entropy as a loss function. Cross-entropy H for a binary variable with true value y_t and predicted estimate y_p is defined with eq. (2). In our case, the target variable is a black-and-white (binary) 2D input image Im, and the loss function is calculated as a mean of binary cross-entropy of all $X \times Y$ pixels p_t i.e. p_p, eq. (3). Early stopping method was used for overfit prevention, where error on the validation set was monitored and training was stopped after 300 epochs of non-decreasing validation set error. No other regularization methods were used in training. The training procedure reduced the loss function value from 0.7 to 0.05 on the training set, i.e. to 0.07 on the test set.

$$H(y_t, y_p) = -(y_t \log(y_p) + (1 - y_t)\log(1 - y_p)) \qquad (2)$$

$$\text{Loss}(\text{Im}) = \frac{1}{\|X\| \, \|Y\|} \sum_{x \in X} \sum_{y \in Y} H(p_t(x, y), p_p(x, y)) \qquad (3)$$

D. Method Validation

Due to the black box nature of both the data and the method, there is no formal proof of zero/minimal information loss, or the feature orthogonality. The autoencoder training is based on

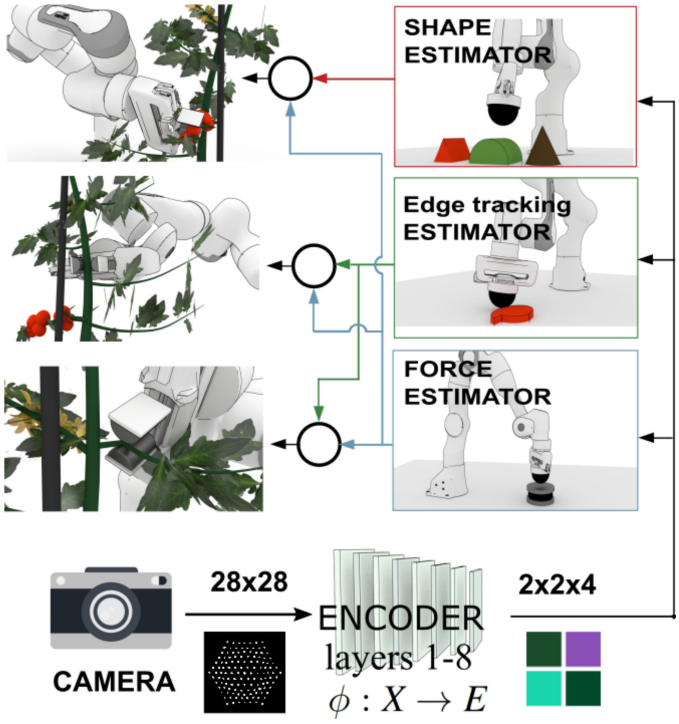


Fig. 4. The TacTip optical sensor captures tactile information in the form of a camera image. The image is transformed into a low-dimensional set of features using a convolutional encoder network. Machine learning models trained for different applications (e.g. shape, force control, edge detection, etc.) are used for tactile information extraction. These can later be combined to accomplish various plant treatment tasks (pruning, fruit picking, fruit testing. etc.)

a loss function minimization. This measure, however, is not necessarily the measure of interest when validating the obtained encoder. We have thus designed a set of experiments, in line with the intended encoder application, where the encoder was validated based on the usability of the features it provided.

Fig. 4 shows a typical deployment scenario of the TacTip sensor. Depending on the particular task, different perception models or *estimators* can be employed. The term *estimator* is used in order to encompass both classification and regression models, and both black-box and statistically-based models. Some examples of perception models are a contact force regressor, or object classifier.

Within this work, different estimators were trained and tested as benchmarks for CNN encoder validation: depending on the classification/regression accuracy of the developed estimators, we evaluate the encoder, i.e. the quality of information it extracts. It is worth noting that in case of a low accuracy of estimators, no conclusions should be made on the encoder functionality, since it is possible that the choice and/or design of estimator are not adequate, or simply, that there is not enough information in the raw sensor image for the desired information extraction.

A CNN encoder model was obtained through CAE training. Then, different contact datasets, represented with sets of $28 \times 28 \times 1$ matrices, were reduced into 16×1 feature vectors using the same pre-trained encoder network, and, depending on the dataset label set, different models were trained to map the input 16×1 features to their respective labels. As we later show in the experiments, the *estimators* managed to successfully predict the given labels on different datasets, using the same subset of features provided by the same encoder, and on a wide variety of labeling problems. We can conclude that the extracted features still carry sufficient information of the contact for the tested use-cases.

IV. Experimental Validation

The optimization procedure during autoencoder training minimizes error function (binary cross-entropy). The final value of the error function, though relevant for the training procedure, is not a clear indicator, nor a direct evaluation measure of the encoder performance. The quality of the extracted features was thus verified using estimators trained on a range of perception problems. The estimators' accuracies are then used as benchmark for validation of the presented feature extraction method.

The benchmark datasets were collected using two TacTip sensors. Since the sensor is not an off-the-shelf product, but rather an early version prototype that is partially manually assembled, differences in sensitivity and compliance parameters can be observed across instances of the same design. Additionally, different components can be used, such as cameras recording sensor interior, or the silicone gel. Despite the perception problems variety, and the structural differences in the sensors used in the study, the identical encoder was successfully deployed on both sensors, thus providing additional validation of the proposed approach.

A. Datasets

Shape-z: The first dataset, Shape-z, was collected for object shape classification task. Sensor images were recorded while in contact with four different shapes (cube, prism side, and 2 cylinders of different radii), at three different indentation depths, spaced at 3 mm, with overall span over 6 mm. For each shape and indentation depth 100 images were recorded, yielding 4 shape classes with 300 images per class. This dataset was also used for autoencoder training.

$\mathbf{r} - \theta$*:* Object edge position and orientation (see Fig. 5) were estimated based on two datasets collected with two distinct TacTip sensors. In the first, $(\mathbf{r} - \theta)_{class}$, the dataset is labeled at 4 displacement values and at 4 different angles. The edge displacement was varied in range $r \in [0, 8]$ mm from the center of the sensor coordinate frame, i.e. within $\sim$20% of the overall sensor diameter (it should be noted that, due to spherical shape of the sensor surface, this roughly corresponds to 50% of sensor surface that can be in contact with an object). The orientation was varied in $\theta \in [0, \pi/2]$. In this dataset, 800 images were collected (50 for each of 16 classes). Similarly, the dataset from [8], here denoted $(\mathbf{r} - \theta)_{regr}$, was used for a regression model learning.

F*:* The third type of dataset was used for contact force F estimator training. Experiments were conducted with two TacTip sensors, each being calibrated with a distinct commercial force/torque sensor. In the first dataset $F_{optoforce}$ a sequence of contacts of the sensor with a flat surface was recorded, where the contact was achieved both in normal direction, and under different angles (pseudo-randomly sampled between 45 and

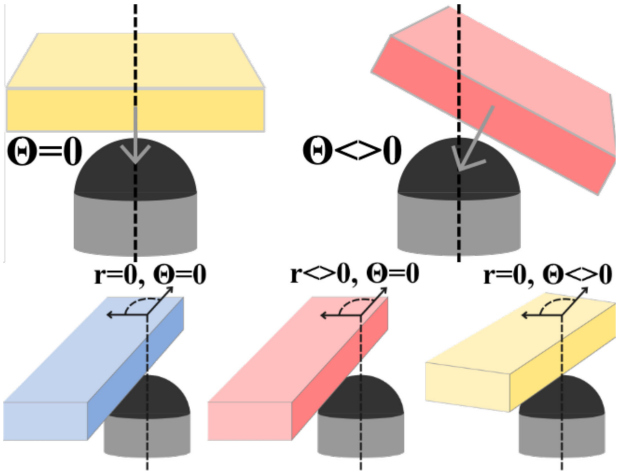


Fig. 5. Image shows Tac Tip sensor in contact with objects. It depicts the radial and angular displacements, r and θ respectively. Contact force normal to sensor surface and under an angle is shown in bottom images.

55 degrees to the surface normal). The Optoforce force/torque sensor was used for dataset target label generation, and forces with absolute values in the range of $[1, 6]$N were measured, varying indentation depths of the sensor surface. The dataset consists of $\sim$450 normal samples, and $\sim$530 samples where sensor normal was inclined w.r.t. the surface normal (see Fig. 5). The second dataset F_{ati} was collected using a second TacTip sensor, and ATI Mini40 for label generation, consisting of x, y, and z contact force components in the sensor local coordinates. Here, 3000 sensor images were recorded with variable sensor orientation and indentation w.r.t. the flat surface, resulting in varying contact force amplitudes.

B. Estimators for Encoder Evaluation

Within classification benchmarking, datasets $(Shape-\mathbf{z})$ and $(\mathbf{r}-\theta)_{class}$ were used. Within regression benchmarking, datasets $(\mathbf{r}-\theta)_{regr}$, $F_{optoforce}$ and F_{ati} were used. Both statistically and neural network based estimators were trained.

Naive Bayes classification: A naive Bayesian classifier was trained, for method comparison to results in [20]. During contact of the sensor and an object edge, a classifier is employed that estimates edge position r_i and orientation θ_j with respect to sensor coordinate frame, classifying them into one out of N_r and N_θ classes. The model is built using histogram method, binning each of the tracked $N_{dim} = 16$ variables (encoder provided image features) into one of the $N_{bins} = 10$ histogram intervals, for each recorded sensor contact. Joint prior $(r, \theta)_i, i \in \{1, \ldots, N_r \cdot N_\theta\}$ for radial displacement and edge orientation is calculated for each contact, from which r and θ estimates are inferred using $\arg\max$ operator. For implementation details, please refer to [20]. Additionally, two independent classifiers are trained for each $r_i, i \in \{1, \ldots, 4\}$ and $\theta_i, i \in \{1, \ldots, 4\}$, independent of the other. A similar inference model is trained for the dataset modelling different object shapes.

NN classification: Simple multilayer perceptrons were trained for classification tasks using Keras [25]. Four classification models were trained with similar architectures, each for a distinct classification dataset (as in Naive Bayes classification), consisting of an input layer, two hidden layers, and an output layer. The input layer was fed with $x \in R^{[16x1]}$ CNN encoder output. The first hidden layer consists of 16 neurons. The number of neurons in the second hidden layer (8/16) was varied depending on the label set size. As is common in multinomial classification, the activation function of the output layer is *softmax*, allowing us to use categorical crossentropy as loss function. The network was trained on batches of size 32, for 500, or 1000 epochs, depending on the label set size. *Adadelta* was used as optimizer.

TABLE II
ACCURACY OF CLASSIFICATION ALGORITHMS

dataset	variable	classNo	Bayes		ClassNN	
			train	test	train	test
$(shape-z)$	shape	4	0.93	0.90	0.97	0.97
$(\mathbf{r}-\theta)_{class}$	r, θ	16	0.99	0.86	1.00	0.90
$(\mathbf{r}-\theta)_{class}$	r	4	0.88	0.87	0.998	0.93
$(\mathbf{r}-\theta)_{class}$	θ	4	0.97	0.94	0.99	0.96

NN regression: Multilayer perceptrons were trained for regression tasks on $(\mathbf{r}-\theta)_{regr}$, $F_{optoforce}$ and F_{ati} datasets, again using Keras [25]. The regression models trained for different datasets all have $x \in R^{[16\times1]}$ CNN encoder output fed to the network input layer. In $F_{optoforce}$ regression network, there are 32 neurons in the first hidden layer, 16 in the second, and a dropout with rate 0.1 between them. In $(\mathbf{r}-\theta)_{regr}$ and F_{ati} regression networks, there are 128 neurons in the first hidden layer, and 64 in the second, and 32 in the third, with dropout layers in between. Finally, the output layer only consists of a single neuron (force amplitude $F_{optoforce}$), two neurons $((\mathbf{r}-\theta)_{regr})$, or three neurons ($F_{ati}$ for (x, y, z) force components), with *linear* activation function. The network weights were trained with *Adadelta* optimizer, with mean squared error as loss function. The networks were trained on batches of size 64, for a maximum of 20000 epochs. Early stopping was used as a regularization method, with patience of 1000 epochs.

C. Results

Table II gives classification algorithms accuracy on classification datasets. The statistically based Bayesian classifier achieved satisfactory results, roughly comparable to the results achieved in [20]. The results cannot be directly compared due to the differences in approaches. In [20] the decision procedure is stricter with evidence accumulation to threshold, whereas here, the prediction is in a single step. However, the results clearly show that as little as 16 black box features enable perception precision comparable to similar method using 254 features [20].

A fundamentally different approach, using a black box classifier based on a feed-forward neural network, yielded better results in the classification task, with an almost perfect classification score in one-step estimation. In both classifier types, the accuracy is higher on the datasets with a smaller number of target classes, which stems from non-bijective mapping of different combinations of stimuli (object properties) to the sensor output image. The lowest accuracy was achieved on the 16 classes

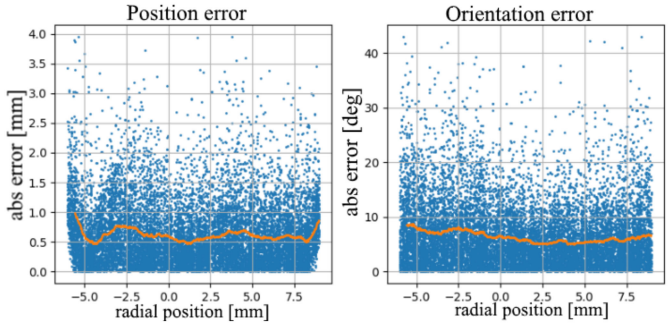


Fig. 6. Edge position and orientation prediction errors over position. The orange curve is the error smoothed with a 500-point window moving-average filter.

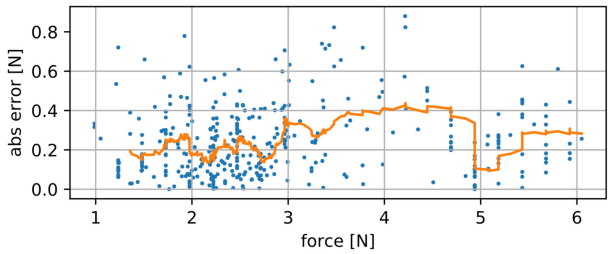


Fig. 7. Force estimation results on $F_{optoforce}$ test set, with MSE of 0.11 N (MSE on training set 0.03 N). Blue points show absolute prediction error, along with a corresponding moving-average filtered orange line (filter window of 20 samples).

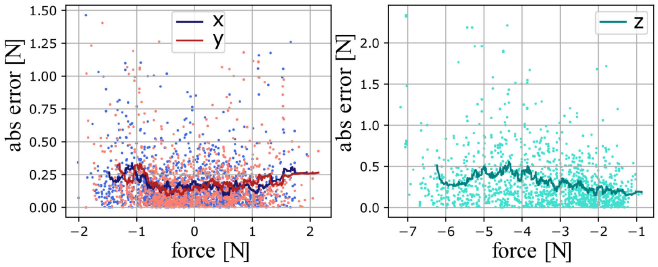


Fig. 8. Force estimation results on F_{ati} test set. Absolute prediction errors for all three components of force are shown w.r.t. target label, along with a 50-sample moving average of error (orange). x and y axes shown together since the results are symmetrical.

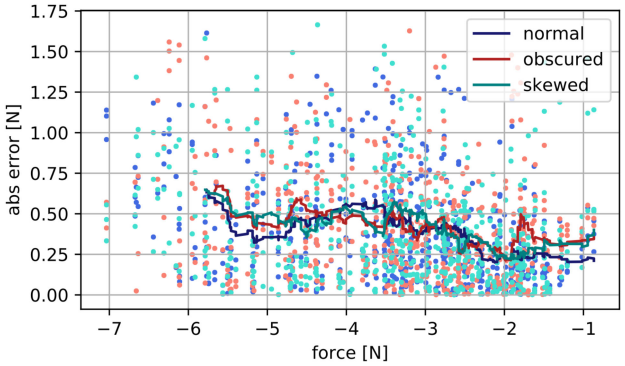


Fig. 9. The encoder generalizes over simulated family of sensors, enabling estimator training on the extracts features. Moving average of the prediction error is shown with 40-sample window.

dataset, but the results are satisfactory, taking into account that the CNN encoder robustly managed large translations and rotations of the object edge w.r.t. sensor coordinate frame, and extracted relevant features for an unseen set of samples.

Neural networks trained for regression tasks also achieved high accuracy. An estimator was trained on the dataset $(\mathbf{r}-\theta)_{regr}$ from [8] for methods comparison. There, a CNN was trained, mapping 28×28 sensor images to labeled values. Figure 6 shows (absolute) prediction errors (filtered using moving average with window 500) for both displacement and orientation, for the sake of comparison. We conclude that our method provides comparable accuracy, while allowing rapid development of perception models due to the simplification in the problem space, robustly handling both unseen features and hardware differences.

Another important result presented here is the contact force estimation. Normal force was estimated in [26] using genetic programming. In [27] an additional pressure transducer in the sensor body was used for analytical force estimation. Here, we show that force can be measured both in normal direction, and under acute angles, using only the features extracted from the sensor camera image. The prediction results on $F_{optoforce}$ are shown on test set in Fig. 7. The low resolution of the commercial sensor ($\sim$0.25 N) in this dataset cannot directly be inferred from the results, since raw sensor output was filtered prior to attributing contact force label to TacTip image sample.

The results on the second contact force dataset, F_{ati}, collected with a different instance of a TacTip sensor, and labeled with a higher resolution force/torque sensor ($\sim$0.01 N), are shown in Fig. 8. Even though the accuracy is slightly lower than in $F_{optoforce}$, the results again show that the encoder trained on one TacTip sensor instance can generalize in feature extraction, over both different perception tasks, and sensors with different mechanical properties. Additionally, we extend the capabilities of the TacTip by measuring the three spatial components of force.

V. Method Discussion and Analysis

In addition to the comparison to the state of the art results, several tests were conducted for method verification. Using the contact force dataset F_{ati}, two additional datasets were generated using image transformations in order to simulate consistent hardware differences. In the first dataset, one quarter of the image was masked, effectively obscuring $\sim$25% of the pins. This transformation simulates various sensor imperfections, for example impurities or camera malfunctions. The second dataset was generated using a skewing operation on the sensor images, simulating variability in sensor body compliance. Throughout the analysis, the datasets are named *normal* (original F_{ati} dataset), *obscured* and *skewed*, respectively.

In the first test, three feed-forward NN (FFNN) estimators were trained for the three contact force datasets. As the input to FFNN in this experiment we used encoder features. Fig. 9 shows the absolute prediction error on the z component of the contact force for all three datasets, along with the moving averages of the

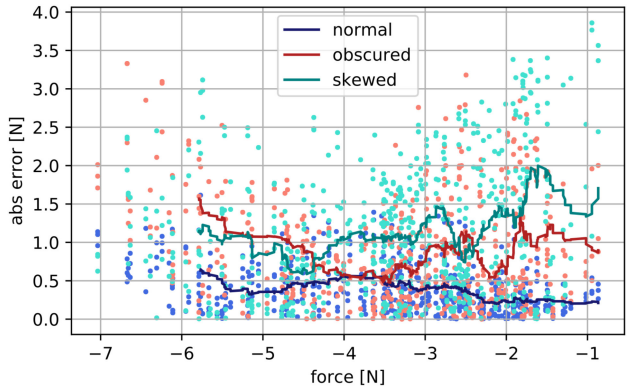


Fig. 10. Moving average of the prediction error is shown (40-sample window) with one encoder and a single estimator generalizing over simulated family of sensors.

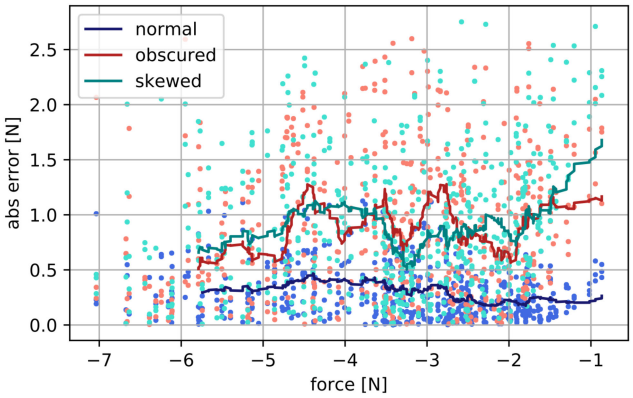


Fig. 11. CNN/FFNN estimator with 28 × 28 image as direct input bypassing encoder generalizes over simulated family of sensors.

errors. These results imply that the encoder generalizes well over differences in the hardware, allowing design and deployment of estimators needed for intended control scenarios.

In the second test, predictions on all three datasets were provided by a single FFNN trained on the *normal* dataset, with encoder features as inputs. Fig. 10 again shows the absolute prediction error on the z component of the contact force for all three datasets. Not unexpectedly, the performance is degraded, but some generalization of the estimator can still be observed, implying generalization of the encoder extracted features.

In the third test, instead of using the Encoder features as inputs, 28 × 28 pixel image is directly used as input. A CNN based estimator was trained on the *normal* dataset, and the prediction results on all three datasets are shown in Fig. 11. The network was designed as a single architecture consisting of a CNN input and FFNN output parts. The architecture of the CNN part is identical to the encoder network, and the FFNN part architecture is the same as that of the standalone trained estimators. As can be seen in Fig. 11, the network achieves slightly more accurate prediction with comparable generalization abilities.

However, this slight improvement comes at the training and deployment time cost. Comparison of three methods (CNN encoder based, CNN estimator, and FFNN for pin position) is given in Table III. The choice of the method will depend

TABLE III
TRAINING TIME AND ACCURACY (R2 SCORE) COMPARISON FOR DIFFERENT INFERENCE METHODS

Network type	Input	Accuracy [%]		Train time [s]
		train	test	
FFNN	Encoder features	95	86	90
CNN+FFNN	Camera image	98	93	1300
FFNN	Pin positions	99	99	3000

on the intended use case, since for rapid prototyping and deployment a readily available model might be more important than a near-perfect precision. In long-term deployment that requires precise measurements, putting time into building other single-task-oriented methods might be beneficial. It should be noted that Table III only compares R2 score accuracy and the training time of different methods, with the time calculated using Keras library on an Intel Core i5 CPU, strictly for comparison purposes. Other factors such as necessary training dataset size, and online computation requirements during deployments were not considered, but would further prefer the presented method in terms of time consumption.

VI. CONCLUSION

The results presented here show that the features extracted from sensor image output, even though drastically reduced in size, still carry sufficient information about the achieved contact to enable accurate touch data inference. The model generalisation was shown through successful application in various perception problems, as well as its robustness to hardware component variability. It has been shown that the model trained on a small classification dataset generalizes well enough to extract features relevant for both classification and regression problems. We have also shown that unrelated physical properties can be estimated from the same set of features, such as object edge orientation and contact force.

The advantages of this approach are two-fold. First, depending on future application scenarios, different estimators will have to be trained that would predict different contact properties of interest. In an alternative approach, i.e. the one used in state-of-the-art solutions, each new application would require training of a new complex perception algorithms, inherently containing feature extraction steps, resulting, as shown here, in unneccessary redundancy, all the while requiring a large training dataset for a complex network training. This method, due to the provided smaller set of denoised pseudo-orthogonal features, enables design of simpler estimators, which can thus be faster trained both thanks to their simpler architectures, and to the smaller dataset required for information inference.

Second, in an application scenario where different contact information is needed simultaneously, e.g. contact force and local object curvature measurements for robot control, using separate complex models for each type of measurement is costly both in terms of communication (transporting large image data), and computation (measurement estimation for each frame), when compared to using a single point dimensionality reduction, further disseminating lower dimensionality data to simpler (faster) estimation models.

The benefits of our proposed approach can be observed in applications such as simultaneous exploration and manipulation of objects, where different features extracted from the same tactile contact can be used on the one hand for object model generation (e.g. SLAM, using geometric object properties), and on the other, for compliant manipulator control (e.g. impedance control, relying on mechanical properties of the object).

As a part of the future work, the developed perception models will be deployed in exploration and manipulation tasks. The contact force measurements will be used for local object exploration, resulting in an object pointcloud map coding local curvature and stiffness within the occupancy grid. This in turn will allow a compliant (stiffness-aware) approach in manipulator trajectory planning. Another line of work will be directed towards manipulator control algorithm design that takes advantage of simultaneous availability of various tactile information.

Thanks to the unsupervised approach of the method, another line of future work is enabled, which will be directed towards building sensors with membranes of more complex shapes, designed for specific tasks. Using this common computer vision method, the tedious task of optimal feature extraction for each new sensor shape can be avoided.

REFERENCES

[1] M. Orsag *et al.*, "Specularia," 2019. [Online]. Available: http://specularia.fer.hr

[2] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robot. Auton. Syst.*, vol. 74, pp. 195–220, 2015.

[3] B. Ward-Cherrier *et al.*, "The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies," *Soft Robot.*, vol. 5, pp. 216–227, 2018.

[4] B. Winstone, G. Griffiths, C. Melhuish, T. Pipe, and J. Rossiter, "TACTIP—Tactile fingertip device, challenges in reduction of size to ready for robot hand integration," in *Proc. IEEE Int. Conf. Robot. Biomimetics.*, 2012, pp. 160–166.

[5] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[6] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017, Art. no. E2762.

[7] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 951–958.

[8] N. F. Lepora, A. Church, C. D. Kerckhove, R. Hadsell, and J. Lloyd, "From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 2101–2107, Apr. 2019.

[9] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of material properties from images," in *Proc. ICRA*, 2019, *arXiv:1803.03435*.

[10] G. Heidemann and M. Schopfer, "Dynamic tactile sensing for object identification," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2004, vol. 1, pp. 813–818.

[11] D. Goger, N. Gorges, and H. Worn, "Tactile sensing for an anthropomorphic robotic hand: Hardware and signal processing," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 895–901.

[12] K. Aquilina, D. A. Barton, and N. F. Lepora, "Principal components of touch," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 1–8.

[13] H. Liu, X. Song, T. Nanayakkara, L. D. Seneviratne, and K. Althoefer, "A computationally fast algorithm for local contact shape and pose classification using a tactile array sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 1410–1415.

[14] M. Madry, L. Bo, D. Kragic, and D. Fox, ''ST-HMP: Unsupervised spatio-temporal feature learning for tactile data," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 2262–2269.

[15] A. Schmitz, Y. Bansho, K. Noda, H. Iwata, T. Ogata, and S. Sugano, "Tactile object recognition using deep learning and dropout," in *Proc. 14th IEEE-RAS Int. Conf. Humanoid Robots*, 2014, pp. 1044–1050.

[16] N. Pestell, J. Lloyd, J. Rossiter, and N. Lepora, "Dual-modal tactile perception and exploration," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 1033–1040, Apr. 2018.

[17] N. F. Lepora, "Biomimetic active touch with fingertips and whiskers," *IEEE Trans. Haptics*, vol. 9, no. 2, pp. 170–183, Apr.–Jun. 2016.

[18] B. Ward-Cherrier, N. Rojas, and N. F. Lepora, "Model-free precise in-hand manipulation with a 3D-printed tactile gripper," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2056–2063, Oct. 2017.

[19] B. Ward-Cherrier, L. Cramphorn, and N. F. Lepora, "Tactile manipulation with a tacthumb integrated on the open-hand M2 gripper," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 169–175, Jan. 2016.

[20] N. F. Lepora, K. Aquilina, and L. Cramphorn, "Exploratory tactile servoing with active touch," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 1156–1163, Apr. 2017.

[21] J. W. James, N. Pestell, and N. F. Lepora, "Slip detection with a biomimetic tactile sensor," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3340–3346, Oct. 2018.

[22] Y. Ito, Y. Kim, and G. Obinata, "Contact region estimation based on a vision-based tactile sensor using a deformable touchpad," *Sensors*, vol. 14, no. 4, pp. 5805–5822, 2014.

[23] K. Sato, K. Kamiyama, N. Kawakami, and S. Tachi, "Finger-shaped gelforce: Sensor for measuring surface traction fields for robotic hand," *IEEE Trans. Haptics*, vol. 3, no. 1, pp. 37–47, Jan.–Mar. 2010.

[24] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson, "Microgeometry capture using an elastomeric sensor," *ACM Trans. Graph.*, vol. 30, 2011, Art. no. 46.

[25] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[26] G. de Boer, H. Wang, M. Ghajari, A. Alazmani, R. Hewson, and P. Culmer, "Force and topography reconstruction using GP and MOR for the TACTIP soft sensor system," in *Proc. Conf. Towards Auton. Robot. Syst.*, 2016, pp. 65–74.

[27] Y. Ito, Y. Kim, and G. Obinata, "Multi-axis force measurement based on vision-based fluid-type hemispherical tactile sensor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 4729–4734.

.4 PUBLICATION 4 - SOFT ROBOTICS APPROACH TO AUTONOMOUS PLASTERING

M.Polic, B.Maric, and M.Orsag Soft robotics approach to autonomous plastering. *IEEE International Conference on Automation Science and Engineering (CASE)*. Lyon, France, 482–487, 2021.

2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)
August 23-27, 2021. Lyon, France

Soft robotics approach to autonomous plastering

Marsela Polic, Bruno Maric and Matko Orsag

***Abstract*— This paper presents an industrial soft robotics application for the autonomous plastering of complex shaped surfaces, using a collaborative industrial manipulator. In the core of the proposed system is the deep learning based soft body modeling, i.e. deformation estimation of the flexible plastering knife tool. The estimation relies on visual feedback and a deep convolution neural network (CNN). The transfer learning approach and specially designed dataset generation procedures were developed in the learning phase. The estimated deformation of the plastering knife is then used to control the knife inclination with respect to the treated surface, as one of the essential control variables in the plastering procedure. The developed system is experimentally validated, including both the CNN based deformation estimation, as well as its performance in the knife inclination control.**

I. INTRODUCTION

The latest growth in the supply of affordable collaborative robots sparked a new surge in digital manufacturing. This time, the focus shifts from big (e.g. automotive) industries, towards small and medium enterprises, which usually work with smaller customer-oriented batches that require frequent changes to the production line. Collaborative robots are key enablers of these new technologies since they allow the unskilled operators to train them quickly, and effectively. However, standard industrial manipulators remain more versatile, cost effective and far easier to buy. This is why we proposed to deploy a standard joint-position controlled industrial manipulator KUKA KR10 in a collaborative way [1]. In the robotization of a specific manufacturing processes in the aerospace industry, we demonstrated a collaborative framework for robotic sanding, covering almost 90% of the sanding requirements across various production stages and products.

The goal of this paper is to extend the capabilities of this framework to another application, in particular, to apply a coat of putty using a standard knife tool 1. Based on the experience of human operators in different stages of the plastering process, the robotic plastering system should be able to control both the contact force and the angle between the knife tip and the treated object. By varying these two variables, the robotic system can perform different plastering strategies, such as applying a thicker layer, or removing excess plaster. However, the plastering knife is a flexible tool, undergoing physical deformation under application of force, which significantly complicates the problem of coupled force and angle control. Recent research results related to robotic plastering [2] show a mechanical setup that ensures the tool remains in a predefined contact with the wall. This system, deploying artificial neural network to process walls and plan the procedure, shows promising results, but is focused on large, flat wall surfaces.

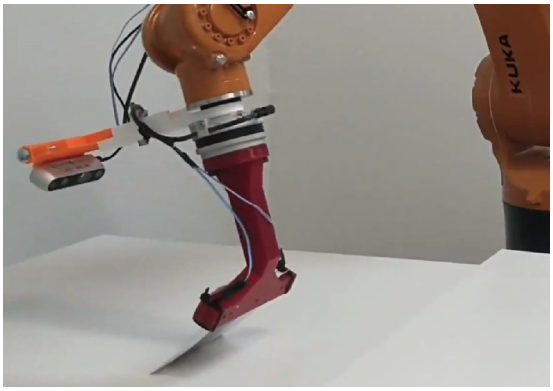


Fig. 1: Flexible plastering knife tool is mounted on the robot end effector, along with the Intel RealSense RGB-D camera and a torque sensor. Plastering is an example of an industrial task that involves deformable object manipulation.

The domain of deformable object manipulation is becoming ever more interesting with the development of sensing and computational capabilities. Robotic plastering, involving a flexible robotic tool, is an example of one such application, among many others in fabrics and clothing industry [3], flexible parts and cables in manufacturing [4], food and agriculture robotics, medical applications [5], and even robotic art [6]. As opposed to rigid body manipulation, the position control of deformable objects is not such a straight-forward task since their shape is affected by physical interactions. A comprehensive overview of deformation sensing, modeling and control in different application is given in [4]. Another compliant approach is to integrate deformation sensing into soft actuator hardware design [7], [8]. Until recent progress in deep learning, most of this existing work relied on analytical model fitting, based on visual and force feedback. In this work, we propose a deep learning based control feedback for manipulation of a deformable object in an industrial task. Similar to other work in the field, we choose a set of deformation features that enable the control to account for tool deformation. One classical approach models position and shape information

Authors are with the LARICS Laboratory for Robotics and Intelligent Control Systems, University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb 10000, Croatia (marsela.polic, bruno.maric, matko.orsag) at fer.hr

Authors would like to thank Frano Petric for fruitful discussion

2021 IEEE 17th International Conference on Automation Science and Engineering (CASE) | 978-1-6654-1873-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/CASE49439.2021.9551597

independently in the deformation features [5]. Here however, these are coupled, and represent the net effect that the position control and force application exert on the flexible robot tool.

Drawing inspiration from our previous work in the visual based tactile sensing, we obtain the control feedback estimates with a convolutional neural network (CNN) [9]. The camera is mounted on the robot arm, recording the controlled deformable object, i.e. the plastering tool. The tool is visually enriched with a set of visual cues, similar to several successful designs in the field of visual based tactile sensing [8], [10], [11]. Thanks to rapid advancements of GPUs, transfer learning has been widely exploited in recent years in classification tasks [12], across different applications, including industrial such as detection of defects in welding [13], [14] and surface defect detection [15], [16]. It has been deployed in robotic plastering as well, training an Inception V3 based architecture for grading of robotic plastering results [17]. Similarly, in [2] the robotic plastering success was evaluated through a custom trained CNN based classification on wall images.

One of the emerging architectures is the MobileNet model [18]. The novelty in this model, the depth-wise separable convolutions, reduced the complexity of computation, and enabled deployment in mobile and embedded platforms. In this work, this efficiency was key to deployment of the MobileNet V2 [19] in the closed force/position control loop of an industrial manipulator. To this end, as an addition to the state of the art, we have extended the transfer learning approach to training a regression model.

II. Shape modelling and control

During product finishing, workers apply protective or decorative coating using a flexible plastering knife tool. In this work Kuka KR 10 robot was equipped with one such off-the-shelf available knife tool in a setup shown in Fig. 1. The mount of the tool attaches to the flexible knife and an Intel RealSense RGB-D camera, mounted so that it captures the complete plastering knife within the camera frame. A torque sensor is deployed to measure contact forces with the treated surface. Inspired by the results from visual based tactile sensing development [9], the plastering tool was enriched with visual cues in the form of regular grid of dots for easier deformation modeling through image analysis.

To ensure complete surface coverage, high level control layers generate the desired tool trajectory and the desired force. These control variables, the applied contact force and the shape of the knife tool, are coupled due to the flexibility of the plastering knife. Namely, the shape of the tool is a result of the contact force acting between the surface and the knife. The shape of the knife is described with the relative position of the tip with respect to the robot flange, and the inclination angle of the knife tip with respect to the treated surface, i.e. Δz, Δx and $\Delta\varphi$ as shown in Fig. 2. The trajectory thus defines desired knife tip waypoints $\mathbf{T}_{d_B^\kappa}$ that describe the poses and inclination

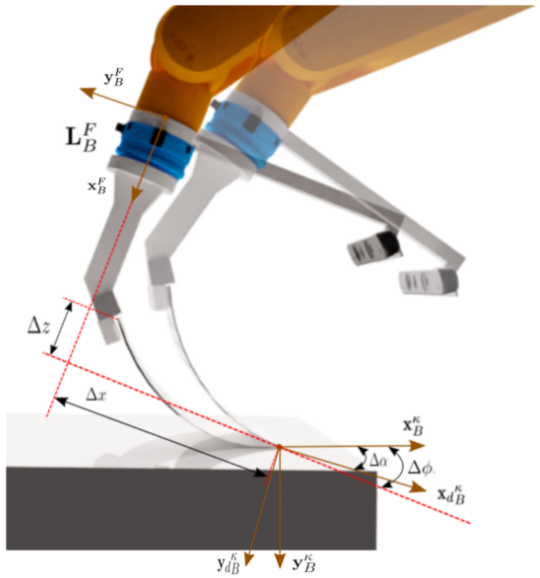


Fig. 2: Knife shape is estimated with deflection variables ΔX, ΔZ, $\Delta\varphi$. The estimated and desired approach vectors of the knife tip are $\mathbf{x}_B^\kappa$ and $\mathbf{x}_{d_B^\kappa}$, respectively. The controller compensates the angle error $\Delta\alpha$ by rotating the robot flange.

angles of the knife, and the desired force, applied to the surface along the local surface normal vector $\mathbf{n}_B$.

Our control system for robotic plastering is decoupled into two subsystems: the knife inclination control, and the impedance based FDCC controller for contact force control. In this work, we focus on the knife inclination angle control, based on robot flange rotation to provide the desired inclination of the knife regardless of the contact force. The FDCC control algorithm combines three main control concepts: Impedance, Admittance and Force Control, and guarantees robot independent compliant behaviour. The approach relies on dynamics simulation, and can thus achieve the same compliant behaviour on different manipulators. The discussion of contact force and trajectory planning remains for future work.

A. Knife inclination control

Planned trajectories rely on the known shape of the treated surface, and an optimal shape of the knife. To make sure that the desired knife tip pose is tracked, we propose the following control strategy. Starting from knife shape shown in Fig. 2, we consider the knife tip angle $\Delta\varphi$ w.r.t. the fixed part of the knife tool. The current and desired knife tangents are denoted as $\mathbf{x}_B^\kappa$, and $\mathbf{x}_{d_B^\kappa}$, respectively. The control strategy is to rotate the robot flange $\mathbf{L}_B^F$ to align $\mathbf{x}_B^\kappa$ with desired $\mathbf{x}_{d_B^\kappa}$, while keeping the knife tip $\mathbf{p}_B^\kappa$ at the same position. In other words, the robot flange is to be rotated around the estimated knife tip position by an angle $\Delta\alpha$, calculated using the small angle approximation of the vector product between the

desired and the actual orientation:

$$\Delta\alpha \sim \frac{\|\mathbf{x}_{d_B^\kappa} \times \mathbf{x}_B^\kappa\|}{\|\mathbf{x}_{d_B^\kappa}\| \, \|\mathbf{x}_B^\kappa\|}. \quad (1)$$

The calculated angle error is used to generate a rotation matrix $\Delta\mathbf{T}$ the axis $\mathbf{z}_B^\kappa$, $\Delta\mathbf{T} = \mathbf{Rot}(\alpha, \mathbf{z}_B^\kappa)$. The new flange pose is corrected using the following nonlinear equation:

$$\mathbf{T}_{B,new}^F(t_{k+1}) = \mathbf{T}_B^\kappa(t_k) \cdot \Delta\mathbf{T}(\alpha) \cdot {\mathbf{T}_F^\kappa}^{-1}, \quad (2)$$

where $\mathbf{T}_B^\kappa$ and $\mathbf{T}_F^\kappa$ denote the desired and actual transformation the knife w.r.t. Δx and Δz.

From the control perspective, angle error $\Delta\alpha$ is a nonlinear function. In this paper, we limit the discussion to the linearized point around the desired shape of the knife, for surface irregularities in a single degree of freedom. For this class of problems, $\Delta\alpha$ takes the following open loop behavior:

$$\Delta\alpha(s) = \mathbf{J}_B^\kappa(\mathbf{q}, \Delta z, \Delta x, \Delta\phi) g_\kappa(s) G_{FDCC}(s), \quad (3)$$

where $\mathbf{J}_B^\kappa(\mathbf{q}, \Delta z, \Delta x, \Delta\phi)$ denotes the Jacobian linearization of the robot pose including plastering knife shape, at joint angles $\mathbf{q}$. The dynamic elasticity of the knife is described with $g_\kappa(s)$, and $G_{FDCC}(s)$ describes the closed loop dynamic of the FDCC impedance control of the flange. The last part of the equation encompasses the dynamic behavior of the robot arm control function, tuned as a PT2 system for a desired behavior.

In effect, we propose controlling $\Delta\alpha$ with a steady set zero reference value. This ensures the knife tip follows the desired tangent approach vector $\mathbf{x}_{d_B}^\kappa$. The linearized approach proposed in the paper enables us to deploy a classic PI control, tuned based on system identification.

III. Knife shape estimate

The deformation of the plastering tool, essential for a successful task execution, is accounted for by the robot controller through a closed control loop. The three deformation features, describing the net effect of the relative knife pose and the exerted force, are obtained from a deep CNN based black-box model. The back-bone of the CNN model is the MobileNet V2 architecture, chosen for its efficiency and suitability for online inference in closed loop control. The history of CNN development is closely related to visual scene analysis, resulting in models trained for object detection and semantic segmentation. As such, it is extensively trained for classification on different benchmark datasets. In this work, we build upon the network pretrained on the ImageNet dataset [20].

Through dataset collection we show other methods capable of measuring the same deformation on point clouds. These can be measured using linear laser scanners or RGBD cameras, like the one used in this work. We show that the proposed method shows results comparable to point cloud methods, and argue that it can be applied to more complex shape deformations, where information cannot be modeled exactly, like in this case.

A. Transfer Learning

We use the transfer learning approach, in which we first remove the last, classification layer of the pretrained network. Then, we add four feed-forward (FF) layers, with 1024, 1024, 512, and 3 neurons instead, as determined in hyper-optimization, where cross-validation was conducted over a varying number and size of added layers. The deeper added layers have a rectified linear activation, while the output layer only has a linear activation function. This way, the three output neurons estimate the continuous values of the deformation features, allowing for modest generalization outside the training labels range.

A training procedure is conducted on a custom dataset, where the weights are adjusted only for the last five layers, while the rest of the network is kept as is. The number of trained layers is determined empirically. The dataset consisted of some 550 marked images, out of which approximately 400 are used for training, and approx. 150 in validation and test. One regularization technique deployed is early stopping, that prevents overfitting by monitoring the loss function on the validation dataset.

B. Dataset generation

The training dataset is generated in a series of experiments, where the tool is deformed under external force. The deformation is recorded with an Intel RealSense D435 RGB-D camera. The ground truth for the three deformation features is extracted from the point cloud recordings. First, the 3D point-cloud data of a single reading is transformed into the 2D reading, as a vertical slice of the data in a predefined patch. This transformation maps the tool shape deformation in 2D space, and produces a mean descriptor of tool slices. Tool shape information can be extracted either with visual cues in the RGB spectrum, or it could be identified at a discontinuity in the derivative along z axis in the local camera frame. With an a priori known length of the knife, the points belonging to the tool are filtered from the starting point next to the mount all the way to the tip. Finally, a third order polynomial curve is fitted to the extracted knife profile.

Two positional deformation features are obtained from the position of the knife tip, and the third describing orientation, is obtained as a derivative of the fitted polynomial at the tool tip with respect to the tool mount. This process is also depicted in Fig. 3.

C. Input preprocessing

Some initial preprocessing is conducted on the images obtained by the Intel RealSense camera, originally in 640×480 resolution, including cropping to the area of interest, brightness and contrast manipulation, and scaling to the MobileNet input size, which is 224×224 pixels in three channels. The result of the preprocessing can be seen in the left side of Fig.4. Another important preprocessing step is the label set normalization of the three predicted variables into the $[0, 1]$ range. This way

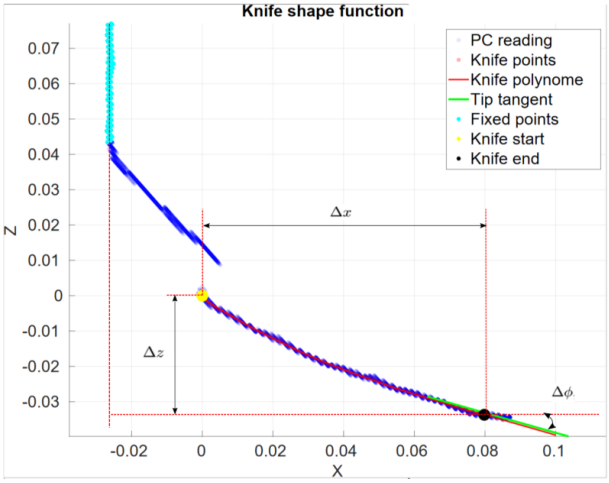


Fig. 3: Point-cloud readings of the central part of the tool, used for label generation. The flexible knife part is identified and fitted with a polynomial, shown with a red line. Polynomial tangent, shown in green, is considered knife tip inclination. Knife deflection is obtained as the angle between tip tangent and fixed tool part (cyan).

they contribute equally to the MSE. The normalization data on the training set is saved for de-normalization step during deployment.

D. Deployment

The efficient MobileNet V2 architecture enables a relatively high measurement rate on a NVidia GeForce GTX 1060 GPU in TensorFlow environment, at $\sim 30-50$Hz. Paired with a conventional Intel RealSense D435 camera, the measurement estimates are available at a rate of $\sim$ 30Hz. However, the outer control loop of the industrial manipulator in the described industrial setup is run at 100Hz. A white-noise linear velocity Kalman filter is deployed for measurement up-sampling through extrapolation. The measurement estimate error, manifested as signal noise, is sufficiently small for the successful execution of the plastering task. However, the inner control loop for robot motion run at 250Hz is sensitive to this error. In order to reduce the small amplitude shaking motion, the Kalman estimate is filtered with a simple low-pass PT1 filter.

E. Sensing results

The network trained via transfer learning is tested on a separate set of test data. The results are shown in Fig. 5 for all the three deformation features of the knife: two distance variables, and one angular. As shown in Fig. 2, the deformation is measured in the local coordinate frame of the end effector, hence the range of the variables does not start from zero. The mean average error accounts for less than 3% of the x and z axis variables range, at sub-millimeter mean absolute errors of $0.61mm$ and $0.93mm$ respectively. Similarly, the mean absolute angular error is at 1.88°, accounting for less than 4% of the variable range. The results are shown sorted by the true label of each variable for clarity. A larger prediction error towards the lower, left half of the variables ranges can be observed. These measurements correspond to the unexcited knife, clearly dominating the dataset in terms of number of measurements. This counter-intuitive behavior is the result of intentional dataset pruning in the training procedure, where the proportion of the unexcited state samples was reduced. Since almost half of the dataset corresponds to the neutral knife position, it was reduced in size to prevent the network from over-fitting to this portion of the data.

IV. Experimental results

The experimental setup includes joint-position controlled industrial manipulator KUKA KR10, equipped with 6-DOF force/torque sensor, Intel Realsense RGB-D camera and specially designed plastering tool. The robot is controlled though the ROS environment, using the impedance based FDCC [21] controller.

We performed system identification for the open loop transfer function 3. Second order approximation of the system dynamics yielded the following transfer function:

$$G_{\Delta\alpha} = \frac{1.63}{0.0043s^2 + 0.1036s + 1} \tag{4}$$

Knowing the linearized transfer function of the system Eq. 3, a PI controller was designed using the standard pole placement technique, resulting with closed loop behaviour with 7% overshoot and $455[ms]$ settling time.

Experiments are conducted with the plastering tool on experimental surfaces, evaluating the precision of knife inclination control during the plastering procedure. The robot is provided with a position trajectory for the knife tip, as well as the desired knife inclination with respect to the treated surface. The position references are effectively filtered through the FDCC controller, since the force control is turned off for the sake of knife tip pose control validation.

The control loop is closed through the neural network provided estimations of flexible tool position and inclination while in contact with the manipulated surface. The controlled variable was the knife inclination, identified as crucial in the plastering process. The other two estimated deformation features were not directly controlled, but were used in robot position reference adaptation, accounting for the tool deformation in the robot position control. The task motion was referenced along the x axis, with z axis reference only provided for realizing contact with the surface. A realistic reference of 45° for the knife inclination was commanded in the experiments.

The experiments were conducted on a piece-wise-flat surface, mimicking a step change unmodeled irregularity of the manipulated surface. The measurements of the experiments in the open loop in Fig. 6 clearly show this irregularity at $t \approx 12s$, corresponding to the more prominent knife deformation upon coming into contact

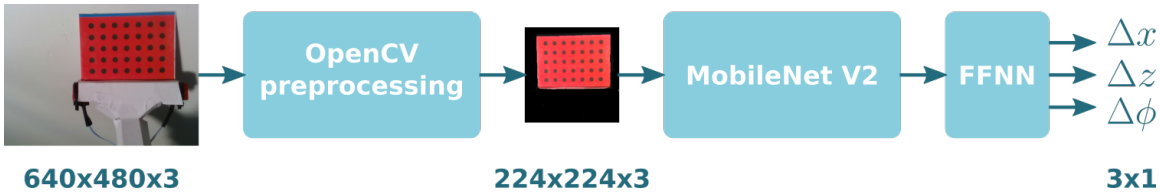


Fig. 4: Full sensing pipeline. The raw RGB camera image is preprocessed with HSV and brightness filters, and cropped into $224 \times 224 \times 3$ as an input into pretrained CNN. Four feed forward layers (FFNN) are trained to estimate the three knife deformation features.

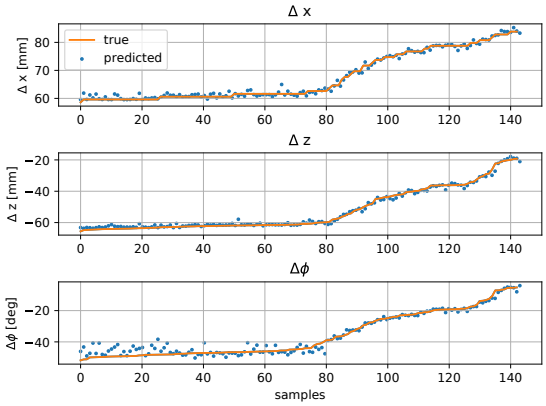


Fig. 5: The prediction results of the trained CNN on the test set. Higher precision for the excited states is due to manual pruning of the unexcited samples.

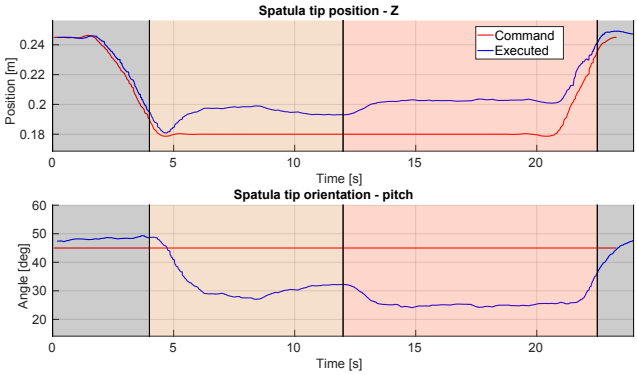


Fig. 6: Showing global position of knife tip in robot base frame during the open-loop motion.

with a protruding surface profile. This step disturbance can be observed in the lower right corner of the Fig. 1. Several repetitions of the experiment were conducted in the closed loop, all of which resulted in stable behavior of the system. Fig. 7 shows how a tuned PI controller reaches the inclination reference. Then, at $t \approx 13.5s$, the robot reached the surface irregularity. The measured inclination error was again accounted for by the robot position control. The results in Fig. 7 show one randomly chosen representative experiment, out of 5 successful consecutive repetitions with the same reference.

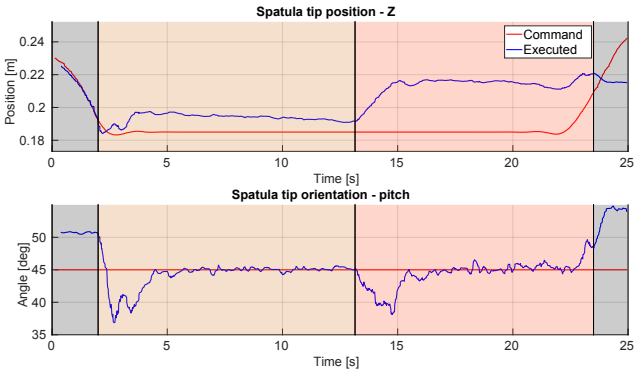


Fig. 7: Showing global position of knife tip in robot base frame during the close-loop motion.

Increased signal-to-noise ratio can be observed in the measured angle signal in the closed loop experiments. Since the system is in contact with the surface even in the case of open loop experiments, the noise should not be attributed to the neural network prediction imprecision, but to the fact that the measurement rate is lower than the robot control rate. Even though upsampled from 30Hz to 100Hz, the higher rate only provided scheduled measurements to the real time robot control. The signal filtering in the upsampling process is however not completely accurate, not taking into account the control signals. Since the amplitude of the measurement noise is low enough not to deteriorate the reference following, this issue was not addressed within this work, but it remains a part of the future work, since this measurement signal also results in a "shaky" behaviour of the robot during task execution. The videos of the experiments can be found at [22].

It is worth noting that the z axis deformation is not directly controlled, and the inclination error compensation results in a larger tracking error of this deformation feature. This would be resolved by closing the control loop for the knife tip position as well.

V. CONCLUSION

In this work, deep learning is deployed in an industrial task involving deformable object manipulation. A CNN is trained in an efficient manner, by exploiting the transfer learning paradigm. From the raw camera images, the neural network predicts the values of the deformation

features of the flexible robot tool. The experimental validation was conducted with a reduced scope compared to the intended industrial application. However, this reduced scope enables us to formally validate the proposed control system.

Even though deep learning has long ago found its' roles in various robotic applications, including industrial, these were mostly at a higher control level, e.g. as decision modules in the state machine. In this work, we show promising results for deployment of deep learning based inference within the lower level robot control, in the position control loop. To validate this, a reduced scope experimental setup was tested, where a single deformation feature of a flexible robot tool was controlled with the feedback signal provided by a neural network model. The experimental results show that the provided control architecture ensures reference tracking, even when submitted to external disturbances.

An important takeaway from the experiments is related to the achieved measurements rate. The presented robotic system is controlled at 100Hz in a real-time control loop. The measurements are therefore to be provided at this same rate. However, the state of the art commercial hardware, in a sensing pipeline involving a consumer camera and PC running CNN inference, only achieves a maximum 30 Hz sensing rate under perfect conditions. The light-weight MobileNet V2 is the state-of-the-art approach, specifically designed for fast and/or embedded inference. However, not even this architecture, paired with the available hardware, can reach the 100Hz inference. Further hardware development is expected to leverage this issue.

Finally, through the cascade control around the FDCC setup, simultaneous position and force control would be achieved in the final deployment scheme. A position controller would account for knife tip position error, and the force controller would be tuned to control the contact force during manipulation. Another part of the future work is to extend the approach to more complex surfaces.

ACKNOWLEDGMENT

This work has been supported by Croatian Science Foundation under the project Specularia UIP-2017-05-4042 [23]

References

[1] B. Maric, A. Mutka, and M. Orsag, "Collaborative human-robot framework for delicate sanding of complex shape surfaces," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2848–2855, 2020.

[2] X. Li and X. Jiang, "Development of a robot system for applying putty on plastered walls," in *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1417–1422, IEEE, 2018.

[3] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849, IEEE, 2018.

[4] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

[5] D. Navarro-Alarcon, H. M. Yip, Z. Wang, Y.-H. Liu, F. Zhong, T. Zhang, and P. Li, "Automatic 3-d manipulation of soft objects by robotic arms with an adaptive deformation model," *IEEE Transactions on Robotics*, vol. 32, no. 2, pp. 429–441, 2016.

[6] L. Scalera, S. Seriani, A. Gasparetto, and P. Gallina, "Watercolour robotic painting: a novel automatic system for artistic rendering," *Journal of Intelligent & Robotic Systems*, vol. 95, no. 3, pp. 871–886, 2019.

[7] P. H. Nguyen, S. Sridar, W. Zhang, and P. Polygerinos, "Design and control of a 3-chambered fiber reinforced soft actuator with off-the-shelf stretch sensors," *International Journal of Intelligent Robotics and Applications*, vol. 1, no. 3, pp. 342–351, 2017.

[8] C. Xiang, J. Guo, and J. Rossiter, "Soft-smart robotic end effectors with sensing, actuation, and gripping capabilities," *Smart Materials and Structures*, vol. 28, no. 5, p. 055034, 2019.

[9] M. Polic, I. Krajacic, N. Lepora, and M. Orsag, "Convolutional autoencoder for feature extraction in tactile sensing," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3671–3678, 2019.

[10] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[11] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[12] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

[13] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects," *IEEE Access*, vol. 8, pp. 119951–119960, 2020.

[14] H. Zhang, Z. Chen, C. Zhang, J. Xi, and X. Le, "Weld defect detection based on deep learning method," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pp. 1574–1579, IEEE, 2019.

[15] Y. Wang, L. Gao, Y. Gao, X. Li, and L. Gao, "Knowledge graph-guided convolutional neural network for surface defect recognition," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pp. 594–599, IEEE, 2020.

[16] J. Zhou, W. Zhao, L. Guo, X. Xu, and G. Xie, "Real time detection of surface defects with inception-based mobilenet-ssd detection network," in *International Conference on Brain Inspired Cognitive Systems*, pp. 510–519, Springer, 2019.

[17] J. Bard, A. Bidgoli, and W. W. Chi, "Image classification for robotic plastering with convolutional neural network," in *Robotic fabrication in architecture, art and design*, pp. 3–15, Springer, 2018.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[21] S. Scherzinger, A. Roennau, and R. Dillmann, "Forward dynamics compliance control (fdcc): A new approach to cartesian compliance for robotic manipulators," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4568–4575, Sep. 2017.

[22] Laboratory for robotics and intelligent control systems youtube channel. Accessed: March 2021.

[23] M. Orsag *et al.*, "Specularia." http://specularia.fer.hr, 2019. Accessed: 2019-09-08.

CURRICULUM VITAE

Marsela Polić received her BSc and MSc degree in electrical engineering and information technology from the University of Zagreb, Faculty of Electrical Engineering and Computing (UNIZG-FER) in 2014 and 2016, respectively. As an undergraduate student, she received two Dean's Awards Josip Lončar, as well as the Ministry of Science and Education scholarship (2011-2014), and the City of Zagreb scholarship (2014-2016) due to her academic success. During the master program, she was awarded Erasmus+ scholarship and finished the third semester at Chalmers University of Technology, Goteborg, Sweden, as an exchange student. Alongside her studies, she gained industrial working experience in student internships at Končar KET, and at Ericsson Nikola Tesla.

Upon graduating, she joined LARICS (Laboratory for Robotics and Intelligent Control Systems) as a research assistant at the Department of Control and Computer Engineering (ZARI) at FER, Zagreb, where she has worked on several international and domestic scientific projects. In the FP7 project ASSISIbf (Animal and robot Societies Self-organise and Integrate by Social Interaction – bees and fish) she participated in the development of a robotic system for interaction with young honeybees, where she investigated collective behaviour of honeybees through development of distributed graph search algorithms. She also participated in the MBZIRC2020 (Mohamed Bin Zayed International Robotics Challenge), where she lead the mobile manipulation team in an autonomous wall building challenge. Her PhD was closely tied to the Croatian Science Foundation funded SpECULARIA - Structured Ecological CULtivation with Autonomous Robots In Agriculture, within which she developed a robotic system for manipulation of plants in structured greenhouse cultivation. She participated in other research projects such as Croatian Science Foundation funded AgroSPARC and research collaboration with Special Hospital for Orthopaedic Surgery Akromion. During her PhD studies, she visited Bristol Robotics Laboratory as a recipient of the 2019 British Scholarship Trust research grant, where she stayed with the Tactile Robotics Group. In 2019, she participated in a research visit at the Aerial Robotics Lab at Imperial College London as a part of the AeRoTwin project.

Her main research interests are in machine and deep learning in 2D and 3D perception, soft robotics, and compliant manipulator control. She is an author or co-author of 5 papers published in peer-reviewed journals and 6 papers presented at international conferences. The full list of publications is given below.

FULL LIST OF PUBLICATIONS

JOURNAL PUBLICATIONS:

1. K.Griparić, T.Haus, D.Miklić, M.Polić, S.Bogdan. A robotic system for researching social integration in honeybees. *PloS one*, e0181977, 2017, IF: 3.0 (Q1).

2. M.Polić, I.Krajačić, N.Lepora, M.Orsag. Convolutional autoencoder for feature extraction in tactile sensing. *IEEE Robotics and Autonomous Letters*, 4(4):3671-3678, 2019, IF: 3.6 (Q1).

3. M.Polić, M.Car, F.Petric, M.Orsag. Compliant Plant Exploration for Agricultural Procedures With a Collaborative Robot. *IEEE Robotics and Autonomous Letters*, 6(2):2768-2774, 2021, IF: 3.74 (Q2).

4. I.Vatavuk, M.Polic, I.Hrabar, F.Petric, M.Orsag, S.Bogdan. Team LARICS at MBZIRC 2020: Autonomous Mobile Manipulation in a Wall Building Scenario. *Field Robotics*, 2021 (accepted), IF: n/a).

5. M.Polić, J.Tabak, M.Orsag. Pepper To Fall: A Perception Method For Sweet Pepper Robotic Harvesting *Intelligent Service Robotics*, Online, 2021, IF: 2.246 (Q3).

CONFERENCE PUBLICATIONS:

1. M.Polić, Z.Salem, K.Griparić, S.Bogdan, and T.Schmickl. Unsupervised optimization approach to in situ calibration of collaborative human-robot interaction tools. *IEEE Evolving and Adaptive Intelligent Systems (EAIS)*. Ljubljana, Slovenia, 1–8, 2017.

2. A.Ivanović, M.Polić, O.Salah, and M.Orsag, and S.Bogdan. Compliant net for AUV retrieval using a UAV. *IFAC-PapersOnLine, 51(29):431–437, 2018*.

3. B.Marić, M.Polić, T.Tabak, M.Orsag. Unsupervised optimization approach to in situ calibration of collaborative human-robot interaction tools. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. Karlsruhe, Germany, 255–262, 2020.

4. V.Lešić, H.Novak, M.Ratković, M.Zovko, D.Lemić, S.Skendžić, J.Tabak, M.Polić, and M.Orsag. Rapid Plant Development Modelling System for Predictive Agriculture Based on Artificial Intelligence. *IEEE International Conference on Telecommunications (ConTEL)*. Zagreb, Croatia, 173–180, 2021.

5. M.Polić, A.Ivanović, B.Marić, B.Arbanas, J.Tabak, and M.Orsag. Structured Ecological Cultivation with Autonomous Robots in Indoor Agriculture. *IEEE International Conference on Telecommunications (ConTEL)*. Zagreb, Croatia, 189–195, 2021.

6. M.Polić, B.Marić and M.Orsag. Soft robotics approach to autonomous plastering. *IEEE International Conference on Automation Science and Engineering (CASE)*. Lyon, France, 482–487, 2021.

ŽIVOTOPIS

Marsela Polić stekla je zvanje prvostupnice, odnosno magistre elektrotehnike i informacijske tehnologije Sveučilišta u Zagrebu, Fakulteta elektrotehnike i računarstva (UNIZG-FER) 2014. te 2016. godine. Kao studentica preddiplomskog studija nagrađena je s dvije Dekanove nagrade Josip Lončar, stipendijom Ministarstva znanosti i obrazovanja (2011.-2014.), te stipendijom Grada Zagreba za izvrsnost (2014.-2016.). Tijekom diplomskog studija nagrađena je stipendijom Erasmus+, te je treći semestar završila na Chalmers University of Technology, Goteborg, Švedska, kao studentica na razmjeni. Uz studij je stekla radno iskustvo u industriji kroz studentske prakse u tvrtkama Končar KET i Ericsson Nikola Tesla.

Po završetku studija zaposlila se u LARICS-u (Laboratorij za robotiku i inteligentne upravljačke sustave) kao asistent na Zavodu za automatiku i računalno inženjerstvo (ZARI) na FER-u u Zagrebu, gdje je sudjelovala na više međunarodnih i domaćih znanstvenih projekata. U FP7 projektu ASSISIbf (Animal and robot Societies Self-organise and Integrate by Social Interaction - bees and fish) sudjelovala je u razvoju robotskog sustava za interakciju s mladim pčelama, gdje je istraživala kolektivno ponašanje pčela kroz razvoj distribuiranih algoritama pretraživanja grafova. Također je sudjelovala na MBZIRC2020 (Mohamed Bin Zayed International Robotics Challenge), gdje je vodila tim za mobilnu manipulaciju u zadatku autonomne izgradnje zida. Njezin je doktorat bio usko povezan s projektom Hrvatske zaklade za znanost (HRZZ) SPECULARIA (Strukturiran ekološki uzgoj primjenom autonomnih robota u staklenicima), u sklopu kojeg je radila na razvoju robotskiog sustava za manipulaciju biljaka u strukturiranom uzgoju u staklenicima. Sudjelovala je i u drugim istraživačkim projektima poput HRZZ projekta AgroSPARC, i u istraživačkoj suradnjoj sa Specijalnom bolnicom za ortopediju i traumatologiju Akromion. Tijekom doktorskog studija 2019 je kao stipendistica zaklade British Scholarship Trusta posjetila Tactile Robotics Group u Bristol Robotics Laboratory. U 2019 je također sudjelovala u istraživačkom posjetu Aerial Robotics Laboratoriju na Imperial Collegeu u Londonu u sklopu projekta AeRoTwin. Njezini glavni istraživački interesi su strojno i duboko učenje u 2D i 3D percepciji, mekana robotika i podatno upravljanje robotskim manipulatorima. Autorica je ili koautorica 5 radova objavljenih u recenziranim časopisima i 6 radova prezentiranih na međunarodnim konferencijama.

COLOPHON

This document was typeset and inspired by the typographical look-and-feel `classicthesis` developed by André Miede, which was based on Robert Bringhurst's book on typography *The Elements of Typographic Style*, and by the `FERElemental` developed by Ivan Marković whose design was based on `FERBook` developed by Jadranko Matuško.