Semi-Autonomous Control of an Exoskeleton using Computer Vision

Bengtson, Stefan Hein

DOI (link to publication from Publisher):
10.54337/aau521483109

Publication date:
2022

Semi-Autonomous Control of an Exoskeleton using Computer Vision

PhD Dissertation

Stefan Hein Bengtson

Dissertation submitted November 17, 2022

Thesis submitted: November 17, 2022

PhD Supervisor: Professor Thomas B. Moeslund
Aalborg University

PhD Co-supervisor: Professor Thomas Bak
Aalborg University

PhD Committee: Associate Professor Claus Brøndgaard Madsen (chair)
Aalborg University Denmark

Professor Tal Oron-Gilad
Ben-Gurion University of the Negev, Israel

Professor Norbert Krüger
University of Southern Denmark, Denmark

PhD Series: Technical Faculty of IT and Design, Aalborg University

ISSN: xxxx-xxxx
ISBN: xxx-xx-xxxx-xxx-x

Published by:
Aalborg University Press
Kroghstræde 3
DK 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by X, 2022

Curriculum Vitae

Stefan Hein Bengtson

Stefan Hein Bengtson received his BSc degree in Networks and Distributed Systems in 2015 and his MSc degree in Vision, Graphics and Interactive Systems in 2017 - both from Aalborg University, Denmark. His Master thesis focused on automatic tracking of fish in 3D using a stereo-vision camera setup.

After finishing his MSc degree in 2017 Stefan joined the Visual Analysis and Perception Laboratory at Aalborg University as a research assistant where he worked with surveillance of traffic intersections using various sensors. In the start of 2018, he joined the newly started EXOTIC project at Aalborg University as a PhD student. During his PhD project, he collaborated closely with the Spinal Cord Injury Centre of Western Denmark which he visited on multiple occasions. Furthermore, Stefan visited the Robotics and Semantic Systems research group at Lund University in Sweden for a few months during his PhD project. During this external research stay he collaborated on applying computer vision for pose estimation of objects, with a focus on symmetric objects.

His main interest is computer vision and how it can be applied to assistive robotics to make them behave in a semi-autonomous manner for improving the human-robot interaction. Another area of interest is robotic vision and especially in terms of doing pose estimation of objects. During his time as a PhD student, he has carried out both teaching and supervision of undergraduate and graduate projects within the areas of robotics, image processing, and computer vision. Both at Aalborg University in Denmark and at Sino-Danish Center in China.

Curriculum Vitae

Abstract

This PhD thesis was carried out as part of the EXOTIC project, funded by Aalborg University from 2018 to 2021. The shared goal of this interdisciplinary project was to research the idea of an intelligently tongue-controlled upper limb exoskeleton for persons with tetraplegia. The main focus of the work presented in this thesis is the application of computer vision for intelligent control in a semi-autonomous manner to make it easier to control the exoskeleton.

A review of existing work on using computer vision for semi-autonomous control of assistive robotics manipulators revealed a tendency of having a clear-cut division of control between the human and the system. This clear division is easy to understand, easy to implement and often improves the objective performance of the system, such as completing predefined tasks faster. However, other studies indicate that such clear-cut schemes may be less satisfying to use, especially for persons with mobile impairments, as it can be experienced as a loss of control when the machine takes over completely.

A semi-autonomous control scheme with an adaptive level of autonomy was hence proposed such that the user will always have a sense of control. This scheme was evaluated against a manual control scheme and a semi-autonomous control scheme based on a more clear-cut division of control. These different control schemes were evaluated across two studies, with the latter one including solely persons with movement impairments in their arms. Both studies indicated a statistically significant improvement across multiple scenarios when using the adaptive scheme instead of the other two schemes. Especially in more complex tasks, where the hand of the exoskeleton needed to be both oriented and positioned in a certain way.

The computer vision applied in the two studies of the semi-autonomous control schemes relied on classical methods, such as detecting objects by color thresholding. This was a deliberate choice to ensure reliable detections of the objects in the studies as the main purpose was to test the semi-autonomous control and not the computer vision. The applied computer vision algorithms would hence fail to work outside the restricted environment of these two studies.

Research on computer vision in less restrictive environments was conducted as well as part of this thesis, namely pose estimation of objects from RGB images where the pose information would be useful for automating grasping of these objects for e.g. an exoskeleton. An existing state of the art approach for doing pose estimation was expanded to alleviate many of its shortcomings, resulting in an increased pose estimation performance, and a significant reduction in memory usage, while at the same time maintaining an inference speed suitable for real-time usage. A custom loss function was proposed as part of the solution which is able to inherently handle symmetric objects which can be an issue when dealing with pose estimation.

Finally, the above approach was expanded even further by using a single shared model for all objects instead of multiple object-specific models. This reduced memory consumption even further while also boosting the pose estimation performance by fine-tuning parts of this shared model.

Resumé

Denne Ph.d.-afhandling er en del af EXOTIC-projektet, der blev finansieret af Aalborg Universitet fra 2018 til 2021. Det fælles mål med dette tværfaglige projekt var at udforske ideen om et intelligent tunge-kontrolleret exoskelet til overkroppen for personer med tetraplegi. Hovedfokus i det arbejde, der præsenteres i denne afhandling, er anvendelsen af computer vision til intelligent semi-automatisk styring af et exoskelet for dermed at gøre det lettere at styre det.

Der blev foretaget en systematisk gennemgang af eksisterende arbejde om anvendelse af computer vision til semi-automatisk styring af robotter, der kan assistere personer med bevægelseshandicap i overkroppen. Denne gennemgang afslørede, at der er en tendens til at vælge at have en klar opdeling af kontrollen, hvor mennesket er ansvarligt for en del af opgaven, imens systemet er ansvarlig for en anden del. Dette valg er let at forstå, let at implementere og denne tilgang vil ofte forbedre systemets objektive ydeevne, f.eks. ved at nogle prædefinerede opgaver kan udføres hurtigere. Andre undersøgelser viser imidlertid, at sådanne klare opdelinger af kontrollen kan være mindre tilfredsstillende for brugerne. Især for personer med bevægelseshandicap, da det kan opleves som endnu et tab af kontrol, når maskinen overtager kontrollen fuldstændigt i nogle situationer.

Der foreslås derfor en semi-automatisk kontrolmetode med et adaptivt niveau af assistance, således at brugeren altid har en følelse af at være i kontrol af exoskelet- tet. Denne adaptive kontrolmetode blev evalueret op mod en manuel kontrolmetode og en semi-automatisk kontrolmetode uden et adaptivt niveau af assistance. Disse forskellige kontrolmetoder blev evalueret igennem to forsøg, hvor sidstnævnte forsøg udelukkende omfattede personer med bevægelseshandicap i armene. Begge undersøgelser viste en statistisk signifikant forbedring på tværs af flere scenarier, når kontrolmetoden med et adaptivt niveau af assistance blev anvendt i stedet for de to andre kontrolmetoder. Det blev især tydeligt i de mere komplekse opgaver, hvor hånden på exoskelettet både skulle både orienteres og placeres på en bestemt måde for at fuldføre opgaven.

De computer vision algoritmer, der blev anvendt i de to forsøg med de

forskellige kontrolmetoder, var baseret på klassiske teknikker, f.eks. detektering af objekter ved hjælp af farve. Dette var et bevidst valg for at sikre en pålidelig detektion af objekterne i forsøgene, da hovedformålet var at teste den semi-automatiske styring og ikke selve computer vision aspektet. De anvendte computer vision algoritmer vil derfor ikke kunne fungere uden for det kontrollerede miljø, der blev anvendt i forsøgene.

Som en del af ph.d.-studiet blev der også forsket i computer vision algoritmer, som vil kunne fungere i mindre kontrollerede miljøer. Nærmere bestemt algoritmer til estimering af objekters placering ud fra farvebilleder, hvilket er nyttige informationer ved automatisering af opgaver, f.eks. når et exoskelet skal gribe fat i et objekt. En af de førende algoritmer inden for området blev brugt som udgangspunkt for yderligere forbedringer, hvilket resulterede i en mere nøjagtig estimering af forskellige objektets placering samt en betydelig reduktion i forbruget af hukommelse. Desuden bibeholdt den forbedrede version af algoritmen et lavt tidsforbrug ved estimeringen, og den er derfor egnet til brug i realtid. Desuden blev der foreslået en specielt tilpasset løsning, som er i stand til at håndtere symmetriske objekter, hvilket normalt ellers kan være et problem.

Endeligt blev ovenstående fremgangsmåde udvidet yderligere, således at en enkelt model kunne trænes til at håndtere flere forskellige objekter i stedet for at man skulle træne en specifik model til hver enkel type af objekt. Dette reducerede hukommelses- forbruget yderligere og forbedrede samtidig modellens evne til at estimere placeringen af objekterne endnu mere nøjagtigt.

Contents

Thesis Details

Thesis Title:	Semi-Autonomous Control of an Exoskeleton using Computer Vision
PhD Student:	Stefan Hein Bengtson
Supervisors:	Prof. Thomas B. Moeslund, Aalborg University Prof. Thomas Bak, Aalborg University

The thesis consists of the following publications:

[A] Mikkel Thøgersen, Muhammad Ahsan Gull, Frederik Victor Kobbelgaard, Mostafa Mohammadi, **Stefan Hein Bengtson**, and Lotte N. S. Andreasen Struijk, "EXOTIC - A Discreet User-Based 5 DoF Upper-Limb Exoskeleton for Individuals with Tetraplegia". In: *Proceedings for the 2020 IEEE 3rd International Conference on Mechatronics, Robotics and Automation*, pp. 79–83, 2020.

[B] Mikkel Berg Thøgersen, Mostafa Mohammadi, Muhammad Ahsan Gull, **Stefan Hein Bengtson**, Frederik Victor Kobbelgaard, Bo Bentsen, Benjamin Yamin Ali Khan, Kåre Eg Severinsen, Shaoping Bai, Thomas Bak, Thomas Baltzer Moeslund, Anne Marie Kanstrup and Lotte N. S. Andreasen Struijk, "User Based Development and Test of the EXOTIC Exoskeleton: Empowering Individuals with Tetraplegia Using a Compact, Versatile, 5-DoF Upper Limb Exoskeleton Controlled through Intelligent Semi-Automated Shared Tongue Control". In: *Sensors*, vol. 22, no. 18, 6919, 2022.

[C] **Stefan Hein Bengtson**, Thomas Bak, Lotte N. S. Andreasen Struijk, and Thomas Baltzer Moeslund, "A Review of Computer Vision for Semi-Autonomous Control of Assistive Robotic Manipulators (ARMs)". In: *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731-745, 2019.

[D] **Stefan Hein Bengtson**, Mikkel Berg Thøgersen, Mostafa Mohammadi, Frederik Victor Kobbelgaard, Muhammad Ahsan Gull, Lotte N. S. Andreasen Struijk, Thomas Bak, and Thomas B. Moeslund, "Computer

Vision-Based Adaptive Semi-Autonomous Control of an Upper Limb Exoskeleton for Individuals with Tetraplegia". In: *Applied Sciences*, vol. 12, no. 9, 4374, 2022.

[E] **Stefan Hein Bengtson**, Hampus Åström, Thomas B. Moeslund, Elin A. Topp, and Volker Krueger, "Pose Estimation from RGB Images of Highly Symmetric Objects using a Novel Multi-Pose Loss and Differential Rendering". In: *Proceedings for the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4618–4624, 2021.

[F] **Stefan Hein Bengtson**, Hampus Åström, Thomas B. Moeslund, Elin A. Topp, and Volker Krueger, "A Shared Pose Regression Network for Pose Estimation of Objects from RGB Images". Presented at: *The 16th International Conference on Signal Image Technology and Internet Based Systems*, 2022.

Additional Papers co-authored during the PhD related to the PhD topic:

- Lotte N. S. Andreasen Struijk, Mostafa Mohammadi, Mikkel Thøgersen, **Stefan Hein Bengtson**, Frederik Victor Kobbelgaard, Muhammad Ahsan Gull, Anne Marie Kanstrup, Michael Gaihede, Helge Kasch, and Thomas B. Moeslund, "Tongue Control of Exoskeletons and Assistive Robotic Arms for Individuals with Tetraplegia". In: *Abstract Book from the 16th Congress of the Nordic Spinal Cord Society : NoSCoS2019*, pp. 51, 2019.

- Max Hildebrand, Frederik Bonde, Rasmus Vedel Nonboe Kobborg, Christian Andersen, Andreas Flem Norman, Mikkel Thøgersen, **Stefan Hein Bengtson**, Strahinja Dosen, and Lotte N. S. Andreasen Struijk, "Semi-Autonomous Tongue-Control of an Assistive Robotic ARM for Individuals with Quadriplegia". In: *Proceedings for the 2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, pp. 157-162, 2019.

- Mostafa Mohammadi, Hendrik Knoche, Mikkel Thøgersen, **Stefan Hein Bengtson**, Muhammad Ahsan Gull, Bo Bentsen, Michael Gaihede, Kåre Eg Severinsen, and Lotte N. S. Andreasen Struijk, "Eyes-Free Tongue Gesture and Tongue Joystick Control of a Five DOF Upper-Limb Exoskeleton for Severely Disabled Individuals". In: *Frontiers in Neuroscience*, vol. 15, pp. 739279, 2021.

- Muhammad Ahsan Gull, Mikkel Thøgersen, **Stefan Hein Bengtson**, Mostafa Mohammadi, Lotte N. S. Andreasen Struijk, Thomas B. Moeslund, Thomas Bak, and Shaoping Bai, "A 4-DOF Upper Limb Exoskeleton for Physical Assistance: Design, Modeling, Control and Performance Evaluation". In: *Applied Sciences*, vol. 11, no. 13, 5865, 2021.

Papers co-authored during the PhD without any connection to the PhD:

- Malte Pedersen, **Stefan Hein Bengtson**, Rikke Gade, Niels Madsen, and Thomas B. Moeslund, "Camera Calibration for Underwater 3D Reconstruction Based on Ray Tracing using Snell's Law". In: *Proceedings for the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1410-1417, 2018.
- Joakim Bruslund Haurum, Anastasija Karpova, Malte Pedersen, **Stefan Hein Bengtson**, and Thomas B. Moeslund, "3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset". In *Proceedings for the 2020 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1–11, 2020.
- Malte Pedersen, Joakim Bruslund Haurum, **Stefan Hein Bengtson**, and Thomas B. Moeslund, "3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset". In: *Proceedings for the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426-2436, 2020.

Preface

This PhD thesis was carried out under the EXOTIC project (EXOskeleton using the Tongue for Intelligent Control) funded internally by Aalborg University. EXOTIC was an interdisciplinary research project including both PhD students and senior members from the Faculty of Engineering and Science, the Technical Faculty of IT and Design, and the Faculty of Medicine at Aalborg University. Furthermore, the Spinal Cord Injury Centre of Western Denmark (VCR) was an external partner in the project. The focus of this PhD project was to demonstrate the feasibility of using computer vision for an intelligent control of the tongue-controlled exoskeleton in the EXOTIC project.

I would like to thank my two supervisors, Thomas and Thomas, for their constant support and for always having my back - no matter what. A big thanks to Lotte, project lead on the EXOTIC project, and the other members of the EXOTIC team, who made this PhD project possible. Especially Mikkel and Mostafa - I really enjoyed our time together, whether it was working on the exoskeleton or playing board games. I would also like to thank my friends from Lund University and namely Hampus for his patience and his eternal struggle to keep our code organized. Furthermore, a big thanks to my colleagues at Aalborg University, who have been a solid base throughout my entire PhD period. I have really valued the opportunity to discuss anything, ranging from the newest tech to the challenges of parenthood. A special thank you to Malte who has been by my side since we started out on our bachelor studies nearly a decade ago.

I would also like to thank my family for always encouraging and supporting my constant need for tinkering, even when it included disassembling and utterly destroying old VHS players at a very young age. A lot of what I have accomplished can also be attributed to them. Also, a great thank you to my in-laws, who have been extremely supportive. Last but not least, I am tremendously thankful to my girlfriend Eva. I am so grateful for everything we have together and especially our wonderful little daughter Frida. Our little family is my source of happiness - you two are my everything.

Stefan Hein Bengtson
Aalborg University, November 17, 2022

Part I

Overview of the Work

Chapter 1

Introduction

Throughout history, humans have constantly been developing technology to make life easier and to accomplish things that were otherwise impossible. Another inherent trait in humans is our social nature, we gather, we bond and we generally care for each other. One obvious use for our technological advances is hence for assistive purposes in order to aid our fellow humans in need of help, such as persons with disabilities.

The intersection of this is *assistive technologies (AT)*, where a product or technology is specifically designed for the purpose of assisting a person with a disability. AT can be used in many different aspects and can take many forms depending on the disability it is designed to alleviate: glasses and hearing aids for persons with sensory impairments, such as reduced vision or hearing. Wheelchairs, crutches and recently exoskeletons, see Figure 1.1, for persons with physical impairments while persons with intellectual impairments can be aided by technologies such as special learning aids.

The diversity of assistive technologies is hence quite wide and some of them are so common in our society to the point where one might not even consider it an assistive technology any longer, e.g., glasses or contact lenses. However, a lot of people would be seriously impaired in their everyday life without their glasses. The need for assistive technologies is already prevalent today as it is estimated that approximately one billion people worldwide could benefit from assistive technologies today, but only a tenth of those have access to it [10]. The need for AT is likely to increase even further due to an ageing world population where the risk of a disability increases when age increases [10].

Many people who require assistance in their everyday life depend on caregivers to help them but this is an unlikely long term solution as caregivers are a limited resource [7] [1]. During the Corona pandemic it became apparent what the lack of caregivers could result in, as multiple Danish associations

Fig. 1.1: The author drinking from a straw with the aid of an upper limb exoskeleton and glasses. Both are examples of an assistive technology but at different stages of maturity. Glasses have existed for centuries whereas exoskeletons have just recently become a possibility for assistive purposes.

for disabled persons reported a major lack of caregivers [3]. In one case, a person with amyotrophic lateral sclerosis (ALS) who had been able to live at home with ALS for the last 15 years with help from family and caregivers had to be administered to the intensive care unit at a hospital due to lack of caregivers [3]. However, the impact of too few caregivers spans wide as it may also affect the partner and the close relatives of the person in need of assistance as they may resort to stepping in as primary caregivers. This may be an acceptable temporary solution but should be avoided in the long term as the relatives will most likely become overloaded [9].

The use of assistive technology can alleviate many of the issues outlined above by making persons with disabilities more independent and thereby reducing the need for caregivers or other assistance. Robotics is a promising technology in this aspect, and especially for persons with physical impairments, as advances in recent decades have removed multiple barriers to assistive robotics, such as price, size and power consumption. Wheelchair mounted robotic manipulators are a reality and are already commercially available, and are estimated to reduce the need for assistance by between 30% [8] and 40% [6] for persons in wheelchairs with upper-limb impairments. The usage of exoskeletons for both rehabilitation and assistive purposes has also gained a lot of attention recently [4]. In particular lower limb exoskele-

tons, where the legs are actuated, have advanced to a point where they are commercially available and starting to see use for rehabilitation purposes at hospitals and clinics. However, exoskeletons for assistive purposes in the home of the user are more uncommon, likely due to the cost and size of some of them. Especially upper limb exoskeletons, where the arms are actuated, are still an active area of research with only a few commercially available options [4].

This PhD project is part of the EXOTIC project, which is an acronym for **EXO**skeleton using the **T**ongue for **I**ntelligent **C**ontrol. The purpose of the project was to create an upper limb exoskeleton which can be used by individuals with complete paralysis of both arms and legs (tetraplegia). Controlling the exoskeleton despite paralysis of all extremities is made possible using a tongue-based interface. The system is further enhanced by an intelligent control scheme, based on computer vision, to help the user in controlling the exoskeleton. The focus of this PhD thesis is on the latter; the vision-based intelligent control of the exoskeleton.

The main concept is the following: by equipping the exoskeleton with a camera it is able to perceive its immediate environment. Computer vision is then used to interpret the camera data to assist the user in controlling the exoskeleton. It is hence similar to modern vehicles that feature multiple sensors in order to assist the driver of the vehicle.

How such a vision-based semi-autonomous control should work is still an on-going topic of research [2]. Furthermore, the context of a tongue-controlled upper limb exoskeleton is still a relatively unexplored area in human-robot interaction (HRI) research. Especially when the intended user of the system is paralyzed in both arms and legs, as this group can be more reluctant to accept a semi-autonomous control scheme because they might re-experience a loss of control [5].

Another aspect is the computer vision algorithms which the semi-autonomous control relies on to be able to automate parts of the control. In the recent decade, deep learning has drastically accelerated as to what is possible in terms of computer vision. However, deep learning has also introduced some new problems, such as training the models, which can be an issue as huge amounts of labelled data are often required. Furthermore, deep learning may in many cases eliminate the need for classic approaches, such as hand-crafted feature descriptors, but it requires careful consideration of both the model design and training parameters such as the loss function.

1 Thesis Structure

This PhD thesis is divided into three main areas of work: the EXOTIC project, human-robot interaction and computer vision for object manipulation. The publications related to this PhD thesis and their relation to the different areas are shown in Figure 1.2, with some of the publications overlapping multiple areas of work. The papers related to the EXOTIC project consists of multiple papers co-authored with the other members of the EXOTIC project. The papers focusing on human-robot interaction overlap both the computer vision and EXOTIC project, as the context for these papers is based on the EXOTIC project while employing computer vision as part of the HRI. The computer vision area contains publications focusing on pose estimations of objects which is useful for automating object manipulation tasks required by EXOTIC project. Finally, a review paper on computer vision for assistive robotics bridges the gap between all three main areas of work.

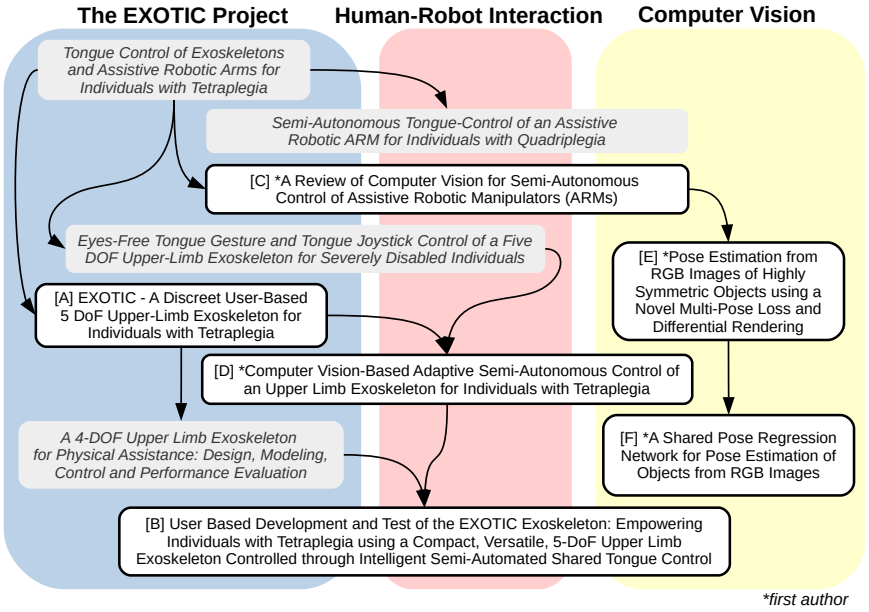


Fig. 1.2: An overview of the publications related to this PhD thesis. The publications are placed based on their association with the three main pillars of this PhD thesis: computer vision for object manipulation, human-robot interaction and the EXOTIC project. The publications included in this thesis are highlighted in white boxes with a solid outline.

The structure of this PhD thesis follows the same structure with the three main areas of work as outlined in Figure 1.2. First the EXOTIC project is described in more detail, as it provides the entire context for this PhD project, based on papers A and B. This includes the anticipated users of the system,

the upper limb exoskeleton and the tongue-based interface. This is followed by a description of the work carried out on human-robot interaction in the context of the EXOTIC project, based on papers C and D. Finally, the work on using computer vision for pose estimation of objects are described on the basis of papers E and F.

References

[1] S. Bedaf, P. Marti, F. Amirabdollahian, and L. de Witte, "A multi-perspective evaluation of a service robot for seniors: the voice of different stakeholders," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 6, pp. 592–599, 2018.

[2] S. H. Bengtson, T. Bak, L. N. S. A. Struijk, and T. B. Moeslund, "A review of computer vision for semi-autonomous control of assistive robotic manipulators (arms)," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731–745, 2020.

[3] E. Færch, "Handicappet kvinde må lade sig indlægge på intensiv hjælpere vil hellere være podere," *TV2 - Nyheder*.

[4] M. A. Gull, S. Bai, and T. Bak, "A review on design of upper limb exoskeletons," *Robotics*, vol. 9, no. 1, p. 16, Mar. 2020.

[5] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[6] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, "Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics*, June 2011, pp. 1–5.

[7] K. M. Marasinghe, "Assistive technologies in reducing caregiver burden among informal caregivers of older adults: a systematic review," *Disability and Rehabilitation: Assistive Technology*, vol. 11, no. 5, pp. 353–360, Sep. 2015.

[8] G. Romer, H. Stuyt, and A. Peters, "Cost-savings and economic benefits due to the assistive robotic manipulator (arm)," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 201–204.

[9] E. W. M. Scholten, A. Kieftenbelt, C. F. Hillebregt, S. de Groot, M. Ketelaar, J. M. A. Visser-Meily, and M. W. M. Post, "Provided support, caregiver burden and well-being in partners of persons with spinal cord injury 5 years after discharge from first inpatient rehabilitation," *Spinal Cord*, vol. 56, no. 5, pp. 436–446, Jan. 2018.

[10] W. H. O. (WHO) and T. W. Bank, *World report on disability 2011*. World Health Organization, 2011.

References

Chapter 2

The EXOTIC Project

The entire context for this PhD thesis was given by the EXOTIC project, which also provided the funding. A more in-depth description of the EXOTIC project along with the research carried out in relation to this is therefore described in the following. Note that the semi-autonomous control along with the computer vision aspects of the EXOTIC project are only described briefly in this chapter as they will be covered more in-depth later in chapter 3 and 4. Furthermore, a brief summary and a short video about the EXOTIC project is available at the official homepage[1].

1 Background - Tetraplegia

The main motivation of the EXOTIC project was to increase the quality of life and independence of persons with tetraplegia, who have limited to no motor function in their lower and/or upper body. A common cause of tetraplegia is injuries to the cervical spinal cord, i.e. the upper part of the spinal cord around the neck area. The degree of paralysis and which body parts are affected depend on the location and the severity of the damage, as illustrated in Figure 2.1. Damage to the higher vertebrates (C1-C6) usually causes tetraplegia, i.e. paralysis of both arms and legs. Injuries to the lower part of the spinal cord usually result in paraplegia as it mainly affects the legs. However, real-life examples of spinal cord injuries (SCI) are usually more complicated in terms of the body parts affected as the nerves within the spinal cord may only be partially destroyed in some cases. Such cases with partial paralysis of all four limbs are commonly referred to as incomplete tetraplegia.

On a worldwide basis it is estimated that between 250,000 and 500,000 persons injure their spinal cord, each year [5]. This is roughly equal to 40

[1] `https://rerob.aau.dk/#proj_Exotic`

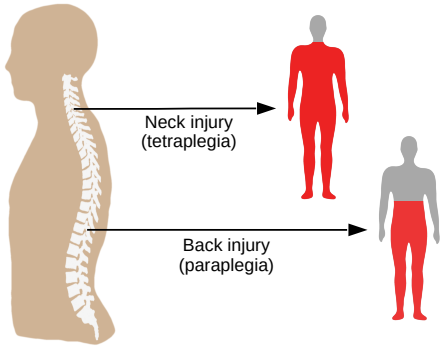


Fig. 2.1: The different levels of paralysis usually occur when damaging different parts of the spinal cord. The illustration assumes a complete injury where the nerves are completely severed at the point of injury. Real-life cases of SCI are usually more complex due to partial injuries where not all nerves are damaged.

to 80 new cases of SCI every year per million per capita on average. Furthermore, it is estimated that approximately one-third of these persons will have to live with some degree of tetraplegia due to their SCI [36]. The average age of persons sustaining an SCI is reported to be 33 years [36], but especially younger adults (females between 15-19 years and males between 20-29 years) and older persons (females older than 60 years and males older than 70 years) are the main contributors to this statistic. These numbers are primarily dominated by males as close to four males sustain an SCI for each female on average.

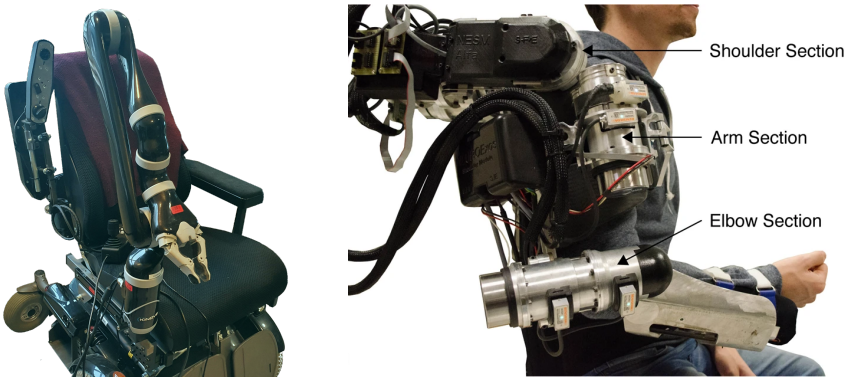
Studies show that the average life expectancy of a person sustaining a severe SCI at 25 years old is reduced by roughly 30% compared to the general population [5] [24]. The life expectancy after a severe SCI leading to tetraplegia is hence relatively high. In the case of a 25 year old person, it equates to roughly another 30 years when considering the average life expectancy for most developed countries.

The relatively high life expectancy is hence one of the main motivating factors for focusing on persons with tetraplegia. Another main motivation factor is the degree of assistance needed, especially for persons with severe tetraplegia with complete paralysis of all four extremities. These persons are dependent on constant assistance from caregivers around the clock, each and every day. The potential for assistive technology is hence extremely high for persons with tetraplegia due to these two factors. Both from a societal-level, in terms of the monetary cost of providing a high level of assistance for multiple decades. But also, on a personal-level, a person with tetraplegia can live a long and full-filling life if given enough support and assistance, for instance through the usage of assistive technologies.

2 Related Work - Assistive Technologies

Looking at assistive technologies for persons with tetraplegia, one of the most common and important technologies is their powered wheelchair [5]. Often operated with a joystick, either using the hands in case of incomplete tetraplegia or using the chin in case of complete tetraplegia, a powered wheelchair makes it possible to regain a lot of the mobility otherwise lost due to a limited function of the lower limbs, i.e. the legs. Powered wheelchairs are also an example of an assistive technology which has reached a point of maturity where it is widely used, at least in developed countries. Unfortunately, assistive devices focusing on the upper limbs, i.e. the arms, have yet to reach the same level of maturity. Persons with limited or no function in their arms and/or hands are hence still dependent on a caregiver to assist them with tasks such as eating.

However, in recent years several stand-alone assistive robotic manipulators (ARMs) have reached the commercial market, like iARM from Exact Dynamics [9] or JACO from Kinova [19], as shown in Figure 2.2a. Such ARMs could replace at least parts of the lost functionality in the arms of a person with tetraplegia and reduce the need for a caregiver. Previous studies on persons with upper limb impairments estimate a reduction in the need for assistance of 30-40% when using a wheelchair-mounted ARM [30] [23]



(a) JACO from Kinova [19] mounted on a wheelchair.

(b) The NESM exoskeleton [8]. Figure adapted from [8].

Fig. 2.2: Example of different assistive robotic manipulators.

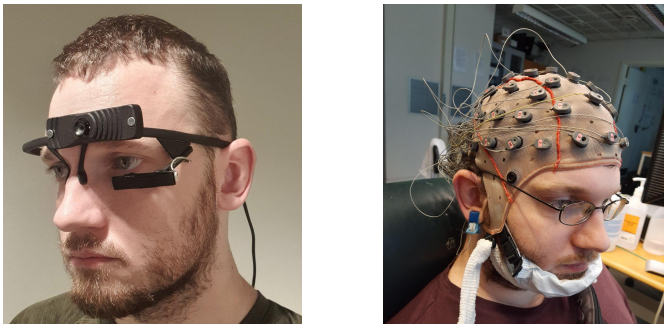
Another promising aspect is recent advances in the development of assistive upper limb exoskeletons [13] [8] [11], as shown in Figure 2.2b. A discerning characteristic of exoskeletons is that it is the actual body of the user that is moved which can be beneficial in several ways. One benefit is that the user may feel a greater degree of ownership of the actions being car-

ried out by e.g., an upper limb exoskeleton as they can see their own arm moving as opposed to a stand-alone ARM. Another benefit of an upper limb exoskeleton, and exoskeletons in general, is the potential for rehabilitation purposes [1] [10]. Additionally, an upper limb exoskeleton has the potential to provide a more integrated and less conspicuous solution than e.g., a stand-alone ARM mounted on a wheelchair. The exoskeleton could even reach a point where it could be worn under clothes, making it less eye-catching, especially as it is the person's own arm that is moving.

However, there is often a clear trade-off between the bulkiness of an upper limb exoskeleton and the support they can provide. The more minimalist upper limb exoskeletons are able to achieve their small size by only actuating one or a few joints, such as the elbow [13]. The rest of the joints in these exoskeletons are often passive and rely on the user having some residual movement, making them unsuitable for persons with tetraplegia where the arm and shoulder are paralyzed. Upper limb exoskeletons capable of fully supporting a paralyzed arm and shoulder are hence often much bulkier, as shown in Figure 2.2b, as several joints need to be actively actuated. A significant aspect of researching the design of upper limb exoskeleton in the EXOTIC project was hence how to keep the size of it minimal while still providing sufficient support for a person with tetraplegia to be able to use it.

A common challenge for both upper limb exoskeletons and ARMs in general is how to enable a person with tetraplegia to interface with the system. In case of incomplete tetraplegia, it may be possible to rely on amplifying the muscle signals by measuring EMG (electromyography) [13]. Other options include a joystick [11] or relying on hand gestures [16], in the case of a stand-alone wheelchair-mounted ARM. However, all of these options are not viable in cases of complete tetraplegia, where the persons cannot move their limbs. This scenario calls for alternative ways of interfacing, such as voice commands [11] [16] [17] or eye movements [10] [37] using e.g. eye tracking glasses as shown in Figure 2.3a. A drawback of these options is that they may be highly susceptible to noise, such as noise from the surroundings or unintentional eye movements, i.e., the Midas touch problem [14]. Some studies reported that speech recognition was the least preferred option by the majority of the participants [11].

Another option is a brain computer interface (BCI) which relies on measuring brain signals through electrodes, either surface-mounted on the head [38] [37] [1] [10] as shown in Figure 2.3b or implanted in the brain [2]. BCI using surface-mounted electrodes can be difficult to work with due to their sensitivity to noise, even in a controlled laboratory setting. However, BCI relying on implanted electrodes does not suffer from these drawbacks and has successfully been used to control a full body exoskeleton [2] but with the obvious drawback being its invasive nature.



(a) Eye tracking glasses. **(b)** A BCI cap.

Fig. 2.3: Examples of the author wearing different viable interfaces for persons with tetraplegia.

Other alternatives include low-tech devices such as chin-operated joysticks and sip-and-puff systems, which have been available for several decades. A drawback of these low-tech devices is that they are somewhat limited in the amount of unique commands they can capture from the user. Especially the sip-and-puff systems where the user is limited to either exhaling or inhaling into the device. This is less of an issue with the chin-operated joystick as continuous control is provided in multiple directions. However, a major drawback of these devices is the aesthetics, as they have to be placed near the mouth or the chin of the user.

Another option is the use of a tongue-based interface, where a touch pad-like device is placed in the roof of the user's mouth and operated by the tongue [33]. This option has multiple benefits, such as providing continuous control, allowing a high throughput of commands and being able to hide it inside the mouth of the person using it. The latter avoids issues with aesthetics and also problems with the device being in the way or obstructing the view, as is the case with e.g. a chin-joystick. Tongue-based control has previously been used successfully for controlling assistive robotic manipulators after a bit of training [32]. An important part of the EXOTIC project was hence to explore this idea of using a tongue-based control interface to enable persons with tetraplegia to control an exoskeleton.

From the above overview it is clear that it is a challenging task to provide an efficient way for persons with tetraplegia to control an exoskeleton or other assistive robotics. A common occurrence for assistive robotics is hence the inclusion of computer vision, in some capability, in an attempt to make the control easier [3].

The idea of using computer vision for part of the control has successfully been demonstrated on multiple occasions for both stand-alone arms, like JACO [15, 17, 38] but also several exoskeletons [8, 22, 28, 29]. The degree

of assistance provided from the computer vision varies from some systems taking full control [8, 22] whereas other approaches assist with sub-parts of the tasks, such as the coarse motion towards the target where the user manually controls the grasping [15]. The majority of all these approaches report an increase in performance when using computer vision as part of the control, such as requiring less time or less effort from the user to complete a certain task. However, despite these performance benefits some users actually appeared to prefer the manual option in order to retain some control, as observed in a study on vision-based control of a robotic arm for persons with SCI [18]. This could indicate that feeling in control is just as important as the objective performance, in terms of e.g. task completion speed, when using computer vision as part of a control scheme. Providing this feeling of control while still assisting the user is hence one of the focal points for the research carried out on computer vision-based semi-autonomous control in the EXOTIC project.

3 The EXOTIC Project

The EXOTIC project expands upon many of these current trends in assistive technology by proposing the novel combination of an upper limb exoskeleton controlled through a tongue-based interface combined with computer vision for more intelligent control. The overall focus of the EXOTIC project was hence to research, develop, and test such a system for use by persons with tetraplegia.

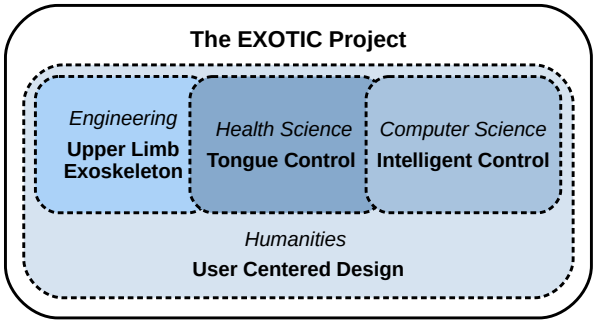


Fig. 2.4: The structure of the EXOTIC project was interdisciplinary and spanned multiple disciplines with a PhD student associated with each.

The complexity of the EXOTIC project also meant that resources from multiple disciplines had to be combined to achieve the goal. Due to the interdisciplinary nature of the project, it involved four PhD students and a postdoc each with their individual areas of expertise, as shown in Figure

2.4, all sharing a common laboratory. The focus of this PhD project was to research how computer vision could be used for intelligent control of the EXOTIC exoskeleton.

However, the other areas of the EXOTIC project are described briefly as well because they provide the context for the work described in this PhD thesis. This dependency between the different areas of the EXOTIC project also meant that the design and implementation of the system, along with planning the different studies and carrying them out were a joint effort. This is also reflected in the list of publications where several of the EXOTIC members co-authored multiple papers together.

4 User Centered Design

During the EXOTIC project it was a high priority to include the actual intended users, i.e., persons with tetraplegia, as much as possible when researching different designs and solutions for the system. User involvement was important to ensure that the research carried out in the EXOTIC project was applicable to real-world scenarios and had actual value for persons with tetraplegia.

An important aspect was hence to identify for which tasks a person with tetraplegia might imagine using an upper limb exoskeleton. These tasks were mapped out during interviews centered around playing a specially designed board game, see Figure 2.5, where the participant had to create a prioritized list of common everyday activities [20].

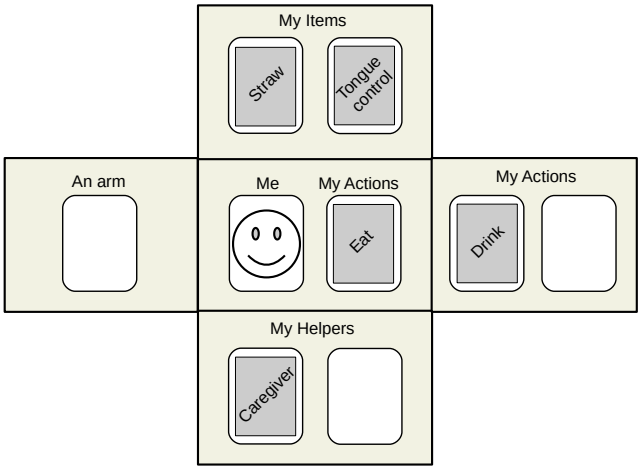


Fig. 2.5: Simplified re-creation showing part of the design game [20]. The game was used during the interviews for mapping out activities where an upper limb exoskeleton could be helpful.

The game also served as a way of making it easier to imagine what an upper limb exoskeleton is and how it could be of use to them. A total of nine persons with tetraplegia took part in these game-based interviews which were conducted one-on-one at the home of the participants [21].

From these interviews, over fifty different activities were identified where the participants imagined that an upper limb exoskeleton would be beneficial. Two activities which were consistently prioritized highly included being able to use the exoskeleton for eating and drinking. However, not necessarily in the traditional sense of sitting down and having a meal but rather prolonged sessions associated with other activities, like watching television or reading a book. One of the interviewed persons elaborated by saying that: *"A meal is on the plate as it is and is eaten in a specific tempo, but when you have a bowl of candy, then you will have to keep saying; one more, one more, one more, one more"* [21]. During the interviews, the participants were also asked to associate what they imagined would be both negative and positive attributes of an exoskeleton. Several of these attributes where associated with the size of the exoskeleton, with the positive attributes mentioning: *"small, light, and smaller at wrist and forearm"* [21] whereas the negative attributes included: *"chunky looking and sharp edges"* [21].

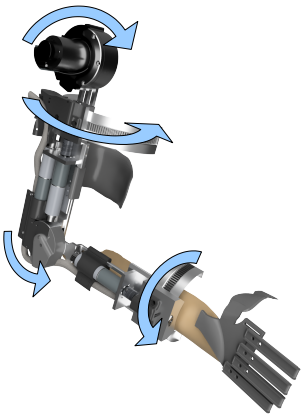
Some of the participants from the interviews also took part in the user-board meetings, where they would provide feedback on the current state of the EXOTIC project along with suggestions for further directions for research. Unfortunately some of these meetings had to be conducted online due to COVID-19 which was not ideal, especially as technical difficulties prevented participants from joining on multiple occasions. The occurrence of the COVID-19 pandemic did in general complicate the process of involving persons with tetraplegia as many of them were considered to have a higher risk of serious illness if contracting COVID-19.

5 Upper Limb Exoskeleton

In paper A we describe the upper limb exoskeleton developed during the EXOTIC project which features four degrees of freedom (DoFs), as illustrated in Figure 2.6a. A fully functional human arm features 7 DoFs and the exoskeleton can hence not reach every configuration that a human arm normally can. The decision to reduce the number of DoFs was mainly to reduce the bulkiness and size of the exoskeleton as a fully articulated upper limb exoskeleton would increase the size of the exoskeleton significantly, especially around the shoulder [12]. The decision to do so was based on the prior interviews of persons with tetraplegia where multiple persons expressed concerns regarding the size of the exoskeleton. However, the DoFs present in the exoskeleton have been selected in order to optimize its workspace for tasks including

eating and picking objects up from a table [7, 34] as they were identified to be some of the main activities desired in the interviews of persons with tetraplegia [20, 21].

The end-effector, i.e. the hand, of the EXOTIC exoskeleton constitutes an additional DoF in the form of a Carbonhand glove from Bioservo Technologies AB [6]. Figure 2.6b depicts a person wearing both the four DoFs EXOTIC upper limb exoskeleton and the Carbonhand glove. The glove enables active actuation of closing the hand of the person wearing it by contracting the thumb, index, and middle finger. Opening of the hand is achieved in a passive manner using elastic bands mounted on the glove. In paper A we demonstrate that this combination of the Carbonhand glove and the EXOTIC upper limb exoskeleton enables its user to carry out tasks such as picking up objects from a table and drinking from a bottle using a straw.



(a) The four DoFs of the exoskeleton. Figure adapted from [4], paper D.

(b) A person wearing the exoskeleton and the Carbonhand. Figure adapted from [4], paper D.

Fig. 2.6: The upper limb exoskeleton developed during the EXOTIC project.

The EXOTIC exoskeleton consists of a rigid frame made using a combination of steel and aluminum, to increase the strength and reduce its weight. Encoders are mounted at each of the four joints, making it possible to derive the position of the exoskeleton using its forward kinematics as also described in paper A. Each joint is powered by its own motor with a separate Proportional Derivative (PD) controller and the exoskeleton can hence be controlled join-by-joint at the lowest level.

However, the input from the user provided through the tongue-based interface (described next in Section 6) consists of velocity commands, instructing the end-effector to move in a certain direction in world-space and with a certain speed. This jogging of the robot was achieved using both the inverse

and forward kinematics of the exoskeleton in order to infer how each joint should move.

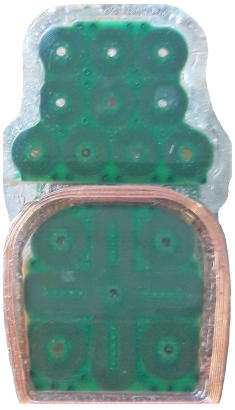
A kinematics solver based on a genetic algorithm [31] was used for the inverse kinematics. Furthermore, only the position of the end-effector was considered when solving for the inverse kinematics as the four DoFs of the exoskeleton are not sufficient to reach any arbitrary position and orientation in a 3D space. The inverse kinematics hence only considered the DoFs at the shoulder, upper arm, and elbow in order to reach a position in 3D. The last DoF around the wrist is controlled independently and directly by the user.

6 Tongue Control

The tongue-based interface used in the EXOTIC project relies on the iTongue system from TKS Technology [35] with our own customized software. The iTongue system consists of an inductive intra-oral tongue interface (ITCI) positioned in the palate of the user's mouth, as depicted in Figure 2.7a, which is held in place using either dental braces or a dental mold. A metal tongue piercing is used to activate the regions of ITCI, where the surface is covered with multiple inductive coils, as depicted in Figure 2.7b. This activation unit, i.e. metal piercing, can also be glued to the tongue for temporary usage, such as testing the system.

The ITCI is connected wirelessly through Bluetooth to an external receiver, shown in Figure 2.7c, which is connected to the EXOTIC exoskeleton. The tongue-based interface can hence be used while the mouth is closed and without any wires protruding from the mouth, making it barely noticeable.

(a) The ITCI in the roof of the mouth and the activation unit on the tongue. Figure adapted from [34], paper B.



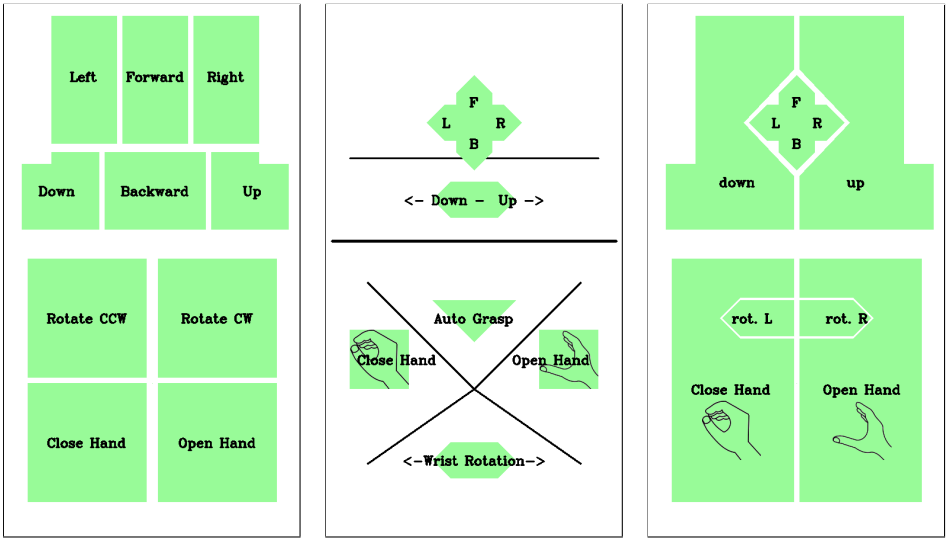
(b) *"The inductive tongue interface (ITCI)."* [4] Figure from [4], paper D.

(c) The wireless receiver. Figure adapted from [27]

Fig. 2.7: Overview of the tongue-based interface.

A wide variety of different control layouts were developed and tested [25, 27] with some of them shown in Figure 2.8. Besides the positioning of the different elements in the layout, the main difference consisted of how the different commands were triggered.

One approach was a button-based control, as shown in Figure 2.8a, where the user would trigger the different commands by moving the activation unit to some pre-defined areas for a certain amount of time. This is supposed to mimic how one would normally interact with a button. Another approach tried to mimic a joystick-based control, as shown in Figure 2.8b. The user would hence position the activation unit at an element in the control layout and afterwards drag it around. For instance, dragging the *"Down - Up"* slider in Figure 2.8b either to the right or left for the hand of the exoskeleton to go either up or down. The third approach is illustrated in Figure 2.8c and tried to mimic a gesture-based control. The user would here have to drag the activation unit across the surface of the ITCI in certain patterns to trigger different commands. For instance, moving the activation unit from left to right for the exoskeleton to move right. This control is hence somewhat similar to the joystick-based control with the exception of not requiring a certain element to be dragged around.



(a) Button-based control [25]. **(b)** Joystick-based control [4]. **(c)** Gesture-based control [27].

Fig. 2.8: Different layouts for the tongue-based interface. The joystick-based layout features an "Auto Grasp"-button as this layout was used later for testing the computer vision-based semi-autonomous control of the exoskeleton.

Evaluation of the different control modes indicated several benefits in favor of using either the joystick- or gesture-based control in favor of the

button-based control [25, 26]. Specifically in terms of reducing the task completion time during the trial but also due to other characteristics such as being able to fit more commands into the limited space of the user's palate. However, no significant difference was found between the joystick- and gesture-based control [27].

Furthermore, no significant difference in performance was found between presenting the user with visual feedback and not showing the user any visual feedback once the user had reached a certain level of proficiency in using the tongue control [27]. This visual feedback being a screen displaying the control layout along with the current position of the activation unit superimposed onto it. This indicates that the tongue-based interface could be used without visual feedback without any negative impact which is ideal as it would reduce the need for an additional screen. Subsequent studies on semi-autonomous control of the exoskeleton, as described later in Chapter 3, hence relied on the joystick-based layout without visual feedback. This is also evident by the "Auto Grasp"-button included in the joystick-based layout (see Figure 2.8b) which was needed by some of the semi-autonomous control schemes described later in Chapter 3.

Finally, all necessary software was developed for use with ROS (Robot Operating System). Both for the exoskeleton but also for the tongue-based interface and the semi-autonomous control described later on. The decision to use ROS for integrating the work of the different EXOTIC team members was paramount in ensuring a fully integrated system which could be used in real-time. This was especially important in order to make it possible to carry out an evaluation of the full system, where a person would actually use the tongue-based interface for controlling the upper limb exoskeleton with assistance from a computer vision-based semi-autonomous control.

7 Evaluation

Previously, in paper A the capabilities of solely the EXOTIC upper limb exoskeleton were demonstrated during a brief pilot test. The main purpose of this pilot was to demonstrate the functionality of the exoskeleton and it was hence carried out without any semi-autonomous control and using a gamepad controller instead of the iTongue system.

In paper B we hence describe and evaluate the fully integrated system, including both the upper limb exoskeleton, the tongue-based interface and using computer vision in a semi-autonomous control scheme. This evaluation was carried out based on ten persons without tetraplegia and three persons with tetraplegia. Several of the participants with tetraplegia had also taken part in both the prior interviews and userboard meetings, mentioned earlier. The tests including persons with tetraplegia took place at the Spinal Cord

Injury Centre of Western Denmark which also assisted with recruiting many of the participants. It should be noted that testing on persons both with and without tetraplegia was approved by the Science Ethics Committee for the Northern Region of Denmark. Furthermore, getting both this ethical approval and the process of recruiting participants for the evaluation was severely complicated by COVID-19 which caused several delays.

The evaluation consisted of the participants carrying out various tasks related to activities of daily living (ADL) which had to be completed using the upper limb exoskeleton while using the tongue-based interface. An overview of the experimental setup is shown in Figure 2.9. Two of these tasks consisted of picking up either a bottle with a straw or a plastic strawberry from a table and bringing it to the mouth. Another task was to pick up a scratching stick and bring it to the side of their face. Finally, the last task consisted of using the exoskeleton to operate a wall mounted light switch. The computer vision-based control would provide assistance in picking up the objects but not when moving towards the face of the user due to safety reasons. It should be noted that the vision-based semi-autonomous control is identical to the *"fixed semi-autonomous control"* as described in paper D and also later in Chapter 3, which both include a more extensive evaluation of different schemes for doing semi-autonomous control.

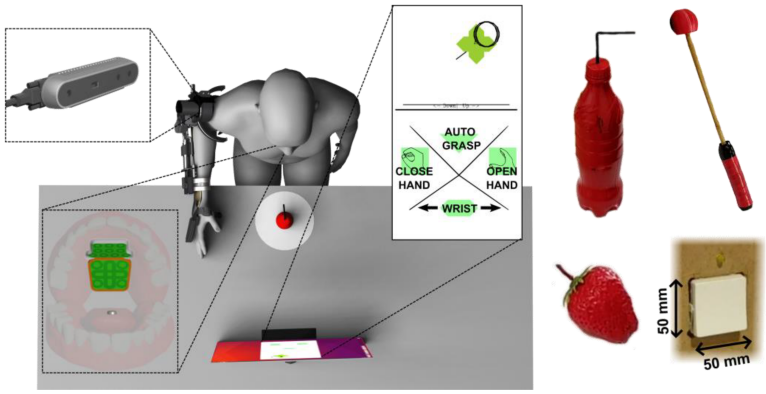


Fig. 2.9: *"Experimental setup overview. The participant was positioned in front of a table with a bottle positioned 10 cm away from the table front. The iTongue system was mounted at the palate of the participant and the activation unit was glued to the tongue. A screen on the table showed dynamic visual feedback of the control layout and the position of the activation unit on the control layout. The objects used for ADL tasks are pictured on the right. From the top left: the bottle, the scratch stick, and the strawberry."* [34] Figure from [34], paper B.

The evaluation showed that all of the participants were able to complete the tasks successfully repeatedly, including the participants with tetraplegia. The average timings for the different tasks and for both participants with and without tetraplegia can be seen in Table 2.1.

Table 2.1: The average time used by the participant when completing the four different tasks. The results are divided into participants with and without tetraplegia. Table adapted from [34], paper B.

Tetraplegia	Bottle	Strawberry	Scratch	Switch
No	38.7 ± 6.1s	62.7 ± 8.54s	70.3 ± 12.0s	$34.30 \pm 10.78s$
Yes	55.4 ± 8.0s	92.3 ± 11.6s	106.7 ± 16.9s	41.39 ± 9.47s

The participants with tetraplegia did in general appear to use more time to complete the various tasks. However, this could be attributed to differences in the amount of training in using the system which differed between the study including persons with tetraplegia and the study including persons without tetraplegia. The same observation was made for paper D and is discussed in more detail later in Chapter 3.

All three of the participants with tetraplegia took part in a semi-structured interview after having participated in the evaluation. In general they appeared positive about the entire idea of the presented system and the EXOTIC project. One of the potential users expressed the following: *"I think there is so much potential in this project. The freedom it would be to be able to pick up a bottle, drink from it yourself, and decide yourself. It would mean a massive difference. Function-wise, I think it is good, about where it should be."* [34]. Another of the participants with tetraplegia added the following: *"It has been great, great to be able to move the arm again it was delightful."* [34]. However, a few points of criticism were also expressed especially about its current appearance: *"With respect to functioning and sound, I wouldnt have second thoughts about using it, (...) but I think it is unattractive"* [34].

The presented system does hence appear to be capable of empowering persons with tetraplegia to a point where they are capable of completing some tasks on their own and potentially raise their quality of life. Furthermore, this was also confirmed in the semi-structured interviews of the participants with tetraplegia where they highlighted the great potential of the presented system and expressed that they could imagine using it in their daily life.

8 Summary

In this chapter it is described how a fully functional tongue-controlled exoskeleton was developed, along with several studies demonstrating that this exoskeleton would allow persons to perform basic tasks such as drinking and eating, using solely their tongue to control it. The final system was evaluated on several persons with tetraplegia despite COVID-19 which severely complicated the entire process. Namely recruiting persons with tetraplegia and getting the protocol for the study approved by the necessary authorities.

The main contributions and outcomes of the EXOTIC project were hence as follows:

- We describe the design and implementation of a 5 DoF upper limb exoskeleton for individuals with tetraplegia in paper A and B. The design is grounded in feedback gathered from potential users [20, 21] along with an analysis of what DoFs are necessary for which tasks [7].
- In paper A we demonstrate that the developed exoskeleton can be used for tasks such as eating snacks and drinking along with a more in-depth analysis of the exoskeletons characteristics [12].
- We designed and tested different control layouts for the tongue-based interface [27]. Furthermore, we demonstrated that these can be used without relying on any visual feedback [27] without any significant decrease in performance.
- In paper B we describe the fully integrated system with the novel combination of an upper limb exoskeleton using a tongue-based interface with an intelligent control scheme based on computer vision.
- Finally, in paper B we also evaluate this system on persons both with and without tetraplegia. These tests and subsequent interviews indicated that the novel combination of an upper limb exoskeleton, a tongue-based control, and computer vision for a semi-automatic control do allow persons with tetraplegia to regain some functionality and could improve their quality of life.

The design of the upper limb exoskeletons has hence been demonstrated to be functional on multiple occasions, also for individuals with tetraplegia. However, the design could still be optimized further and in particular its size and weight. One avenue for further research could hence focus on creating a hybrid exoskeleton, where some parts of the exoskeleton were replaced by soft robotics. For instance, the use of tendons to actuate parts of the exoskeleton. Other ideas for future work could be how to improve the tongue-based interface, and namely how to create a less invasive interface which does not require the use of a tongue piercing. Combining tongue-based control with other methods for controlling the exoskeleton could also prove beneficial, such as integrating the use of BCI as part of the control. This would also allow persons with reduced tongue movement to make use of the exoskeleton as well.

Finally, the benefit of equipping the system with a level of intelligence was briefly demonstrated in paper B. This idea of intelligent control is explored even further in the next chapter.

References

[1] M. Barsotti, D. Leonardis, C. Loconsole, M. Solazzi, E. Sotgiu, C. Procopio, C. Chisari, M. Bergamasco, and A. Frisoli, "A full upper limb robotic exoskeleton for reaching and grasping rehabilitation triggered by MI-BCI," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*. IEEE, Aug. 2015, pp. 49–54.

[2] A. L. Benabid, T. Costecalde, A. Eliseyev, G. Charvet, A. Verney, S. Karakas, M. Foerster, A. Lambert, B. Morinière, N. Abroug, M.-C. Schaeffer, A. Moly, F. Sauter-Starace, D. Ratel, C. Moro, N. Torres-Martinez, L. Langar, M. Oddoux, M. Polosan, S. Pezzani, V. Auboiroux, T. Aksenova, C. Mestais, and S. Chabardes, "An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration," *The Lancet Neurology*, vol. 18, no. 12, pp. 1112–1122, Dec. 2019.

[3] S. H. Bengtson, T. Bak, L. N. S. A. Struijk, and T. B. Moeslund, "A review of computer vision for semi-autonomous control of assistive robotic manipulators (arms)," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731–745, 2020.

[4] S. H. Bengtson, M. B. Thøgersen, M. Mohammadi, F. V. Kobbelgaard, M. A. Gull, L. N. S. A. Struijk, T. Bak, and T. B. Moeslund, "Computer vision-based adaptive semi-autonomous control of an upper limb exoskeleton for individuals with tetraplegia," *Applied Sciences*, vol. 12, no. 9, p. 4374, Apr. 2022.

[5] J. Bickenbach, A. Officer, T. Shakespeare, P. von Groote, W. H. Organization, and T. I. S. C. Society, *International perspectives on spinal cord injury / edited by Jerome Bickenbach ... [et al]*. World Health Organization, 2013.

[6] Bioservo Technologies AB, "Carbonhand," accessed: 2022-09-09. [Online]. Available: https://www.bioservo.com/healthcare

[7] E. Casanova-Batlle, M. de Zee, M. Thøgersen, Y. Tillier, and L. N. Andreasen Struijk, "The impact of an underactuated arm exoskeleton on wrist and elbow kinematics during prioritized activities of daily living," *Journal of Biomechanics*, vol. 139, p. 111137, 2022.

[8] S. Crea, M. Nann, E. Trigili, F. Cordella, A. Baldoni, F. J. Badesa, J. M. Catalán, L. Zollo, N. Vitiello, N. G. Aracil, and S. R. Soekadar, "Feasibility and safety of shared eeg/eog and vision-guided autonomous whole-arm exoskeleton control to perform activities of daily living," *Scientific Reports*, vol. 8, no. 1, p. 10823, Jul 2018.

[9] Exact Dynamics, "iARM," accessed: 2022-09-08. [Online]. Available: http://iarmrobot.com/overview.shtml

[10] A. Frisoli, C. Loconsole, D. Leonardis, F. Banno, M. Barsotti, C. Chisari, and M. Bergamasco, "A new gaze-BCI-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1169–1179, Nov. 2012.

[11] M. Gandolla, S. D. Gasperina, V. Longatelli, A. Manti, L. Aquilante, M. G. D'Angelo, E. Biffi, E. Diella, F. Molteni, M. Rossini, M. Gföhler, M. Puchinger, M. Bocciolone, F. Braghin, and A. Pedrocchi, "An assistive upper-limb exoskeleton controlled by multi-modal interfaces for severely impaired patients: development and experimental assessment," *Robotics and Autonomous Systems*, vol. 143, p. 103822, Sep. 2021.

[12] M. Gull, M. Thøgersen, S. Bengtson, M. Mohammadi, L. Struijk, T. Moeslund, T. Bak, and S. Bai, "A 4-dof upper limb exoskeleton for physical assistance: Design, modeling, control and performance evaluation," *Applied Sciences*, vol. 11, no. 13, Jun. 2021.

[13] M. Hosseini, R. Meattini, G. Palli, and C. Melchiorri, "A wearable robotic device based on twisted string actuation for rehabilitation and assistive applications," *Journal of Robotics*, vol. 2017, pp. 1–11, 2017.

[14] R. J. K. Jacob, "Eye tracking in advanced interface design," in *In W. Barfield & T. A. Furness (Eds.), Virtual Environments and Advanced Interface Design*. University Press, 1995, pp. 258–288.

[15] H. Jiang, J. P. Wachs, and B. S. Duerstock, "Integrated vision-based robotic arm interface for operators with upper limb mobility impairments," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, June 2013, pp. 1–6.

[16] H. Jiang, T. Zhang, J. P. Wachs, and B. S. Duerstock, "Enhanced control of a wheelchair-mounted robotic manipulator using 3-d vision and multimodal interaction," *Computer Vision and Image Understanding*, vol. 149, pp. 21–31, Aug. 2016.

[17] H. W. Ka, C.-S. Chung, D. Ding, K. James, and R. Cooper, "Performance evaluation of 3d vision-based semi-autonomous control method for assistive robotic manipulator," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 2, pp. 140–145, Mar. 2017.

[18] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[19] KINOVA, "Robotic arms series," accessed: 2022-09-08. [Online]. Available: https://assistive.kinovarobotics.com/product/jaco-robotic-arm

[20] F. V. Kobbelgaard, S. Bødker, and A. M. Kanstrup, "Designing a game to explore human artefact ecologies for assistive robotics," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Oct. 2020.

[21] F. V. Kobbelgaard, A. M. Kanstrup, and L. N. S. A. Struijk, "Exploring user requirements for an exoskeleton arm insights from a user-centered study with people living with severe paralysis," in *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 2021, pp. 312–320.

[22] C. Loconsole, F. Stroppa, V. Bevilacqua, and A. Frisoli, "A robust real-time 3d tracking approach for assisted object grasping," in *Haptics: Neuroscience, Devices,*

Modeling, and Applications, M. Auvray and C. Duriez, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 400–408.

[23] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, "Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics*, June 2011, pp. 1–5.

[24] J. W. Middleton, A. Dayton, J. Walsh, S. B. Rutkowski, G. Leong, and S. Duong, "Life expectancy after spinal cord injury: a 50-year study," *Journal of the International Spinal Cord Society (ISCoS)*, no. 50, pp. 803–811, 2012.

[25] M. Mohammadi, H. Knoche, B. Bentsen, M. Gaihede, and L. N. S. A. Struijk, "A pilot study on a novel gesture-based tongue interface for robot and computer control," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, Oct. 2020.

[26] M. Mohammadi, H. Knoche, and L. N. S. A. Struijk, "Continuous tongue robot mapping for paralyzed individuals improves the functional performance of tongue-based robotic assistance," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2552–2562, Aug. 2021.

[27] M. Mohammadi, H. Knoche, M. Thøgersen, S. H. Bengtson, M. A. Gull, B. Bentsen, M. Gaihede, K. E. Severinsen, and L. N. S. A. Struijk, "Eyes-free tongue gesture and tongue joystick control of a five DOF upper-limb exoskeleton for severely disabled individuals," *Frontiers in Neuroscience*, vol. 15, Dec. 2021.

[28] M. Nann, F. Cordella, E. Trigili, C. Lauretti, M. Bravi, S. Miccinilli, J. M. Catalan, F. J. Badesa, S. Crea, F. Bressi, N. Garcia-Aracil, N. Vitiello, L. Zollo, and S. R. Soekadar, "Restoring activities of daily living using an eeg/eog-controlled semi-autonomous and mobile whole-arm exoskeleton in chronic stroke," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2314–2321, 2021.

[29] V. W. Oguntosin, Y. Mori, H. Kim, S. J. Nasuto, S. Kawamura, and Y. Hayashi, "Design and validation of exoskeleton actuated by soft modules toward neurorehabilitation-vision-based control for precise reaching motion of upper limb," *Frontiers in neuroscience*, vol. 11, pp. 352–352, Jul 2017.

[30] G. Romer, H. Stuyt, and A. Peters, "Cost-savings and economic benefits due to the assistive robotic manipulator (arm)," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 201–204.

[31] S. Starke, N. Hendrich, and J. Zhang, "Memetic evolution for generic full-body inverse kinematics in robotics and animation," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 3, pp. 406–420, 2019.

[32] L. N. S. A. Struijk, L. L. Egsgaard, R. Lontis, M. Gaihede, and B. Bentsen, "Wireless intraoral tongue control of an assistive robotic arm for individuals with tetraplegia," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, Nov. 2017.

[33] L. Struijk, E. Lontis, M. Gaihede, H. Caltenco, M. Lund, H. Schiøler, and B. Bentsen, "Development and functional demonstration of a wireless intraoral inductive tongue computer interface for severely disabled persons," *Disability and Rehabilitation: Assistive Technology*, vol. 12, no. 6, pp. 631–640, 2017.

[34] M. B. Thøgersen, M. Mohammadi, M. A. Gull, S. H. Bengtson, F. V. Kobbelgaard, B. Bentsen, B. Y. A. Khan, K. E. Severinsen, S. Bai, T. Bak, T. B. Moeslund, A. M. Kanstrup, and L. N. S. Andreasen Struijk, "User based development and test of the exotic exoskeleton: Empowering individuals with tetraplegia using a compact, versatile, 5-dof upper limb exoskeleton controlled through intelligent semi-automated shared tongue control," *Sensors*, vol. 22, no. 18, 2022.

[35] TKS, "Itongue," 2022, accessed: 2022-09-08. [Online]. Available: https://tks-technology.dk/produkter/

[36] M. Wyndaele and J. J. Wyndaele, "Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey?" *Journal of the International Spinal Cord Society (ISCoS)*, no. 44, pp. 523–529, 2006.

[37] H. Zeng, Y. Wang, C. Wu, A. Song, J. Liu, P. Ji, B. Xu, L. Zhu, H. Li, and P. Wen, "Closed-loop hybrid gaze brain-machine interface based robotic arm control with augmented reality feedback," *Frontiers in Neurorobotics*, vol. 11, Oct. 2017.

[38] Z. Zhang, Y. Huang, S. Chen, J. Qu, X. Pan, T. Yu, and Y. Li, "An intention-driven semi-autonomous intelligent robotic system for drinking," *Frontiers in Neurorobotics*, vol. 11, p. 48, 2017.

Chapter 3

Human-Robot Interaction

In the previous chapter, it was described how the EXOTIC upper limb exoskeleton could be controlled using a tongue-based interface. However, using the tongue for control can be challenging at times and it does require some practice [5]. An important part of the EXOTIC project, and the focus of this thesis, was hence to enhance this control through the use of computer vision to allow the system to assist the user in controlling the exoskeleton. The user would hence control the exoskeleton in a semi-autonomous manner, where parts of the control are carried out autonomously, i.e. semi-autonomous control.

The benefits of using computer vision in such a semi-autonomous scheme for controlling assistive robotics were briefly demonstrated in paper B. Computer vision is especially useful in these scenarios as it allows the system to gather information about the current state of the world, such as nearby objects. This information can be used both to infer what the user is intending to do but also how to accomplish a certain task, like how to interact with a certain type of object. However, an important aspect of any semi-autonomous control is how the control is shared between the human and the machine, which will be discussed in more detail in the following sections.

1 Man Versus Machine?

In paper C, an extensive review was carried out on existing works on computer vision-based semi-autonomous control of assistive robotic manipulators (ARMs) and the different approaches were categorized based on the characteristics of the employed semi-autonomous control. One of the main characteristics is the level of autonomy of the machine when aiding the human, which can be viewed on a scale [7], as shown in Table 3.1. The extremes

on this scale range from the human being in complete control to the machine being in complete control.

However, characterizing the semi-autonomous behavior of an entire system using a single level of autonomy is not feasible. In paper C we hence also proposed to use a four-stage model [7] when analyzing the different systems in the review. These four different stages are illustrated in Figure 3.1 along with common examples of what each stage encompasses.

It was hence possible for a system to exhibit a high level of autonomy in one stage and a lower level in another stage. For example, the user drawing a bounding box around the object to interact with would be considered a low level of autonomy for the decision selection stage. However, the same system could then take over full control of grasping the selected object resulting in a high level of autonomy for the action implementations stage.

This four-stage model with the associated level of autonomy is of course by no means a perfect model which can encompass every small detail of the analyzed systems. However, it was a way to provide an overview of general trends in the existing systems using computer vision for semi-autonomous control of ARMs. One of such trends identified in paper C was a tendency for the hand-over between human and machine to be very clear-cut, such that either the human was in control or the machine was in control, not both at the same time.

The exception being a single system [6] in the review which relied on an established framework [3] where control inputs from both the user and from the machine are continuously blended together. The idea being the following; if the machine has a high confidence that it can assist the user it would provide a high level of assistance and vice versa. A benefit of this approach for semi-autonomous control is that it allows the system to adapt its behavior based on the situation. For instance, a situation with only one

Table 3.1: *"The different levels of autonomy. Table adapted from [7]."* [1]
Table from [1], paper C.

Levels of autonomy
1) The system offers no assistance.
2) - offers a complete set of decisions/actions.
3) - narrows down the selection to a few.
4) - suggests one alternative.
5) - executes the suggestion if the human approves.
6) - allows the human a restricted time to veto before executing.
7) - executes automatically, then necessarily informs the human.
8) - informs the human only if asked.
9) - informs the human only if it, the system, decides to.
10) - decides everything, ignoring the human.

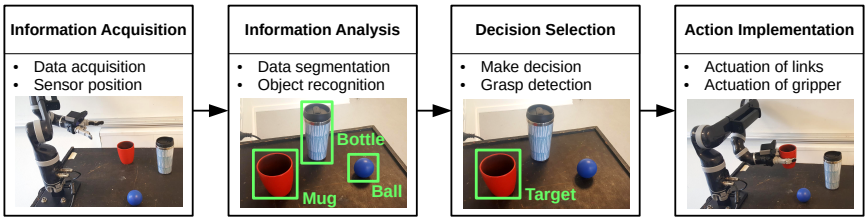


Fig. 3.1: *"The four-stage model originally proposed by [7], with examples of the tasks associated with each individual stage. The figure is adapted from [8]."* [1] Figure from [1], paper C.

object in front of the user makes it likely that the user will want to interact with that object versus a scenario with multiple objects, where it is more difficult to predict the intention of the user.

Another important aspect of this adaptive level of autonomy is how the human will never give up control completely as opposed to the other approaches, where there was a clear handover in control between human and machine. While these clear-cut approaches may perform better, in terms of e.g. task completion speed, it may not always be the best option. In a study on autonomous control of a wheelchair-mounted robotic arm for persons with spinal cord injuries the majority of the participants reported higher satisfaction when allowed to control the robotic arm manually [4]. It is not difficult to imagine that having to relinquish control of the robotic arm may have some resemblance to the loss of control, one may feel when sustaining a spinal cord injury in the first place, at least to some extent.

This issue is likely even more crucial for an upper limb exoskeleton than a standalone robotic arm, as it is the user's own arm and hand being moved around. For an exoskeleton, it will also be much harder to disassociate oneself from what the exoskeleton is doing in order to concentrate on doing other things, which is less of an issue with a standalone robotic arm mounted on e.g. a wheelchair. This also underlines another important aspect when designing semi-autonomous control for an exoskeleton; the user will always be physically present and involved in the task. Disregarding any input from the user, at least in some parts of the control, can hence be seen as a waste of resources. The human wearing the exoskeleton might as well be involved in the control at all times, such that e.g. small corrections can be made. After all, the ultimate goal of a semi-autonomous control for an exoskeleton is to be able to mimic all the small things we humans do unconsciously when e.g. picking an object up from a table.

2 Semi-Autonomous Control of an Upper-body Exoskeleton

In paper D we expand upon the findings from paper C by designing and evaluating an adaptive semi-autonomous control for the EXOTIC upper limb exoskeleton. The main idea of this adaptive control scheme was to assist the user without completely taking away control from the user. This approach was evaluated against a fixed semi-autonomous control scheme and against a manual control of the exoskeleton. The evaluation was carried out across two studies, where one group of participants had tetraplegia and the second group did not.

The semi-autonomous control for the exoskeleton was designed based on the same general framework [3] as also used by the only approach with an adaptive level of autonomy [6] identified during the review in paper C. An outline of this framework is illustrated in Figure 3.2, where input from the user U is blended with the predicted goal P from the system to actuate an assistive robotic manipulator. It is hence a general framework which can be used in different contexts. For instance, teleoperation of a robotic manipulator, using either a BCI [6] or the pose of the user as input for the system [3].

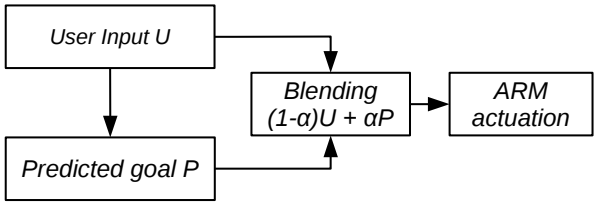


Fig. 3.2: *"Arbitration between the user U and the goal P predicted by the system using linear blending. The figure is adapted from [3]."* [1] Figure from [1], paper C.

In paper D, the context was hence slightly different as the ARM being actuated was the EXOTIC upper limb exoskeleton. Furthermore, a tongue-based interface was used for user input and computer vision was used to predict the most likely goal of the user.

An important aspect of the framework in Figure 3.2 is the arbitration factor α, controlling the blending between the user input and the predicted goal. While the blending itself happens in a linear fashion it is possible to change the behavior of the semi-autonomous control radically depending on how the arbitration factor changes depending on the confidence of the system. This dependency can be illustrated as an arbitration curve, depicting how the arbitration factor changes as a function of the confidence of the system. The arbitration curve can be any arbitrary function and this framework can hence

encompass radically different semi-autonomous control schemes. Examples are shown in Figure 3.3, which depicts the three different arbitration curves used for the control schemes evaluated in paper D.

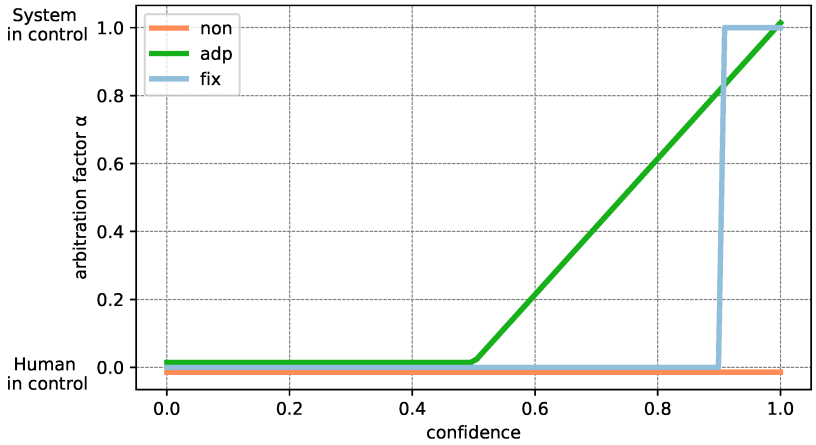


Fig. 3.3: *"The behavior of the different control schemes is illustrated using an arbitration curve. The non-autonomous control (non) is fixed at $\alpha = 0$ as the human is always in control. The curve for the adaptive semi-autonomous control (adp) is given by the function $\alpha = \max(0, 2\rho - 1)$. The fixed semi-autonomous control (fix) is characterized by a sudden jump from the human being in control to the system being in control, which is triggered when the user presses the "auto grasp" button."* [2]
Figure from [2], paper D.

The three depicted arbitration curves in Figure 3.3 are indicative of the following three control schemes:

- **Non-Autonomous Control** (*non*), where the arbitration curve is fixed at $\alpha = 0$ for all levels of confidence. The human will hence always be in control, and it is hence essentially just manual control.
- **Adaptive Semi-Autonomous Control** (*adp*), where the arbitration curve is fixed at $\alpha = 0$ as well but only until a certain level of confidence is reached. Once this confidence threshold is reached the arbitration factor increases linearly with the confidence. This arbitration curve is hence an example of a semi-autonomous control with an adaptive level of autonomy.
- **Fixed Semi-Autonomous Control** (*fix*), where the arbitration curve is fixed at $\alpha = 0$ until a certain confidence level is reached as well. However, once this threshold is reached the transition from the human being in control to the machine being in complete control happens instantly, as signified by the sudden spike. This behavior is identical to a lot of existing systems employing a clear-cut strategy for semi-autonomous control, where either the human or the machine is in complete control.

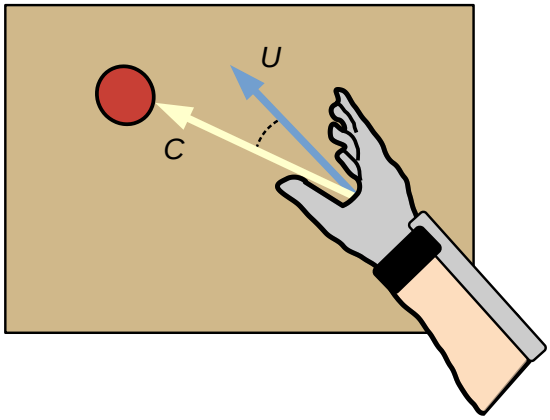


Fig. 3.4: The idea of the confidence measure depicted in 2D for illustration purposes. The user is instructing the exoskeleton to move forward, resulting in the vector U. An object has been found by the computer vision module, resulting in the vector C, spanning from the end-effector of the exoskeleton, i.e. the hand, to the position of the object. The confidence of the system depends on the angle, i.e. similarity, of these two vectors.

A common trait for all three arbitration curves is how they are a function of the confidence of the system. Being able to estimate this confidence is hence an important prerequisite for using this blending based approach for the semi-autonomous control. A direction-based approach was used to estimate the confidence of the system, as illustrated in Figure 3.4. This confidence measure was essentially based on the angle between the direction vectors for user input and for the object detected by the computer vision module. If the user steered directly towards the object, the angle between these two vectors would be small and it would result in a high confidence value, and vice versa.

A computer vision module enabled the system to detect the various objects in the scene which was necessary for both calculating the above confidence measure but also for inferring how to grasp the different objects. The general pipeline of the computer vision employed for paper D is shown in Figure 3.5. A classical approach relying on color segmentation was used as the basis for detecting the different objects. This approach was used to ensure stable and consistent detections while conducting experiments with semi-autonomous control of the exoskeleton. More sophisticated methods based on deep learning were considered but ultimately discarded to avoid introducing more uncertainty into the results than necessary. This decision was justified by the fact that the purpose of paper D was to test the semi-autonomous control schemes and not computer vision algorithms.

Finally, the last stage of the computer vision module was to calculate how to grasp each of the detected objects. This grasp detection was accomplished

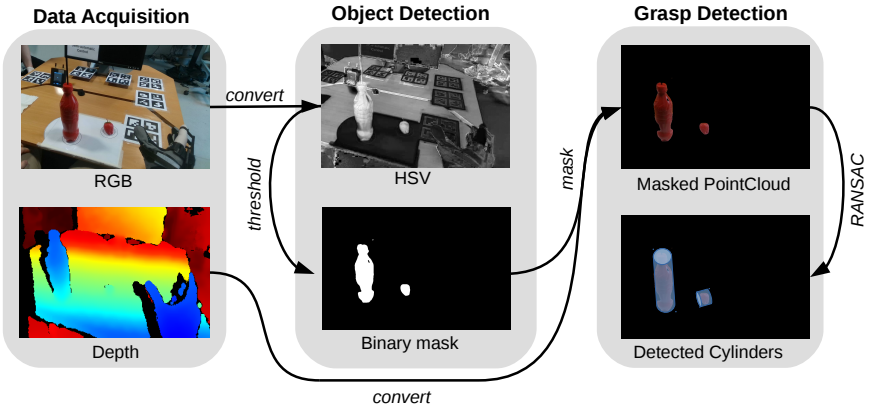


Fig. 3.5: *"Overview of the pipeline for the computer vision module. An RGB-D camera (Intel RealSense D415) is mounted at the shoulder joint of the exoskeleton and captures both RGB and depth information from the area in front of the user. The object detection relies on the RGB data where objects are detected using color thresholding. The depth information is masked based on the detected objects and then converted to a point cloud. Cylinder-like shapes are then detected in the resulting masked point cloud using an RANSAC-based algorithm. Finally, the detected cylinders are converted to grasp poses for the exoskeleton using a rule-based approach."* [2] Figure from [2], paper D.

using a rule-based approach relying on the assumption that each object could be approximated by a cylinder. This assumption was true for some objects but a bit of a stretch for others. Nevertheless, preliminary tests confirmed that the above approach worked for grasping all the objects included in the experiments.

3 Evaluation

The system described above was designed based on the findings of paper C which indicated that an adaptive approach for the semi-autonomous control of the EXOTIC exoskeleton could prove beneficial. It was therefore necessary to evaluate whether or not this assumption was true, which resulted in the following three hypotheses:

- *"**Hypothesis 1:** The adaptive semi-autonomous control is better than the non-autonomous control." [2]*
- *"**Hypothesis 2:** The fixed semi-autonomous control is better than the non-autonomous control." [2]*
- *"**Hypothesis 3:** The adaptive semi-autonomous control is better than the fixed semi-autonomous control."* [2]

These three hypotheses were evaluated across two studies; study A containing 10 participants without tetraplegia and study B containing 7 partic-

ipants with tetraplegia. In both studies, the participants had to use the exoskeleton to grasp a predefined object and then lift it. This was tested using the tongue-based control with the three different control schemes outlined above. An example of the test setup is shown in Figure 3.6 along with the two different objects that the participants were asked to grasp; a strawberry and a bottle. In some cases, both objects were present in the scene at the same time and in other cases only the object to grasp was present in the scene. This choice was intentional to test the impact of the system having to decide between multiple objects.

However, the cases with multiple objects were not tested for study B as it was found necessary to cut each session of experiments shorter due to fatigue of the participants. This fatigue was not due to the use of the presented system or the EXOTIC exoskeleton but instead it was a matter of fatigue from traveling. Many of the participants with tetraplegia in study B unfortunately had to travel for several hours prior to the experiment.

(a) *"The test setup."* [2]
Figure adapted from [2], paper D.

(b) *"The objects for grasping."* [2]
Figure adapted from [2], paper D.

Fig. 3.6: *"An overview of the test setup. (**a**) The participant is placed in a wheelchair with the exoskeleton attached to the right arm in the starting position. In front of the participant is a table with different objects to grasp. (**b**) The two objects used in the studies for the task of grasping: a plastic strawberry and a plastic bottle. The objects can be placed on the two predefined positions marked on the table below the objects."* [2]

During the two studies, the performance of the three control schemes was measured based on the time it took to complete the tasks, the number of commands to do so and the length of travel for the end-effector in the Cartesian 3D workspace. Furthermore, two questionnaires were used to assess the users' perception of the different control schemes in terms of their intuitiveness and how they affected the perceived difficulty of the tasks.

The three control schemes were compared in a pair-wise manner to identify any statistically significant difference between using them. A description of this statistical analysis can be found in paper D.

3.1 Study A - Results

The results from study A are shown in Table 3.2, which depicts the pair-wise differences between the tested control schemes, along with their associated hypotheses. The adaptive semi-autonomous control does in general appear to be performing the best whereas the non-autonomous control generally performs the worst.

Table 3.2: *"Study A - Pair-wise comparison between the three tested control schemes. The mean percentage-wise increase in performance for each comparison is reported, where positive numbers denote an improvement (i.e. reduction) in favor of the hypothesis. (...) Significant results supporting the hypothesis are marked with green while results supporting the hypothesis but lacking significance is marked with yellow. Results marked with red does not support the hypothesis."* [2].
Tables adapted from [2], paper D.

(a) *"Hypothesis **H1**, adaptive semi-autonomous control (adp) is better than non-autonomous control (non)."* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	**37%**	**37%**	**33%**
Strawberry Single	**58%**	**60%**	**38%**
Bottle Multi	31%	17%	**30%**
Strawberry Multi	**43%**	**44%**	16%

(b) *"Hypothesis **H2**, fixed semi-autonomous control (fix) is better than non-autonomous control (non)."* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	**18%**	−4.1%	**23%**
Strawberry Single	**46%**	**37%**	25%
Bottle Multi	14%	−26%	19%
Strawberry Multi	**49%**	**35%**	**28%**

(c) *"Hypothesis **H3**, adaptive semi-autonomous control (adp) is better than fixed semi-autonomous control (fix)."* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	23%	**40%**	13%
Strawberry Single	**22%**	**36%**	**17%**
Bottle Multi	20%	**39%**	13%
Strawberry Multi	−10%	13%	−15%

The adaptive semi-autonomous control outperforms the non-autonomous control scheme across all metrics and all scenarios, as seen in Table 3.2a. All of these improvements are statistically significant with the exception of three cases for the scenarios with multiple objects present. This strongly supports the hypothesis of the proposed adaptive semi-autonomous control being superior to manual control, at least when considering the objective metrics measured.

The difference is less pronounced in the pair-wise comparison between the fixed semi-autonomous control and the non-autonomous one, as shown

in Table 3.2b. There are even a few cases where using the fixed semi-autonomous control instead of the manual one will negatively impact the performance of the user. Specifically in terms of the required commands needed to complete the scenarios consisting of grasping a bottle. However, neither of these two cases carries any statistical significance, unlike the seven statistically significant cases where the fixed semi-autonomous control does indeed increase the performance. These results imply the plausibility of the hypothesis that fixed semi-autonomous control is superior to manual control.

Finally, the pair-wise differences between the two semi-autonomous control schemes are reported in Table 3.2c. The adaptive scheme is in a few cases performing worse than the fixed one. Specifically in the cases when a strawberry has to be grasped while multiple objects are present in the scene. However, the adaptive scheme outperforms the adaptive one in all the other cases and with half of these cases being statistically significant. The majority of significant improvements of using the adaptive scheme are found when measuring the number of commands needed to complete the different tasks or when the task involves a strawberry. This difference in the number of commands used could be explained by the fact that the user has to press the "Auto Grasp"-button, as illustrated earlier in Figure 2.8b, in order to activate the fixed semi-autonomous control scheme. Furthermore, it was observed that the seamless nature of the adaptive semi-autonomous control scheme worked really well for scenarios with a strawberry as these would require both rotating and positioning the hand of the exoskeleton in order to successfully grasp the strawberry. The user would in these cases mainly focus on the positioning while the adaptive semi-autonomous control ensured the correct rotation of the hand in the mean time. The hypothesis of the adaptive scheme being preferable to the fixed one is clearly true in some cases. Furthermore, there are no statistically significant results to support the opposite.

The gathered objective metrics were supplemented by two questionnaires to assess both the intuitiveness (INTUI) of the control schemes and the perceived difficulty of the tasks with the different control schemes (NASA-TLX). Both of the semi-autonomous control schemes were rated significantly better on both questionnaires as compared to their non-autonomous counterpart. There was no statistically significant difference between the adaptive and fixed scheme for either of the two questionnaires. These results do support earlier observations from the object metrics in Table 3.2 quite well as the difference between the adaptive semi-autonomous control versus the manual control was more pronounced than the difference between the two semi-autonomous control modes.

3.2 Study B - Results

The results for study B are summarized in Table 3.3 and only contain results for the cases with a single object present in the scene, as mentioned earlier.

The results from study B exhibit many of the same trends as observed earlier in study A but with less statistical significance in general. For instance, both of the semi-autonomous control modes outperform the non-autonomous one in many cases with approximately half of them being statistically significant as well. However, the fixed semi-autonomous control schemes appear to perform slightly better as it improves performance in all cases in Table 3.3b as compared to study A, where using the fixed scheme decreased performance in a few cases. The difference between the two different semi-autonomous control modes also appear less pronounced in Table 3.3c as compared to study A.

Table 3.3: *"Study B - Pair-wise comparison between the three tested control schemes. The mean percentage-wise increase in performance for each comparison is reported, where positive numbers denote an improvement (i.e. reduction) in favor of the hypothesis. (...) Significant results supporting the hypothesis are marked with green while results supporting the hypothesis but lacking significance is marked with yellow. Results marked with red does not support the hypothesis."* [2].
Tables adapted from [2], paper D.

(a) *Hypothesis **H1**, adaptive semi-autonomous control (adp) is better than non-autonomous control (non).* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	**41%**	43%	41%
Strawberry Single	**54%**	**56%**	33%

(b) *Hypothesis **H2**, fixed semi-autonomous control (fix) is better than non-autonomous control (non).* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	**53%**	42%	**53%**
Strawberry Single	**54%**	50%	**31%**

(c) *Hypothesis **H3**, adaptive semi-autonomous control (adp) is better than fixed semi-autonomous control (fix).* [2]

	time (seconds)	commands (integer)	cartesian (meters)
Bottle Single	−21%	0.37%	−21%
Strawberry Single	−0.29%	12%	2.6%

The same lack of significance was also observed in the results from the questionnaires as well, where no statistical significance can be found between any of the control modes for either of the two questionnaires. Part of this difference in statistical significance between the two studies could likely be attributed to fewer participants in study B (7 persons) compared to study A (10 persons). Another difference between the two studies is also the training-

phase for learning to use the system. In study A the participants had more time to train tongue-based control on the real exoskeleton whereas the training phase was shorter for study B and the training was done on a simulated version of the exoskeleton. Both of these discrepancies between the two studies were a matter of prioritizing available resources and namely time, but should be addressed in future experiments.

4 Summary

This chapter has described how the use of computer vision often is beneficial in various semi-autonomous control schemes for assistive robotic manipulators, as outlined in paper C. Furthermore, it has been identified that there is a general trend of using very clear-cut strategies for arbitrating between the human and the system, where the level of autonomy is pre-defined. However, it was also found that such solutions may not be the best choice for persons with movement impairments as it can result in a feeling of not being in control. A semi-autonomous control relying on an adaptive level of autonomy was hence developed for the tongue-controlled EXOTIC exoskeleton and evaluated across two studies, as described in paper D. The main contributions and outcomes of these two pieces of work can be summarized as:

- In paper C we systematically reviewed existing approaches for using computer vision for semi-autonomous control of assistive robotic manipulators.
- Furthermore, in paper C we also identified current trends and highlighted short-comings in the current state of the work. Namely the static nature of many semi-autonomous control schemes where it was theorized that a more adaptive scheme would be better. Especially for persons with tetraplegia or other movement impairments.
- We elaborated on this finding in paper D where we proposed a computer vision-based semi-autonomous control scheme for the EXOTIC upper limb exoskeleton. This control scheme was characterized by being adaptive such that the level of assistance provided by the system would vary depending on the scenario.
- Furthermore, this adaptive semi-autonomous control scheme was also evaluated in paper D against a semi-autonomous control scheme with a fixed level of autonomy and a fully manual control scheme, i.e. without any autonomy. This evaluation consisted of a study including ten persons without tetraplegia and another study including seven persons with tetraplegia.

- The findings in paper D strongly suggest that the idea of using computer vision for an intelligent control of an upper limb exoskeleton is indeed beneficial. Both objective metrics (e.g. task completion time and commands used) improved significantly in many cases and the users' experience of using the system improved as well.
- Finally, the results in paper D also support the hypothesis from paper C of an adaptive semi-autonomous control scheme being superior to one with a fixed level of autonomy. Comparing these two schemes did in several cases result in a significant difference in favor of the adaptive approach. Most of the cases lacking significance were also in favor of the adaptive scheme with fewer cases favoring the fixed approach. However, all cases in favor of the fixed scheme lacked significance.

Looking at the results from the two studies, many of the same trends were identified for both of them. However, the results from study B were generally found to be lacking statistical significance which is likely due to having fewer participants when compared to study A. A suggestion for future work would hence be to re-conduct study B with more participants and preferably also with the scenarios containing multiple objects which had to be cut. It could also be considered if the experiments could be conducted either in the home of the participants or at a nearby location if study B was to be repeated. This would hopefully reduce the fatigue from traveling in some of the participants with tetraplegia and also allow for more time for training to use the real exoskeleton.

Looking at the results from study A, which included scenarios with both single and multiple objects in the scene, it is clear that both semi-autonomous control schemes perform worse in scenarios with multiple objects. A part of this difference could be linked to the intent prediction which relies on a rather basic direction-based approach, where the system assumes that the user would point the exoskeleton arm towards the object of interest. This assumption may not be true in all cases and it may therefore be beneficial to look into improving the approach for the intent prediction.

Finally, another option for future improvements would be to focus on the computer vision part which was tailored to the exact tasks present in the studies. The computer vision pipeline in its current state would hence struggle if presented with objects where there is little to no resemblance to the objects used in the experiments. For instance, the assumption of being able to fit a cylinder to the detected object in order to infer a grasp pose using the rule-based approach employed in paper D. An alternative could be to train a neural network to estimate the pose of the different objects, with the pose of the objects being useful in determining how to interact with them, e.g. grasp them. This does however introduce new problems, such as training the network in terms of obtaining the required training data and formulating

a meaningful loss function. Furthermore, such a pose estimation approach needs a low inference time to be useful in the context of controlling the EXOTIC exoskeleton. Having to wait several seconds for the computer vision module would likely impact the performance of the semi-autonomous control severely. These issues and many others are addressed in the following chapter where a neural network is trained for the task of doing pose estimation of objects.

References

[1] S. H. Bengtson, T. Bak, L. N. S. A. Struijk, and T. B. Moeslund, "A review of computer vision for semi-autonomous control of assistive robotic manipulators (arms)," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731–745, 2020.

[2] S. H. Bengtson, M. B. Thøgersen, M. Mohammadi, F. V. Kobbelgaard, M. A. Gull, L. N. S. A. Struijk, T. Bak, and T. B. Moeslund, "Computer vision-based adaptive semi-autonomous control of an upper limb exoskeleton for individuals with tetraplegia," *Applied Sciences*, vol. 12, no. 9, p. 4374, Apr. 2022.

[3] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[4] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[5] M. Mohammadi, H. Knoche, M. Thøgersen, S. H. Bengtson, M. A. Gull, B. Bentsen, M. Gaihede, K. E. Severinsen, and L. N. S. A. Struijk, "Eyes-free tongue gesture and tongue joystick control of a five DOF upper-limb exoskeleton for severely disabled individuals," *Frontiers in Neuroscience*, vol. 15, Dec. 2021.

[6] K. Muelling, A. Venkatraman, J.-S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to brain computer interface controlled manipulation," *Autonomous robots*, vol. 41, no. 6, pp. 1401–1422, 2017.

[7] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[8] B. Pitzer, M. Styer, C. Bersch, C. DuHadway, and J. Becker, "Towards perceptual shared autonomy for robotic mobile manipulation," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 6245–6251.

Chapter 4

Computer Vision for Object Manipulation

In the previous chapter it was outlined how computer vision can be used for different aspects of the control of assistive robotics. Namely, the four-stage model consisting of information acquisition, information analysis, decision selection and action implementation to analyze existing systems. The work presented so far has mainly focused on the last stage (i.e. action implementation) in terms of a blending-based adaptive semi-autonomous approach for controlling an upper limb exoskeleton, as outlined in paper D. This work mainly relied on classical computer vision methods, where color thresholding was used to detect objects in the information analysis stage and a rule-based approach was used to infer a grasping pose in the decision selection stage.

In this chapter, other ways of approaching the decision selection stage will be discussed. Namely, the work carried out as part of this PhD project on how to estimate the pose of an object, which can form the basis for how to manipulate an object. For instance, how to grasp said object. This is especially important in the context of EXOTIC as one of the main purposes of the upper limb exoskeleton is to be able to manipulate objects.

1 Pose Estimation of Objects

Pose estimation is the process of estimating both the orientation and position of an object, as shown in Figure 4.1. Sometimes also referred to as 6D pose estimation, where the orientation (α, β, γ) contributes with 3D and the position (x, y, z) contributes with another 3D as well. The problem of pose estimation of objects is a problem which has been studied for several decades [12]. However, despite a lot of progress in recent years, especially due to deep learning, it still remains a challenging problem [9].

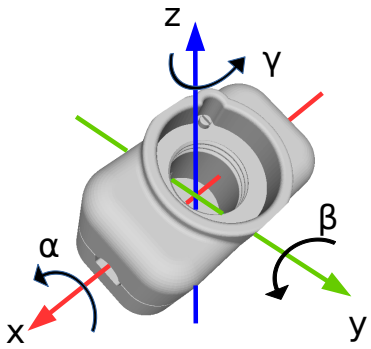


Fig. 4.1: The pose of an object given by its orientation (α, β, γ) and position (x, y, z).

Pose estimation is especially useful in the context of object manipulation as how to e.g. grasp an object depends on both its position and orientation. If the pose of an object is known it can be used to infer how to grasp the object, for instance from some pre-defined grasp poses associated with each object.

Until recent years, the state-of-the-art for pose estimation was dominated by key-point based approaches [15] [6]. These approaches relied on hand-crafted features [3] calculated from depth information which were then used to match against CAD models of different objects in order to infer both the object type and pose.

A few years ago, deep learning-based approaches started to surpass these key-point based approaches in terms of performance, as also highlighted by the benchmark for 6D object pose estimation (BOP) challenge [9]. Some of the deep learning-based approaches trained their model in an end-to-end fashion to directly regress the pose [11] [10]. Others trained a feature extractor [14] which was subsequently used to match against a codebook of known objects with known poses. All of these approaches are primarily based on RGB information as input, where depth information is only an optional input used for post-refinement of the estimated poses. For instance, using an iterative closest point (ICP) algorithm [16] to minimize the error between the observed depth and a CAD model of the object.

There has hence been a clear change in the input modality, from relying solely on depth information to using RGB information with depth as an optional additional input. This distinct change in the input modality could be a matter of deep learning not being straightforward to use for depth information with many of the operations being optimized for RGB data. It could also be a matter of the RGB information being less noisy, and e.g. having sharp edges, than the depth information which can be more susceptible to noise, for instance, sunlight interfering with depth sensors relying on the IR spectrum.

A drawback commonly stressed in terms of deep learning-based approaches is the vast amounts of labeled data needed for training. This issue is also prevalent for the mentioned deep learning-based approaches but was circumvented by either training solely [14] or primarily [10] [11] on synthetic data generated from CAD models.

Recently, there has been a trend of merging the idea of key-point based approaches with deep learning. These approaches rely on deep learning for learning descriptors for fragments on the surface of the object [4] [7]. All of these approaches primarily rely on RGB information as input, like the previous deep learning-based approaches, with depth information only being used in a post-refinement step. Similarly to the key-point based approaches relying on depth information [15] [6], these fragment-based approaches require matching the extracted descriptor with a reference model in order to infer the pose. This process can be rather costly in terms of time, ranging from an average of 0.75 seconds per image containing multiple objects (using a Tesla P100 [7]) to an average of 2.2 seconds per object (using an RTX2080 [4]). For comparison, the deep learning-based approach relying on codebooks for inference [14] is 3-4 times faster than these approaches [7]. However, the codebook-based approach is also less accurate and it is hence a trade-off between speed versus accuracy. Which of the two should be preferred is of course dependent on the use-case.

Considering the context of this PhD project, where the pose estimation is intended to assist in controlling an exoskeleton, it was deemed that speed should be preferred over accuracy. This is based on the consideration that the system should be responsive for the user controlling it. Another consideration was that the human is still present in the control loop of the upper limb exoskeleton, as outlined previously in Section 2. The human can hence help with minor corrections, as long as the estimated pose is not totally off and not having the highest possible pose accuracy is therefore less of an issue.

An obvious drawback of the codebook-based approach [14] despite its low inference time is the codebooks. The codebooks are generated object-wise by sampling $\approx$ 65.000 poses where a feature vector is computed for each and stored for use during inference. This requires the space of poses to be discretized and also introduces a memory consumption which scales linearly with the number of objects that the system has to handle.

In paper E we show that each codebook can be replaced entirely by a small neural network for pose regression, consuming $\approx$ 40 times less memory, while improving the accuracy of the pose estimates. Furthermore, a novel loss function based on differentiable rendering is introduced in order to train the pose regression network. This loss function is designed such that it inherently can handle ambiguities caused by symmetries which are an inherent problem in pose estimation.

This work is further expanded in paper F where a single pose regression

network is shown to be sufficient for multiple different objects. This approach improved the pose estimation accuracy even further while reducing the memory footprint once again.

2 Pose Ambiguities due to Object Symmetries

Symmetries are an inherent problem in pose estimation, as they can cause ambiguities in gauging how correct an estimated pose is. An example of such could be a cylinder-shaped object, as shown in Figure 4.2a, where the appearance and physical properties of the object do not change when rotating the object around its major axis. However, rotational symmetries like the one in Figure 4.2a are due to the object being truly symmetric. It is hence possible to predefine these symmetries such that a pose error function can account for them.

An example of such is the two pose error metrics, *Maximum Symmetry-Aware Surface Distance* and *Maximum Symmetry-Aware Projection Distance*, found in the BOP benchmark for 6D object pose estimation [9]. Both functions rely on a set of predefined global symmetric transformations (such as rotating a cylinder around its major axis) which the error function should apply such that the final error is minimized.

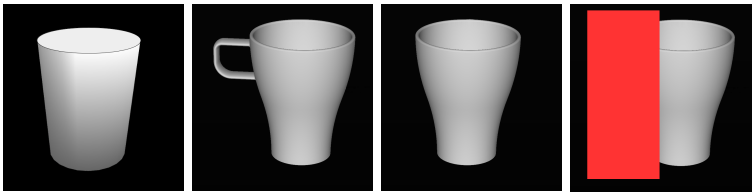


Fig. 4.2: *"(a) Rotationally symmetric objects should be treated equally independent of angle around its major axis. Examples of how symmetries can occur for a mug with a handle. (b) Handle visible, no pose ambiguity. (c) Self-occluded due to a slight rotation and (d) occluded by another object, both of these have ambiguities in pose."* [1] Figures from [1], paper E, © 2021 IEEE.

However, relying on predefined symmetries will fail to encompass situations where an object can appear symmetric without being it. This is depicted in Figure 4.2b-d, where a coffee mug has no apparent symmetries when the handle is showing (b) but once the handle is hidden, either due to self-occlusion (c) or occlusion by another object (d), it appears symmetric. The likelihood of such situations occurring is especially high for objects with little to no texture to help solve these ambiguities.

Scenarios with apparent symmetries, like the ones depicted in Figure 4.2c-d, are hence problematic when evaluating pose estimation approaches. But it is also problematic when training learning-based approaches for pose esti-

mation as such approaches require an error function in order to improve and learn. If this error function behaves inconsistently, due to symmetries, the model being trained will have to learn to account for this as well. In the best case, the model will use a lot of resources in order to learn to account for it. In the worst case, the model will fail to encompass the concept of symmetries and achieve sub-par performance as a result of this.

3 Pose Error based on Visual Similarity

Another way to approach the issue of handling symmetries when doing pose estimation is to compare poses based on their visual similarity. This exact same idea is also the underlying basis for the *Visible Surface Discrepancy* (VSD) metric used in the BOP challenge [9].

In paper E we formulate a novel loss function based on the same idea of comparing poses through their visual similarity, as shown in Figure 4.3. This loss function is then used to train a neural network for doing pose regression on top of the encoder from the codebook-based approach [14], such that the codebook is entirely replaced by a small pose regression network. All the training data is still synthetic and generated from CAD models of the different objects, similarly to the codebook-based approach.

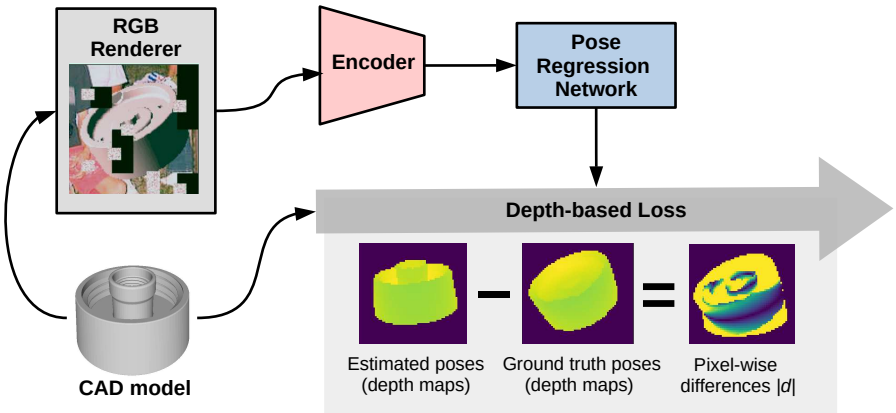
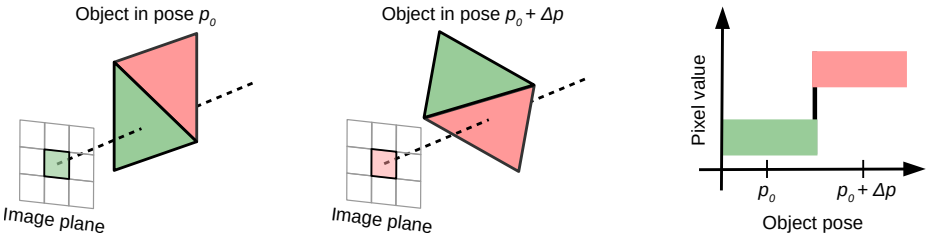


Fig. 4.3: Overview of the components in our proposed pipeline for training a pose regression network. The network is trained from synthetic data in the form of augmented renders of the objects from their respective CAD models. A pre-trained encoder [14] is used as a feature extractor and its output is used as input for our pose regression network. During training, a depth-based loss function is used for comparing the visual similarity between the ground truth and the estimated poses. Figure adapted from [1, 5], paper E, and [5], paper F, © 2021/2022 IEEE.

An important part of training the pose regression network is hence our custom loss function which relies on comparing depth maps. These depth maps are produced using a differentiable depth renderer [13], such that the loss function is still differentiable which is required for doing back-propagation when training the network. This is necessary as rendering images of e.g. a CAD model is not differentiable per default in a common rendering pipeline. The main issue occurs when relying on rasterization to convert a 3D scene, containing e.g. a CAD model, into a 2D raster of pixels, i.e. an image. During this process it is necessary to check every object in the scene to determine how they are associated with the pixels in the image, as illustrated in Figure 4.4a.



(a) Rasterization of a scene with an object consisting of a few faces, i.e. triangles.

(b) Rasterization of the scene after applying a small change to the pose of the object.

(c) The value of a single pixel plotted as a function of the pose of the object in the scene.

Fig. 4.4: Illustration of the rasterization step in a traditional rendering pipeline and why it can be problematic for training a neural network. **(a)** The rasterization step serves to identify the intersection between faces in a scene and pixels in an image. **(b)** Applying a small change to the scene, e.g. changing the orientation of an object slightly, may cause intersection between faces and pixels to change drastically. **(c)** Pixel values may as a result of the small change in pose suddenly and drastically change as well. This sudden jump in the pixel values is not easily differentiable which is problematic for training a neural network. Note that the small change in pose has been exaggerated for illustrative purposes. Figure adapted from [13].

However, even small changes to the scene may cause a drastic change of the individual values of each pixel in the image as they may suddenly overlap with a different face, i.e. triangle, in the scene. For instance, a small change in the orientation of an object in the scene, as shown in Figure 4.4b, will change the intersection between the faces in the scene and the pixels in the image. Such a small change in the scene may hence result in a sudden and drastic change in the value of one or more pixels, as also illustrated in Figure 4.4c. This sudden jump from one value to another is not easily differentiable and hence not suitable for training a neural network. The differentiable renderer used in both paper E and F avoids this issue by aggregating the k closest faces for each pixel instead of just considering a single face as done in traditional rasterization.

4 Multiple Views to Escape Local Minima

The depth map-based loss function did however tend to get stuck in local minima. This issue is sought illustrated in Figure 4.5, where an object is rotated in terms of its elevation and azimuth and the resulting depth-based loss is plotted. The visual appearance of the correct pose *(a)* is nearly identical when rotating the object 180^o around an axis as seen in *(c)*. This is reflected in the loss landscape, with *(a)* and *(c)* being distinct minima. If the pose regression network should predict a pose corresponding to *(c)* instead of *(a)* it is deemed acceptable due to their high visual similarity. However, there are also two distinct minima at *(b)* and *(d)*, when the object is rotated roughly 90^o around one axis. These minima are more problematic as they can be near impossible to escape, judging by the loss landscape, and the resulting pose would likely be off to the point where grasping the object would fail. If the weights of the pose regression network just happen to get trapped at one of these local minima it will likely not reach any of the other and more desirable solutions.

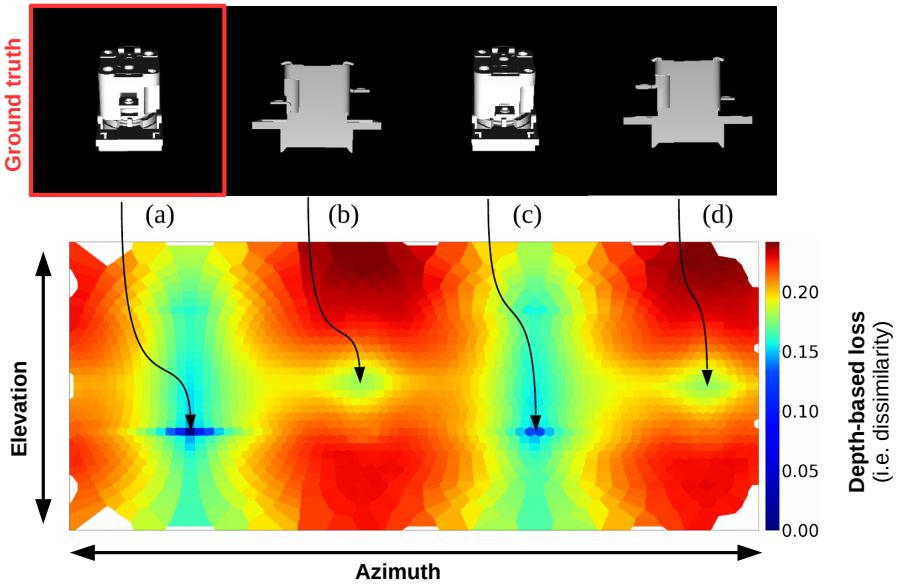


Fig. 4.5: Plot of the loss landscape for a semi-symmetric object. *"(...) Rotations in the image plane are omitted to get a 2D visualisation. The global minimum (ground truth) pose is (a), and its 180 degree semi-symmetry is (c). The two most isolated non-symmetry local minima are given by (b) and (d). The loss landscape is visualized in (e), ignoring in-plane rotations."* [1].
Figure adapted from [1], paper E, © 2021 IEEE.

In paper E, we hence expanded our depth-based loss to produce not one but multiple pose estimates in order to counteract these problematic minima. The solution is illustrated in Figure 4.6, where the pose regression network now outputs n pose estimates, where the difference in depth, $|d|$, is calculated for each in relation to the ground truth. Furthermore, the network also outputs a confidence w_n associated with each of the n poses, which is also used as a weight term when summing up the depth differences across all n poses. However, all poses will contribute with a minimum fixed value of γ to ensure that back-propagation occurs for even zero confidence pose estimates. During inference, this estimated confidence w_n is also important as the final output of the pose regression network will hence be selected as the pose estimate with the largest confidence associated.

Finally, an additional term, $L_{pose}(\hat{P})$, is added to the loss function to ensure that all the n predicted poses are not too similar. If all the predicted poses were allowed to be identical or near identical it would defeat the purpose of introducing multiple pose estimates in the first place. This term in the loss can also be viewed as a way for the pose estimates to repel each other, such that multiple pose estimates do not get stuck in the same local minimum.

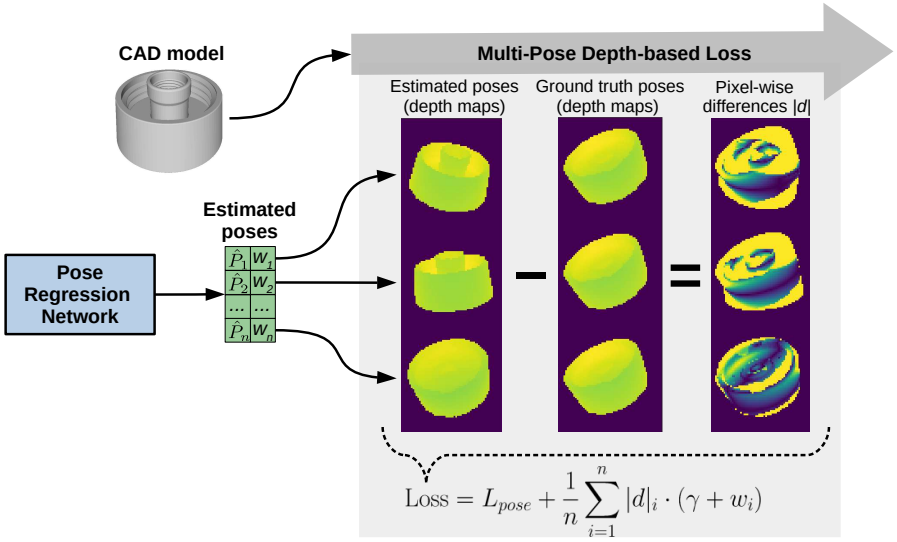


Fig. 4.6: Overview of the depth-based loss function expanded to output n pose estimates to avoid getting stuck in local minima. A confidence, w_n, is estimated by the pose regression network for each of the n pose estimates. The final loss is calculated by summing the depth-based loss for each pose estimate while weighted by its associated confidence, w_n. All poses will as a minimum contribute with γ to the final loss to ensure back-propagation even for zero confidence poses. The term, L_{pose}, ensures that the estimated poses are not too similar. Figure adapted from [1], paper E, © 2021 IEEE, and [5], paper F, © 2022 IEEE.

During evaluation, it was found that having the model output $n = 10$ pose estimates instead of $n = 1$ increased performance noticeably (from 53.10% to 62.34% pose recall), surpassing the codebook-based approach [14] (60.77% pose recall). Introducing these additional pose estimates do impose some computational burden but it is mainly related to the rendering of the depth maps for the loss function. However, this is only required during training and the computational burden during inference is hence negligible. The inference time is $\approx 6.2ms$ per object and it is hence slightly faster than the ≈ 7.0 ms per object measured for the codebook-based approach [14] (both running on an GTX1060). Our approach does hence preserve a fast inference time, which is the main benefit of the codebook-based approach, while at the same time increasing the pose estimation accuracy and reducing the memory footprint by magnitudes (≈ 40 times less memory needed).

5 Shared Pose Regression Network

The work and results presented in paper E relied on training a pose regression network separately for each object, similarly to how a codebook was generated for each object in the codebook-based approach [14]. In paper F we hence expand our approach further by showing that only a single pose regression network is needed instead of training multiple object-specific ones.

This new shared pose regression network relies on the exact same network architecture as paper E with the only exception of expanding the output layer of the network to always produce pose estimates for all the different objects, as illustrated in Figure 4.7. Another important addition to the pipeline for the shared network is a masking scheme which is applied to the output of the network to ensure that only the pose estimates for the correct object ID are left as output.

The size of the network is hence slightly larger due to these multiple outputs when compared to a single network from paper E. However, this slight increase is heavily outweighed by only needing one shared network instead of multiple object-specific ones. In the case of the T-LESS dataset [8], containing 30 different objects, using the shared pose regression network from paper F instead of the multiple networks from paper E would reduce memory usage by $\approx 51\%$ and by $\approx 98\%$ when comparing to the codebook-based approach [14]. The computational burden of applying the masking scheme is also negligible, as reflected by an inference time of $\approx 6.4ms$ per image crop. This is only slightly slower than the $\approx 6.2ms$ achieved by using multiple object-specific networks, making it usable for real-time purposes.

Furthermore, during evaluation of the shared pose regression network it was found that fewer data samples are needed for training the shared pose regression network than for training multiple object-specific ones. In the case of

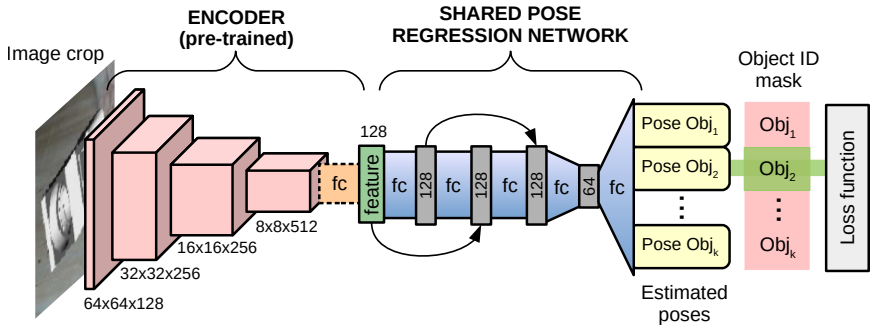


Fig. 4.7: *"The architecture of the shared posed regression network, including the pre-trained encoder used as feature extractor [14]. The network is nearly identical to the original pipeline [1] and includes several skip connections as these were found to increase performance. The exception is the final layer of the network which has been modified to output multi-pose estimates for all the k different object categories, regardless of the class ID of the input. Additionally, a masking scheme is introduced to ensure that only the estimated poses for the correct object ID is propagated further in the pipeline. It is assumed that the object ID is available from a prior detection step."* [5] Figure from [5], paper F, © 2022 IEEE.

the T-LESS dataset, with 30 objects, each object-specific network was trained using 2 million samples per object for a total of 60 million samples. However, the shared pose regression network only required 20 million samples during training to surpass the performance of the multiple object-specific networks (a pose recall of 63.13% versus 62.34%). Finally, the superior performance of the shared pose regression network was found to be largely attributed to fine-tuning parts of the encoder used for feature extraction in the pipeline. Doing a similar fine-tuning step for the multiple object-specific networks is in theory possible. However, such a fine-tuning process would be more complicated due to having multiple networks. It would either require fine-tuning a separate encoder for each network or fine-tuning the same encoder jointly on all the object-specific networks at once. Our proposed pipeline in paper F offers a less complicated pipeline, allowing for easy fine-tuning of the encoder, along with fast inference, a better pose recall and reduced memory consumption.

6 Summary

This chapter presented our work on pose estimation of objects from RGB images. Our initial pipeline proposed in paper E utilizes a novel loss function for dealing with symmetric objects and was used to train multiple pose estimation networks in order to replace object-specific codebooks in a state of the art pose estimation approach. This work was subsequently improved further in paper F by training one single pose estimation network for all objects in-

stead of multiple object-specific ones. The main outcomes and contributions for this part can be summarized as:

- In paper E we improved a state of the art approach by having it rely on multiple small neural networks for doing pose estimation instead of having to rely on multiple codebooks. This improved performance both in terms of pose recall but also other important characteristics, such as reducing the inference time and memory usage.
- A novel loss function for training the neural networks for doing pose estimation was also proposed in paper E. This loss function relies on a differentiable renderer to compare the visual similarity of an object in different poses which allows it to inherently account for any symmetries present in an object. Such symmetries have otherwise proven to be troublesome for learning-based approaches, such as neural networks, as it introduces ambiguities in the ground truth pose.
- The work from paper E was improved further in paper F, where the multiple object-specific networks were replaced with one single neural network. This reduced the memory usage even further while still maintaining the low inference time. Furthermore, using a single network instead of multiple ones allowed parts of the pipeline to be fine-tuned even further which increased the pose recall slightly above using multiple object-specific networks.

The pipelines presented in both papers E and F solely estimate the orientation of objects, and not the full pose including translation. The next logical step for future work would hence be to incorporate the ability to predict the translation of objects as well into the pipeline. The original codebook-based pipeline [14] estimates this translation by comparing the size of the bounding boxes for the detected objects against renderings produced from their respective CAD models. This approach does work but it is also quite sensitive to noise in the detected bounding boxes. A better alternative would hence be to expand this idea with an additional step for fine-tuning the detected bounding boxes prior to using them for estimating the translation [11].

Another obvious option for future work could be to integrate a tracker [2] into the pipeline to filter away erroneous pose estimates as one can likely assume that the pose of an object does not change drastically in a split second. This idea is further supported by our pipeline being able to produce multiple pose estimates within a short time span.

References

[1] S. H. Bengtson, H. Astrom, T. B. Moeslund, E. A. Topp, and V. Krueger, "Pose estimation from RGB images of highly symmetric objects using a novel multi-pose loss and differential rendering," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep. 2021.

[2] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "PoseRBPF: A rao–blackwellized particle filter for 6-d object pose tracking," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1328–1342, Oct. 2021.

[3] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2010.

[4] R. L. Haugaard and A. G. Buch, "Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6749–6758.

[5] S. Hein Bengtson, H. Aastrøm, T. B. Moeslund, E. A. Topp, and V. Krueger, "A shared pose regression network for pose estimation of objects from rgb images," Oct. 2022, accepted at the 16th International Conference on Signal Image Technology and Internet Based Systems.

[6] S. Hinterstoisser, V. Lepetit, N. Rajkumar, and K. Konolige, "Going further with point pair features," in *ECCV*, 2016, pp. 834–848.

[7] T. Hodan, D. Barath, and J. Matas, "EPOS: Estimating 6d pose of objects with symmetries," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[8] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," *WACV*, 2017.

[9] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6d object localization," in *ECCV Workshops*, 2020.

[10] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, *CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation*. Springer International Publishing, 2020, pp. 574–591.

[11] Z. Li, G. Wang, and X. Ji, "CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[12] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[13] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.

[14] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *CVPR*, June 2020.

[15] J. Vidal, C. Lin, and R. Martí, "6d pose estimation using an improved method based on point pair features," in *ICCAR*, 2018, pp. 405–409.

[16] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 119–152, Oct. 1994.

Chapter 5

Conclusion

The work presented in this thesis was conducted as part of the EXOTIC project which is based on the idea of creating a tongue controlled upper limb exoskeleton for individuals with tetraplegia. The main contribution of this thesis was to improve the tongue control of the exoskeleton through a semi-autonomous control scheme based on computer vision. The base assumption being that the addition of this vision-based intelligent control would make it easier to control the upper limb exoskeleton.

A tongue-controlled upper limb exoskeleton was designed and implemented as part of the EXOTIC project in a collaboration with multiple other PhD students. The exoskeleton was designed to include five carefully selected degrees-of-freedom (DoFs) in order to minimize both its weight and size, while still enabling the user to perform the desired tasks. These desired tasks were mainly found to concentrate around eating and drinking, and namely snacking. The tongue-based interface enabled persons with tetraplegia to control the exoskeleton despite being completely paralyzed from the neck and down. Extensive evaluation of this exoskeleton demonstrated that it enabled a person to complete the desired tasks using only their tongue for controlling it.

An initial review of existing literature supported the initial hypothesis of using a computer vision-based semi-autonomous scheme to improve the tongue-based control of the exoskeleton. Our review also revealed a clear tendency to use very clear-cut schemes for arbitrating control between the human and the system, at least in the context of robotic manipulators for assistive purposes. This led to the hypothesis that a semi-autonomous control scheme with a more adaptable level of autonomy would be more suitable for an upper limb exoskeleton, and especially for individuals with tetraplegia. The hypothesis was tested across two studies on computer vision-based semi-autonomous tongue-based control of the EXOTIC upper limb exoskele-

ton. The first study included persons without any paralysis whereas the second study included individuals with impaired or no movement in their right arm and hand. These studies also served to verify the initial base assumption of computer vision-based semi-autonomous control being beneficial for a tongue-controlled upper limb exoskeleton. The result of the studies showed significant improvements when using a vision-based semi-autonomous control over a completely manual approach. Furthermore, the adaptive approach for the semi-autonomous control was in several cases found to result in a significant improvement over the more fixed approach, where the level of autonomy is static. This was especially true for the first study, which also had a better base in terms of proving statistical significance as it included more participants. These observations do suggest that the hypothesis of the adaptive scheme being superior is true in at least some cases. Classical computer vision algorithms were applied in both studies to ensure a stable system with high repeatability to make it easier to test the semi-autonomous control schemes. The used approaches were only applicable due to the controlled nature of the conducted studies.

Research on using computer vision in a less constrained scenario was hence conducted as well. The main focus of this research was pose estimation of objects which could potentially be used in the semi-autonomous tongue-based control of the EXOTIC exoskeleton to infer how to interact with different objects. A pose regression network was proposed and tested based on a state of the art pose estimation approach. This evaluation indicated both an improvement in terms of the ability to correctly estimate the pose of different objects but also in terms of a reduction in both memory usage and inference time. A custom loss function was proposed and used for training this pose regression network which was designed to account for the problem of handling symmetric objects during training. Other approaches would either completely ignore the problem or rely on manually predefined symmetries for each object. The proposed loss inherently handles symmetries by relying on the visual similarity of objects in the different poses and does not require these manually predefined symmetries. This work was expanded further by showing how a single shared pose regression network could replace the need for having multiple object-specific pose regression networks.

The work presented in this thesis also gives rise to plenty of opportunities for future work. The work on the single shared pose regression network could be extended even further by having it estimate the translation of the objects in addition to just their orientation. One approach could be to estimate this translation based on the known size of the CAD models of each object in relation to the size of the object as perceived by the camera. Another promising extension for future work in terms of doing pose estimation of objects could be to include temporal data. Using a tracking scheme it would be possible to

integrate multiple pose estimates in order to increase the quality of the pose estimates by smoothing out noise and discarding erroneous pose estimates. Such an approach is especially promising given the low inference time of our current approach which would allow it to produce multiple estimates within a short time span.

In terms of the tongue-controlled EXOTIC exoskeleton, future work could include developing a hybrid version both in terms of the actual exoskeleton but also in terms of the modalities used for interfacing. The tongue-based control could be supplemented by other ways of interfacing, such as using BCI (brain computer interface) which would enable persons with limited to no control of the tongue to use the system as well. Furthermore, parts of the exoskeleton could be actuated using a tendon-based approach which would likely reduce the size of it.

The proposed semi-autonomous control scheme produced promising results during both of the conducted studies. However, conducting similar studies with even more participants in the future would likely solidify many of these findings. Furthermore, these two studies did also indicate that it would be beneficial to improve upon the existing method for estimating the intention of the user. The implementation used for intent prediction in the conducted studies relied solely on the direction of the end-effector consisting of the user's hand. It would likely be beneficial to partly base this intent prediction on prior observations as well, for instance, the hand often being in a certain orientation and/or position when going for a certain object.

This idea of using prior observations could be achieved by learning from demonstration as part of the semi-autonomous control scheme. The system would hence continually learn what to interact with and how to do it based on observing how a person is using the exoskeleton. The context of an exoskeleton would be perfect for such an approach as the human would be present at all times and available to correct the system. Having to demonstrate how to do various things, in order for the system to learn, would hence be a natural part of using the system and not associated with extra work for the human. This whole idea agrees well with the main idea of both the EXOTIC project and the proposed adaptive semi-autonomous control where input from the user is considered as a valuable resource not to be disregarded. After all, the main essence of the EXOTIC project and hence this thesis was to help people with tetraplegia. The presented work has by no means completely solved all problems that a person with tetraplegia might encount but it has advanced our understanding of how using computer vision for semi-autonomous control of a tongue-controlled exoskeleton can be part of the solution.

Part II

Papers

Paper A

EXOTIC - A Discreet User-Based 5 DoF Upper-Limb Exoskeleton for Individuals with Tetraplegia

Mikkel Thøgersen, Muhammad Ahsan Gull, Frederik Victor Kobbelgaard, Mostafa Mohammadi, Stefan Hein Bengtson, and Lotte N. S. Andreasen Struijk

The paper has been published in the
2020 IEEE 3rd International Conference on Mechatronics, Robotics and Automation, pp. 79–83, 2020.

Abstract

Complete, high spinal cord injuries can lead to a condition known as tetraplegia wherein the body is paralyzed from the neck down. Individuals with tetraplegia are greatly limited in their independence and quality of life. Due to the paralysis, these individuals are bound to a wheelchair and require a high level of assistance throughout their everyday. To enable these individuals to regain some of their lost mobility, exoskeletons holds a great potential. Therefore, this work explores the requirements and design choices that went into creating the EXOTIC exoskeleton, an upper-limb exoskeleton designed for individuals with tetraplegia, which enables the user to drink and eat, while maintaining a relatively small form factor. Finally, the available workspace is simulated and visualized, and a pilot test of the basic functionality shows that picking up an item and transferring it to the mouth takes 41 seconds on average.

1 Introduction

Individuals who have suffered a spinal cord injury may become paralyzed, which entails reduced mobility and, in approximately half of the cases, results in tetraplegia; paralysis from the neck down. This condition greatly limits the independence and quality of life of these individuals [1]. Additionally, tetraplegia results in a constant need for assistance and thus very little privacy. A study by Maheu et al. [2] found that introducing an assistive robotic device could reduce the need for assistance by up to 41%, while increasing the level of independence for individuals with tetraplegia. Yet, fully assistive upper limb exoskeletons are mostly used for rehabilitation purposes [3–6]. Rehabilitation exoskeletons are not necessarily limited in the physical space they occupy, nor the aesthetics they afford, and thus they often lead to bulky exoskeletons that are not fit for assistive applications.

For an upper-limb exoskeleton to be feasible and acceptable to a user, a reduction of the physical dimensions of the current exoskeletons are necessary. Ongoing efforts to reducing bulkiness have taken various approaches: 1) moving actuators towards or beyond the root of the exoskeleton and transferring the actuation forces through tendons [7]; 2) Creating soft-exoskeletons, often with actuation mechanisms removed from the exoskeleton or embedded through pneumatics or similar [7, 8] and finally; 3) Reducing the degrees of freedom in underactuated designs [9].

These opportune design choices each have their merits and flaws. While moving actuators towards the root can greatly reduce both weight and appearance of the extremity of an exoskeleton, it comes at potential overhead in terms of design, as forces must be guided to where the actuation is needed. The transfer of forces, often achieved using Bowden cable transmission de-

signs, leads to some design challenges in terms of friction and control [10]. Others have attempted to transfer the forces through pulley systems, with success in control, but with implications on physical size and appearance [11].

The most opportune design would likely be a soft exoskeleton in terms of physical size and discreteness. Lessard et al. [7] investigated such a system and while they achieve a remarkably small design, such solutions present major problems with regards to control of the arm, as it is difficult to obtain positional feedback from the joints in soft robotics and, thus, it hinders closed-loop control.

Finally, under-actuation has obvious limitations that directly affect the available workspace, however, this approach is taken in most exoskeleton designs, as the human arm has 6 degrees of freedom (not including displacements of the shoulder), which requires equally many actuators and consequently leads to increased bulk. Some exoskeletons have been developed for assistive applications and disabled individuals, and yet, the physical extent remains problematic [5, 12–14] or the assisted degrees of freedom (DoFs) are not suitable to assist individuals with tetraplegia, as these individuals require some wrist and hand actuation for the exoskeleton to be useable [8, 15, 16].

In this paper, we propose a new, compact exoskeleton design that targets individuals with tetraplegia and others with severe disability in the upper limbs. Furthermore, this paper shows preliminary experimentation with the proposed EXOTIC exoskeleton and perspectives and considerations for future iterations.

2 Methods

2.1 User driven design

In the design of the presented exoskeleton, a user-driven approach was taken. Users were involved through interviews and through design games. Five users participated in the investigations [17]. The most important insights found during the investigation were the following [17, 18]: 1) Eating had a high priority. However, this was not confined to eating a whole meal, but rather it was snacking in front of e.g. the television or in the garden. Currently, the users have to ask a helper every time they would like another piece of food; 2) Drinking was found to be just as important as eating, both in a social situation, but likewise to snacking over an extended period, i.e. being able to sip of a drink; 3) The ability to scratch an itch (in the facial region); 4) Turning pages on newspapers and books and being able to grab reading materials themselves; 5) Personal grooming, especially the act of brushing teeth and shaving.

When asked about the physical appearance and size of an exoskeleton for

the arm, the following was found: 1) Three out of five prioritized functions over form, i.e. it was more important to them that an exoskeleton should be versatile and enable many tasks, as compared to the physical appearance of the exoskeleton; 2) Two out of the five where heavily concerned with the appearance of the exoskeleton, the remaining participants likewise expressed their apprehension with respect to appearance. Especially, the exoskeleton should be constructed in such a way that it would minimize further stigmatization; 3) Finally all participants stated that the donning and doffing had to be as simple as possible, as helpers and caregivers must be able to don and doff with relative ease and speed.

In addition to user desires, clinical considerations have to be taken into account for individuals with tetraplegia, as tetraplegia may result in autonomic dysreflexia (an autonomic uncontrolled blood pressure increase that can occur from e.g. irritations/pressure on the skin). As this autonomic response is dangerous and can be lethal in rare cases, care must be taken to ensure that excessive pressure and irritation are diminished.

2.2 Biomechanical considerations

In the research and design of the presented exoskeleton, under-actuation of the shoulder joint was considered, and a biomechanical feasibility study was conducted to ascertain the impact of under-actuating the shoulder [19]. In this research, Casanova et al. [19] tested a 3D-printed exoskeleton model where the shoulder abduction degree of freedom could be restricted. Able-bodied participants were asked to perform a set of tasks: drinking with a straw, eating a chocolate bar and pouring water into a cup while their shoulder abduction was either free to move, constrained to the resting pose (i.e. perpendicular to the ground), or constrained to a 10 degree offset outwards from the resting pose. Their movements were tracked using a motion capture system, and later these motions were analyzed in the biomechanical modelling software AnyBody. From the movements, both relative joint angles and joint reaction forces where analyzed and compared. The results showed that all tasks could be completed with a fixed shoulder abduction angle and that only the water pouring task resulted in exaggerated joint reaction forces at the wrist when shoulder abduction was constrained. Based on these findings, the abduction joint of the shoulder can be omitted, for the most important ADLs requested by the users.

2.3 Design of the EXOTIC exoskeleton

A rigid exoskeleton design with geared motors was chosen as it provides a simple control solution with well-defined joint rotations, as compared to

Fig. A.1: Image of the proposed exoskeleton here shown to successfully move the users hand to grasp a bottle and consecutively move it to the users mouth to sip from the straw.

soft exoskeletons, yet it is possible to achieve a relatively compact design as shown in Fig. A.1.

For anatomical compatibility with most individuals, the exoskeleton is designed to be adjustable in the upper arm link and at the lower braces, which can be adjusted along the axis of the lower arm. Physical stoppers were implemented to limit the joint movement range to be within the range of the corresponding human joint. To prevent potential autonomic dysreflexia reactions, custom-made orthopedic braces where used for the arm rest and wrist rest to provide a distribution of the forces applied to the user, see Fig. A.1.

Two DOF exoskeleton shoulder joint: The shoulder flexion/extension joint consists of a maxon EC-flat type motor (maxon Group AG) with a harmonic drive (Harmonic Drive LLC). The shoulder abduction/adduction joint was omitted in order to reduce the size of the exo as the results from [19] indicated that it could be omitted for the most typical tasks (e.g. drinking, snacking) without incurring significantly different joint torques. Instead, an adjustable joint is used, which can be fixed to a selected angle. This joint is located before the shoulder flexion/extension joint on the exoskeleton mount. To further remove bulk from the shoulder, the external/internal rotation of the upper arm was moved to a half-circular joint, a mechanism similar to [3], [5], but drastically smaller. The joint consists of a half-circular dove-tail guide attached to the shoulder joint and a movable sled onto which the rest of the exoskeleton is attached. The joint is actuated by a maxon EC-4pole motor with a planetary gearing located along the upper arm which actuates the joint through a gear and teeth along the half-circular guide, see Fig. A.1 and Fig. A.2. Previous exoskeletons have used full-ring joints working in a similar manner [12]; however, a full-ring would require that the users arm is

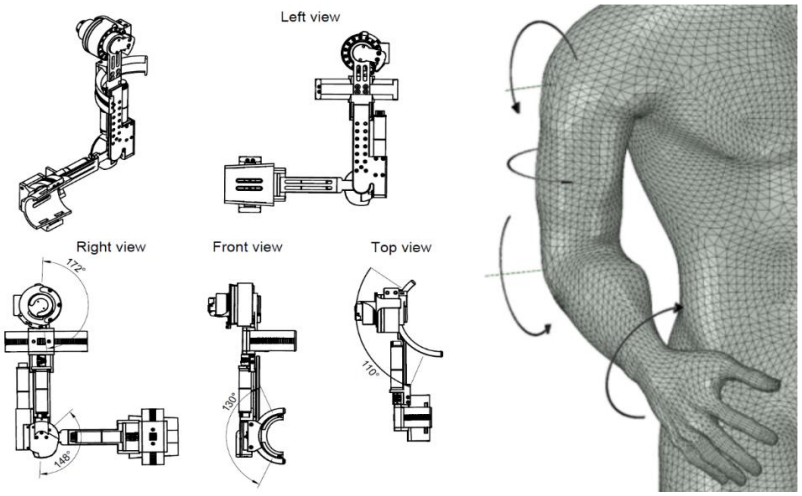


Fig. A.2: Schematic overview of the basic build of the exoskeleton. On the right are schematics of the exoskeleton build together with the range of motion of each joint, and on the left is a 3D rendering of a human arm with depictions of the actuated joints and their axes of rotation.

put through the ring, which would be problematic given that the exoskeleton must be relatively easy to don and doff in order to be used in domestic settings. Instead, a brace carries the upper arm from the exoskeleton and this combination enables donning by only lifting the arm into the exoskeleton.

One DOF exoskeleton elbow joint: The elbow joint is actuated through another EC-4pole motor with a planetary and a worm gear. The lower arm link has another ergonomic brace attached to it, which carries the lower arm, see Fig. A.1.

One DOF exoskeleton wrist joint: To achieve maneuverability of the hand, the wrist joint is actuated through another half-circular joint with a brace attached, allowing for wrist rotations. To ensure that the wrist follows the exoskeleton wrist joint, a single Velcro strip extends from the brace around the brace and wrist, which thereby distributes the force on the ergonomic braces, see Fig. A.1.

One DOF exoskeleton glove: Finally, to facilitate hand closing, a tendon based soft-exoskeleton glove (BioServo Carbonhand) is mounted on the hand. Hand opening is provided using passive elastic bands attached to the wrist brace on rigid guides that extend out to the middle flange.

2.4 Exoskeleton control

All joints incorporate absolute encoders, allowing for direct joint angle feedback. To control the exoskeleton with end effector-based control, a Robot

Operating System (ROS) interface was implemented (ROS Kinetic) using a set of packages: The MoveIt! [20] package was used to provide inverse kinematics and closed-loop trajectory planning and control; The RViz viewer [21] package, was used visualize the state of the exoskeleton in real-time (A visualization of the exoskeleton, viewed in RViz, is visible in the background in Fig. A.1); To add real-time jog control of the exoskeleton, the jog_control [22] package was used and controlled through velocity commands.

To establish the connection from the ROS controllers to the motor drivers (maxon EPOS4 Compact 50/8 CAN) a custom interface was created, which translates the joint position commands from MoveIt! to the appropriate CAN-bus commands, which are sent using a USB to CAN adapter (USB-CAN-SI-M, TITAN Electronics Inc.) enabling update frequencies at approximately 100 Hz. The high-level position commands are fed to the EPOS modules, which implement a PID controller that use incremental encoders as the feedback signal. Further, the incremental encoders enable the use of sinusoidal commutation for smooth motion.

2.5 Analysis and initial testing

A workspace analysis was conducted in which the inverse kinematics were tested for feasible solutions within the maximum reach of the exoskeleton (configured for an average human arm; upper arm length, from shoulder (Acromion) to the elbow (Olecranon) 38 cm and lower arm from elbow (Olecranon) to the wrist (Ulna head) 27.5 cm).

The main load applied to the exoskeleton joints during operation will be applied to the shoulder joint. Therefore, a small strength test was performed by attaching weights, corresponding to the approximate weight of an arm (5 kg) and an additional payload of 1 kg at the end effector, resulting in a combined required torque of 15.6 Nm. While keeping the exoskeleton in a horizontal pose, the load metrics (current and calculated torque) were observed, to verify that they were within nominal values. These measures were extracted through the EPOS module.

To verify that the exoskeleton can perform the tasks that were prioritized by the users, a set of tasks were arranged in a pilot test. The user, an able-bodied co-author of this paper, was seated in a wheelchair in front of a table with the exoskeleton mounted on the right arm and hand. The tasks consisted of the most requested ADLs found in the initial user investigation, namely: drinking and eating. A bottle of water with a straw, a banana and a strawberry were placed in front of the user. Starting with the exoskeleton in a resting pose, the objective was to pick up the objects, one-by-one, transfer them to the mouth and put them back on the table in turn.

As control input for the exoskeleton, a generic gamepad was used to provide Cartesian end-effector control with a maximum speed set to 0.04 m/s.

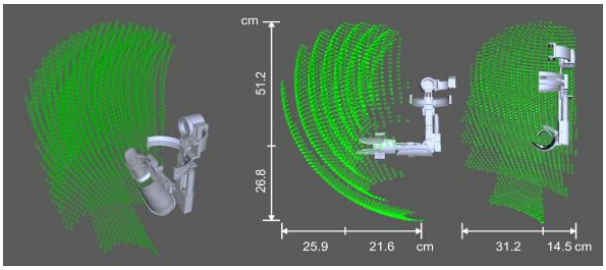


Fig. A.3: Visualization of the valid workspace of the exoskeleton. Green dots correspond to a pose, wherein the end-effector (the approximate position of a hand in the exoskeleton) has a valid inverse kinematics solution. Axes indicate the extent of the workspace. The mid-point on the axes correspond to the position of the end-effector when the exoskeleton is in the pose depicted in grey (the resting pose).

Wrist rotation control and hand closing and opening was enabled through a D-pad also located on the generic gamepad. The exoskeleton was controlled with the left hand, while the right hand and arm was mounted in the exoskeleton. The test was approved by the local ethical committee.

3 Results

3.1 Exoskeleton workspace and load

From a resting pose (upper arm link perpendicular to the ground, lower arm link parallel to ground) the exoskeleton end effector was able to reach 31.2 cm to the right (across the body midline), 25.9 cm forwards, 51.2 cm upwards and 26.8 cm downwards. A visualization of the workspace is shown in Fig. A.3.

Testing of the shoulder motor assembly showed that the exoskeleton could safely apply a torque of 15.6 Nm without exceeding nominal values.

3.2 Pilot testing

Images from the pilot test are shown in Fig A.4. Timings of the tasks are shown in Table A.1. The average time to pick an item up and transfer it to the mouth of the user was 41s. Finally, the workspace allowed for the hand to be moved into close proximity to the head, which might enable the user to scratch itches on the front and side of the head. A video of the exoskeleton attached to a mannequin is available at: `https://www.youtube.com/embed/L-jhidyzWiM?mute=1`

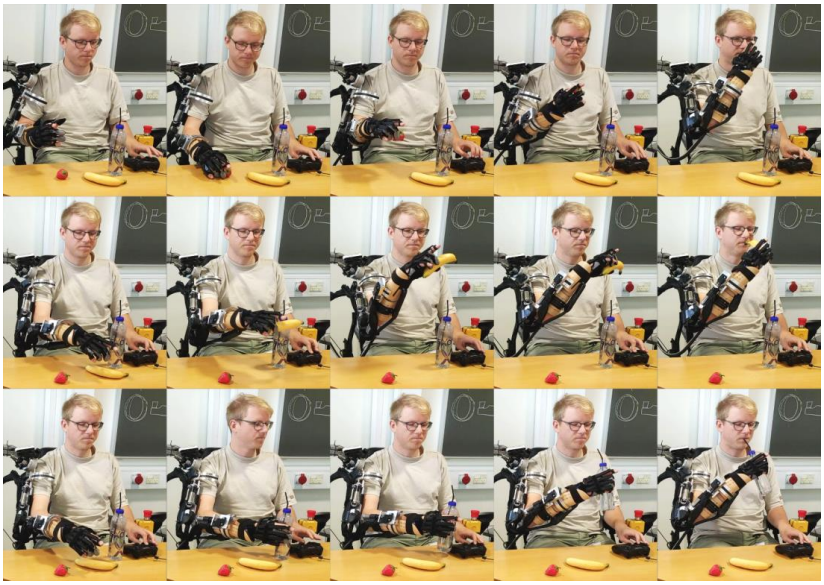


Fig. A.4: The three tasks tested with the exoskeleton. Each row of images shows the progression of each task. From the top: 1) The strawberry task, 2) the banana task, and 3) the bottle task.

4 Conclusion

In this work we have presented a new exoskeleton design with a focus on a small physical structure, while remaining functional for individuals with a severe disability of the upper limb. The exoskeleton has been tested in the lab, showing the ability to solve tasks that were set as the objective for this work by the target user group.

In this paper a gamepad control was used to test the workspace and functionality of the exoskeleton, however, this control modality is not possible to use for individuals with tetraplegia. This paper focused on the exoskeleton, however, future work will focus on the control input available to such user groups and how these control modalities can be optimized to help users in performing their desired tasks.

Table A.1: Timings of the exoskeleton tasks performed.

Task	Time to grasp [s]	Time to mouth [s]
Strawberry task	16	39
Banana task	13	21
Bottle task	20	14

5 Acknowledgements

The authors would like to acknowledge the following professors at Aalborg University for their contributions and guidance: Shaoping Bai, Dept. of Materials and Production; Anne Marie Kanstrup, Dept. of Planning; Thomas Bak, Dept. of Automation and Control; Thomas B. Moeslund, Dept. of Architecture, Design and Media Technology.

References

[1] M. Franceschini, B. D. Clemente, A. Rampello, M. Nora, and L. Spizzichino, "Longitudinal outcome 6 years after spinal cord injury," *Spinal Cord*, vol. 41, no. 5, pp. 280–285, Apr. 2003. [Online]. Available: https://doi.org/10.1038/sj.sc.3101457

[2] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, "Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics*, June 2011, pp. 1–5.

[3] T. Nef, M. Mihelj, G. Kiefer, C. Perndl, R. Muller, and R. Riener, "ARMin - exoskeleton for arm therapy in stroke patients," in *2007 IEEE 10th International Conference on Rehabilitation Robotics*. IEEE, Jun. 2007, pp. 68–74. [Online]. Available: https://doi.org/10.1109/icorr.2007.4428408

[4] B. Kim and A. D. Deshpande, "An upper-body rehabilitation exoskeleton harmony with an anatomical shoulder mechanism: Design, modeling, control, and performance evaluation," *The International Journal of Robotics Research*, vol. 36, no. 4, pp. 414–435, Apr. 2017. [Online]. Available: https://doi.org/10.1177/0278364917706743

[5] R. A. R. C. Gopura, K. Kiguchi, and Y. Li, "SUEFUL-7: A 7dof upper-limb exoskeleton robot with muscle-model-oriented EMG-based control," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2009, pp. 1126–1131. [Online]. Available: https://doi.org/10.1109/iros.2009.5353935

[6] M. A. Gull, S. Bai, and T. Bak, "A review on design of upper limb exoskeletons," *Robotics*, vol. 9, no. 1, p. 16, Mar. 2020.

[7] S. Lessard, P. Pansodtee, A. Robbins, L. B. Baltaxe-Admony, J. M. Trombadore, M. Teodorescu, A. Agogino, and S. Kurniawan, "CRUX: A compliant robotic upper-extremity exosuit for lightweight, portable, multi-joint muscular augmentation," in *2017 International Conference on*

Rehabilitation Robotics (ICORR). IEEE, Jul. 2017. [Online]. Available: https://doi.org/10.1109/icorr.2017.8009482

[8] D. Sasaki, T. Noritsugu, and M. Takaiwa, "Development of active support splint driven by pneumatic soft actuator (ASSIST)," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005, pp. 520–525. [Online]. Available: https://doi.org/10.1109/robot.2005.1570171

[9] T. Otsuka, K. Kawaguchi, H. Kawamoto, and Y. Sankai, "Development of upper-limb type HAL and reaching movement for meal-assistance," in *2011 IEEE International Conference on Robotics and Biomimetics*. IEEE, Dec. 2011, pp. 883–888. [Online]. Available: https://doi.org/10.1109/robio.2011.6181399

[10] E. Brackbill, Y. Mao, S. Agrawal, M. Annapragada, and V. Dubey, "Dynamics and control of a 4-dof wearable cable-driven upper arm exoskeleton," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, May 2009, pp. 2300–2305. [Online]. Available: https://doi.org/10.1109/robot.2009.5152545

[11] J. C. Perry, J. Rosen, and S. Burns, "Upper-limb powered exoskeleton design," *IEEE/ASME Transactions on Mechatronics*, vol. 12, no. 4, pp. 408–417, Aug. 2007. [Online]. Available: https://doi.org/10.1109/tmech.2007.901934

[12] G. Ivanova, S. Bulavintsev, J.-H. Ryu, and J. Poduraev, "Development of an exoskeleton system for elderly and disabled people," in *2011 International Conference on Information Science and Applications*. IEEE, Apr. 2011. [Online]. Available: https://doi.org/10.1109/icisa.2011.5772334

[13] R. Sanchez, E. Wolbrecht, R. Smith, J. Liu, S. Rao, S. Cramer, T. Rahman, J. Bobrow, and D. Reinkensmeyer, "A pneumatic robot for re-training arm movement after stroke: Rationale and mechanical design," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 500–504. [Online]. Available: https://doi.org/10.1109/icorr.2005.1501151

[14] B. Kim, H. In, D.-Y. Lee, and K.-J. Cho, "Development and assessment of a hand assist device: GRIPIT," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, Feb. 2017. [Online]. Available: https://doi.org/10.1186/s12984-017-0223-4

[15] M. H. Rahman, T. K. Ouimet, M. Saad, J. P. Kenne, and P. S. Archambault, "Development and control of a wearable robot

for rehabilitation of elbow and shoulder joint movements," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*. IEEE, Nov. 2010, pp. 1506–1511. [Online]. Available: https://doi.org/10.1109/iecon.2010.5675459

[16] Y. Shimizu, H. Kadone, S. Kubota, K. Suzuki, T. Abe, T. Ueno, Y. Soma, Y. Sankai, Y. Hada, and M. Yamazaki, "Voluntary ambulation by upper limb-triggered HAL® in patients with complete quadri/paraplegia due to chronic spinal cord injury," *Frontiers in Neuroscience*, vol. 11, pp. 1–12, Nov. 2017. [Online]. Available: https://doi.org/10.3389/fnins.2017.00649

[17] F. V. Kobbelgaard, A. M. Kanstrup, and L. N. S. A. Struijk, "Exploring user requirements for an exoskeleton arm insights from a user-centered study with people living with severe paralysis," in *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 2021, pp. 312–320.

[18] F. V. Kobbelgaard, S. Bødker, and A. M. Kanstrup, "Designing a game to explore human artefact ecologies for assistive robotics," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Oct. 2020.

[19] E. Casanova-Batlle, M. de Zee, M. Thøgersen, Y. Tillier, and L. N. Andreasen Struijk, "The impact of an underactuated arm exoskeleton on wrist and elbow kinematics during prioritized activities of daily living," *Journal of Biomechanics*, vol. 139, p. 111137, 2022.

[20] Ioan A. Sucan and Sachin Chitta, "MoveIt." [Online]. Available: https://moveit.ros.org/

[21] R. Haschke, W. Woodall, D. Gossow, D. Hershberger, and J. Faust, "rviz - ROS wiki." [Online]. Available: https://wiki.ros.org/rviz

[22] R. Tajimaaaaa, "jog_control - ROS wiki." [Online]. Available: https://wiki.ros.org/jog_control

Paper B

User Based Development and Test of the EXOTIC Exoskeleton: Empowering Individuals with Tetraplegia Using a Compact, Versatile, 5-DoF Upper Limb Exoskeleton Controlled through Intelligent Semi-Automated Shared Tongue Control

Mikkel Thøgersen, Mostafa Mohammadi, Muhammad Ahsan Gull, Stefan Hein Bengtson, Frederik Victor Kobbelgaard, Bo Bentsen, Benjamin Khan, Kåre Eg Severinsen, Shaoping Bai, Thomas Bak, Thomas B. Moeslund, Anne Marie Kanstrup, and Lotte N. S. Andreasen Struijk

The paper has been published in
Sensors, vol. 22, no. 18, 6919, 2022.

The layout has been revised.

Abstract

This paper presents the EXOTIC - a novel assistive upper limb exoskeleton for individuals with complete functional tetraplegia that provides an unprecedented level of versatility and control. The current literature on exoskeletons mainly focuses on the basic technical aspects of exoskeleton design and control while the context in which these exoskeletons should function is less or not prioritized even though it poses important technical requirements. We considered all sources of design requirements; from the basic technical functions to the real-world practical application. The EXOTIC features: (1) a compact, safe, wheelchair-mountable, easy to don and doff exoskeleton capable of facilitating multiple highly desired activities of daily living for individuals with tetraplegia; (2) a semi-automated computer vision guidance system that can be enabled by the user when relevant; (3) a tongue control interface allowing for full, volitional, and continuous control over all possible motions of the exoskeleton. The EXOTIC was tested on ten able-bodied individuals and three users with tetraplegia caused by spinal cord injury. During the tests the EXOTIC succeeded in fully assisting tasks such as drinking and picking up snacks, even for users with complete functional tetraplegia and the need for a ventilator. The users confirmed the usability of the EXOTIC.

1 Introduction

Each year, between 250,000 and 500,000 individuals worldwide are believed to suffer a spinal cord injury (SCI) [1]. In Northern America alone, SCI cases account for approximately 250,000 individuals [2], of which more than half suffer from tetraplegia [3] with all four limbs being affected. When a high-level complete SCI occurs, the injured individual may not be able to move from the neck down (complete functional tetraplegia). This devastating condition can lead to a loss in quality of life [4] as well as a reduced life expectancy [1]. Further, approximately 22% of individuals with SCI suffer from depression and premature death [5].

Furthermore, individuals with tetraplegia often require constant care from health professionals and a single individual with complete functional tetraplegia may need a team of up to eight caregivers. Meanwhile, human resources from the health care sector are increasingly sparse as the ageing world demography is bound to require more resources within the next decades [6]. This poses an urgent need for solutions that empower individuals with severe disabilities while potentially freeing resources in the health care sector.

An approach to achieve this goal is to apply robotic assistive technologies. A study by Maheu et al. [7] found that introducing an assistive robotic manipulator (ARM) for individuals with upper limb disabilities could reduce the need for help by up to 41%. Consequently, upper arm exoskeletons

hold great potential as assistive devices for individuals with severe paresis or paralysis as a tool to regain some independence [8].

Recently, significant advances have been seen in upper limb exoskeleton (ULE) design [9]. However, generally applicable systems with a critical level of usability that enables domestic assistance of multiple activities of daily living (ADL) for individuals with complete functional tetraplegia are still lacking. To be able to support a user with complete or severe functional tetraplegia in performing multiple ADLs in domestic settings, several critical attributes of the exoskeleton system should be considered. Interdisciplinary development of ULEs including user involvement, involvement of clinicians, biomedical engineers, mechanical engineers, and electrical engineers may help reveal the multifaceted nature of these attributes. Based on our experience from such an interdisciplinary study, we have identified several critical attributes for the domestic use and the potential adoption of ULEs in the everyday life of individuals with complete functional tetraplegia. These attributes are summarized in the onion model shown in Figure B.1. For an assistive ULE to be fully useful as a domestic assistive device, it should support a large variety of motions and ADLs in a manner where the user is always in control and without compromising the social identity, the safety, or the health of the user. Further, the system should be fast to mount and calibrate and it should be mobile.

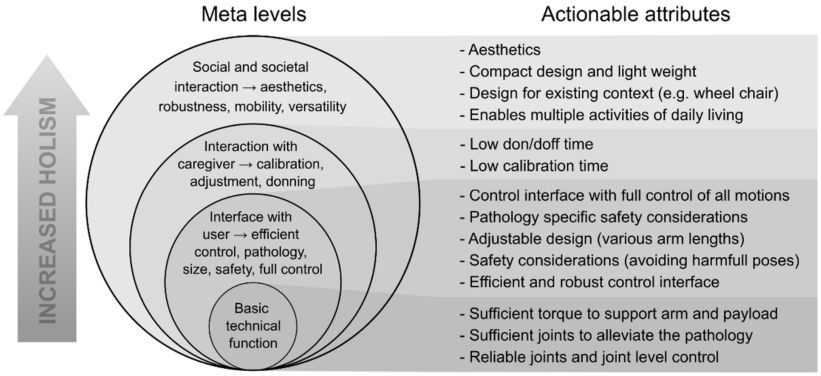


Fig. B.1: Onion model representing the meta levels of design considerations for assistive exoskeleton design. To the right, attributes of each level are specified into actionable attributes. Note: the figure should be read bottom up.

The existing literature on upper arm exoskeletons focuses mostly on exoskeletons for industrial or rehabilitation purposes [10]. Rehabilitation exoskeletons are often advanced high degree of freedom (DoF) exoskeletons that could provide good support in assistive applications, but they mostly suffer from two major obstacles: "bulkiness" and lack of mobility [9, 11–14]. However, there are exceptions such as the Recupera [15, 16], which is mo-

bile and compact but protrudes significantly from the wearers arm. Often, industrial exoskeletons only actuate a subset of DoFs needed to perform ADLs, which renders them incapable of supporting individuals with complete tetraplegia.

A few systems have been developed for physical assistance, especially for individuals with weakened physique [9, 17–19]. However, only three systems have been found which focus on providing assistance and which can potentially support ADL tasks for individuals with tetraplegia: The HAL-UL [20], the "mobile wearable upper-limb exoskeleton" [21], and the NESM exoskeleton [22, 23]. However, the HAL-UL lacks a wrist supination/pronation joint, which means that it can only grab objects that are standing upright, thereby limiting the number of tasks it can perform. Contrary to the HAL-UL, the mobile wearable upper-limb exoskeleton has four DoFs and is compact, but the order and type of joints in the kinematic chain limit the range of motion due to singularities such that if the elbow is flexed, it is impossible to pronate/supinate the wrist and a similar problem is true for the upper arm internal/external rotation. Additionally, both exoskeletons have been created with power assistance in mind rather than complete arm support and they require active movement of the arm or muscle activity to operate them, which individuals with complete functional tetraplegia are not capable of.

The NESM exoskeleton [22, 23] has been designed for individuals with stroke and features five DoFs in the arm and four DoFs in the hand and wrist enabling a good range of motion and full arm assistance. While it can likely perform many ADLs for individuals with tetraplegia, it has a considerable size and weight. Although Crea and Nann et al. [22, 23] showed good results testing it on individuals with stroke, it used a combined encephalography (EEG) and electro-oculography (EOG) brain machine interface (BMI) that, despite its efficient recognition of commands, essentially acted as a start signal to a preprogrammed movement. In addition to being used within rehabilitation, this system has great potential in assistance of users with, e.g., late-stage Amyotrophic Lateral Sclerosis (ALS) or locked-in syndrome, in which the interface options are limited. A similar exoskeleton has been presented by Barsotti et al. [24], but it suffers from the same shortcomings in relation to control interface and is not wheelchair-mountable. Table B.1 summarizes some of the attributes of these exoskeleton systems.

While exoskeletons hold great promise for individuals with severe disabilities, a paradox arises as the disabilities at the same time make it harder for the individuals to operate such assistive technologies. For example, an individual with paralysis or severe paresis in the arms is not able to use, e.g., arm movements, a joystick, or pushbuttons as control input for an exoskeleton.

Currently existing interfaces for individuals with complete functional tetraplegia consist of either chin sticks [25], sip and puff, voice activation, eye-

tracking [26], BMI [22, 23, 27, 28] or combinations thereof [29]. Common for these interfaces is that they exhibit one or more of the following: being indiscreet; being inefficient in terms of the time needed to activate a command; being inflexible in the sense that they only allow for very limited number of command inputs; or they require substantial and or repeated calibration.

To increase the efficiency of assistive devices, autonomous functionality is often applied [30, 31], which can perform some portion of the motions for a given task. Indeed, this can increase performance and accuracy [31, 32]. Kim et al. [31] found that the way in which these functions are invoked is likely of high importance, especially for individuals with physical disabilities. Individuals with SCI found it less satisfying to initiate autonomous execution of a task with an ARM, as opposed to completing the given task manually, due to the reduced sense of agency. This despite a worse performance using the manual control. Hence, autonomous functionality can increase performance, but the way in which it is implemented is of great importance, as users prefer to be in control at all times [31].

Finally, minimal attention has been given to uncovering the opinions, desires, and concerns of potential end-users of the exoskeletons [33, 34]. Therefore, we have previously developed methods to engage users in the design of the exoskeleton arm and conducted interviews and workshops with nine adults living with severe tetraplegia [35, 36]. Complemented by user experiences from experiments with a tongue controlled assistive robot [37], these interviews showed that the users prioritized being able to drink and eat on their own. Not necessarily in the sense of consuming a whole meal but rather during repetitive activities such as eating fruit or candy (snacking) while, e.g., watching television; this to avoid constantly having to ask for assistance to get yet another piece. Another important insight was that the time necessary to don and doff the exoskeleton arm was of high priority as the addition of further time might down-prioritize mounting the exoskeleton and potentially lead to abandonment [34, 38]. Regarding size and function, the interviewees prioritized function but emphasized that a balance between the function and size was paramount [36,38].

Apart from user desires, special attention must be paid to tetraplegia as a medical condition. In particular, individuals with SCI-related tetraplegia are prone to exhibit autonomic dysreflexia (AD); a sudden attack in which the blood pressure rises dramatically and which can be fatal in rare circumstances [39]. AD can be triggered by pressure applied to the skin or excessive skin stretching. In addition, individuals with SCI may suffer from spasticity. These issues may add requirements to safety and the attachment between the exoskeleton and the user.

In this paper, we present the EXOTIC exoskeleton system, see Figure B.2, in which we have strived to include all the considerations and challenges listed in each layer of the model presented in Figure B.1. To evaluate the

Fig. B.2: The EXOTIC exoskeleton system. An individual commands the EXOTIC exoskeleton using a tongue interface shared with computer vision based semi-automation. The individual shown in the image has given a written consent to the use of the image.

proposed system, it was tested with ten able-bodied individuals and three individuals with severe to complete functional tetraplegia. The tests comprised of four ADL tasks inspired by a qualitative investigation of the desires and needs of individuals with severe tetraplegia [38]. The EXOTIC exoskeleton system is the first exoskeleton system empowering users with complete functional tetraplegia to perform arbitrary motions and multiple prioritized ADLs independently and efficiently while continuously being in control of the exoskeleton through intelligent shared tongue control. Table B.1 compares existing ULEs with the proposed EXOTIC exoskeleton system.

2 System Design

2.1 Overview

The EXOTIC exoskeleton system consisted of three main elements: (1) the exoskeleton; (2) a tongue control interface (TCI); and finally, (3) a computer vision guiding system.

2.2 Exoskeleton design

The core design goals for this exoskeleton were to fully assist upper arm motions enabling simple ADLs for users with tetraplegia and at the same time reduce the “bulkiness” and accommodate the user desires. This included a relatively easy donning and doffing and mitigating the risk of provoking AD through ergonomic mounting.

Table B.1: An overview of existing exoskeletons and which attributes they fulfill/include. Green shading indicates that the given attribute was fulfilled/included. Selection criteria: (1) must fully support a paralyzed arm, (2) must implement arm (min. shoulder and elbow) actuation and hand (grab) actuation, (3) must be tangible/tested (i.e., no simulations), (4) must be wheelchair mountable (within reason). Abbreviations: VI = based on visual inspection of images, NA = not announced, NR = not relevant, CFT = complete functional tetraplegia, ADL = activities of daily living. *Powered wheelchairs use 24 V power sources.

Attributes: / Studies:	Enables Multiple ADLs in Individuals w. CFT	Efficient, Robust, and Continuous Control Interface Useable by Individuals w. CFT	Individuals w. CFT Can Control All Motions Fully	Calibration Time	Designed for Existing Context *	Computer Vision Based Semi-Automation	Compact and Light Exoskeleton Design	Aesthetical Concerns (Social Context)	CFT Pathology Specific Safety (AD)	Safety Considerations (Safe Human Operation)	Tests Performed	Adjust-Able Design	Basic Technical Functions
The EXOTIC exoskeleton system for tetraplegia (This study)	Multiple ADLs demonstrated in individuals w. CFT	Intraoral, efficient (0.76 s to start command [36]), robust interface (not affected by environment and commercially available).	Full manual control over all DoFs	Short tongue interface calibration (<1 min)	Wheelchair mountable, wheelchair battery-powered (24 V).	Simultaneously available shared control w. semi-automation based on computer vision and manual control	Compact, Motors parallel to arm, inner rotational joints encircle arm (3.7 kg)	"Invisible" interface, Compact exoskeleton design	Two strap wrist/palm. Open Orthopaedic braces. Loose mounting to avoid AD.	Physical stoppers to limit joints, wheelchair battery operated, current limited, open-brace design (easy to pull out), only moves while being commanded.	Four ADLs on ten able-bodied and three individuals w. CFT.	Adjust-able link lengths	4 arm DoFs, 1 hand DoF
EEG/EOG semi-autonomous exoskeleton for stroke (Nann et al. [22], Crea et al. [23])	One ADL demonstrated in individual w. chronic stroke. (drinking task)	EEG/EOG-based interface (1.43 s to initialize command). EEG is sensitive to multiple factors (EMI, biological)	Semi-automated state-based control only	Longer than EXOTIC (6 min and 18 s calibration time)	Wheelchair mounted. Power requirement NA.	Only operates with semi-automation based on computer vision.	Larger than the EXOTIC system (13 kg)	VI: Visible, protruding interface, larger exoskeleton.	VI: Two strap, arm/hand, open brace design.	Physical stoppers to limit joints, open-brace design (easy to pull out), veto signal (users able to send a stop signal), SEA joints.	One ADL on seven able-bodied and five individuals w. chronic stroke. None w. CFT.	Adjust-able link lengths	5 arm DoFs, 4 hand DoFs
Recupera exoskeleton for stroke rehabilitation (Kirchner et al. [15] Kumar et al. [16]).	No ADLs demonstrated (only exercises).	NR/Relies on residual movement	NR/Relies on residual movement	NR/No relevant user interface to assess.	Wheelchair mounted. Custom 48 V batteries for power	NA/NR	VI: Compact, but protruding significantly from lower arm (4.3 kg)	VI: Compact but protruding. No relevant user interface to assess.	VI: Two strap, upper/lower arm, open brace design.	Physical stoppers, battery operated, current limited. No relevant user interface to assess.	Exercises w. one able-bodied individual one w. chronic stroke. None w. CFT.	Adjust-able arm links	5 arm DoFs, 1 hand DoF
HAL-UL exoskeleton for assisting the elderly (Otsuka et al. [20])	One ADL demonstrated in able-bodied individual. (drinking task)	NR/Relies on residual movement	NR/Relies on residual movement	NR/No relevant user interface to assess.	VI: Likely wheelchair mountable. Power requirement NA.	NA/NR	VI: Compact, Motors parallel to arm, inner rotational joints encircle arm (weight NA)	VI: Compact exoskeleton design. No relevant user interface to assess.	VI: Wrist strap. One open, one closed brace	Physical stoppers. No relevant user interface to assess.	One ADL on an able-bodied individual. None w. CFT.	NA	4 arm DoFs, 1 hand DoF

To remedy the problem of "bulkiness", various strategies can be employed such as: using Bowden-tube cable drives to move the actuation mechanisms towards the base frame or completely off the exoskeleton [40, 41]; using flexible materials with discreet pneumatic actuators or cable drives [42]; or simply by reducing the DoFs [17]. Each of these approaches has its advantages but comes with limitations. Flexible exoskeletons can be compact, to the point where they can be worn underneath clothes [43], and thus could be an ideal solution to the problem of "bulkiness". However, they suffer from non-linear actuation, and challenges in getting positional feedback from the joints make reliable closed-loop control difficult. A reduction of the DoFs has obvious implications on the flexibility of the exoskeleton but provides a simple way to reduce the "bulkiness" at the cost of function. The human arm has seven main DoFs, which would ideally be needed in an exoskeleton arm to achieve a workspace similar to that of the human arm. However, to reduce the "bulkiness" of our exoskeleton, we chose to focus on the gross motions of the arm that are necessary for simple ADLs. These motions can be performed mainly through the shoulder extension/flexion, upper arm internal/external rotation, elbow flexion/extension, and finally the wrist supination/pronation. Whereas the omission of the shoulder abduction/adduction is limiting the flexibility, a previous biomechanical study conducted at our lab [44] investigated the effects of locking the shoulder abduction/adduction joint and showed that the joint was not necessary to perform simple ADLs. Besides the obvious advantage of reducing the "bulkiness" around the shoulder gained from omitting the shoulder adduction/abduction, this also has other advantages. For instance, it ensures that the arm of the user stays within the fixed boundaries of the wheelchair, thus avoiding potential harm to the immediate environment and the user. A passive abduction/adduction shoulder joint was added to afford a more natural pose. This joint was fixed to have an abduction angle of 20°.

To mitigate the medical concern regarding AD, a set of three ergonomic braces were used, of which two were manufactured by an orthopedist (SAHVA A/S, Brøndby, Denmark). As opposed to strapping the arm in, the orthopedic braces "carried" the arm instead, see Figures B.1, B.3 and B.4, akin to how an end-effector arm support supports the arm [45]. This provided a "loose" connection between the human arm and the exoskeleton, allowing the arm to adjust in the exoskeleton, thus reducing the risk of applying excessive pressure or twisting to the skin of the user, which can cause AD. Additionally, this ensured a relatively fast donning and doffing during which the users arm was simply lifted into the three ergonomic braces. Two straps were used at the wrist and palm to secure the position of the hand relative to the exoskeleton.

The exoskeleton frame was created using a custom fabricated aluminum (7075) frame with variable link lengths to accommodate different arm lengths.

To actuate the upper arm rotational joint, elbow flexion/extension joint, and the wrist pronation/supination joints, three motors with planetary gears were fitted (EC-4pole, Maxon motor ag, CH), see Figure B.3. Actuation of the shoulder flexion/extension was realized through a more powerful motor (EC-i40, Maxon motor ag, CH) and a strain wave gear (Harmonic Drive LLC, MA, USA). The shoulder and elbow flexion/extension joints were actuated directly on the joint axis, whereas the rotational joints located on the axis of the arm were actuated through two dovetail, half-circular ring designs with the actuator actuating a gear that turned a semi-circular teeth-set [46], see Figure B.2 and Figure B.4.

Table B.2: DenavitHartenberg parameters of the EXOTIC exoskeleton. Parameters L_u and L_f correspond to the lengths of the upper arm and forearm, respectively, which are both adjustable.

Link	a_i	α_i	d_i	θ_i
1	0	$\pi/2$	0	$\pi/2 - \theta_1$
2	0	$\pi/2$	L_u	$\pi + \theta_2$
3	0	$-\pi/2$	0	θ_3
4	0	0	L_f	θ_4

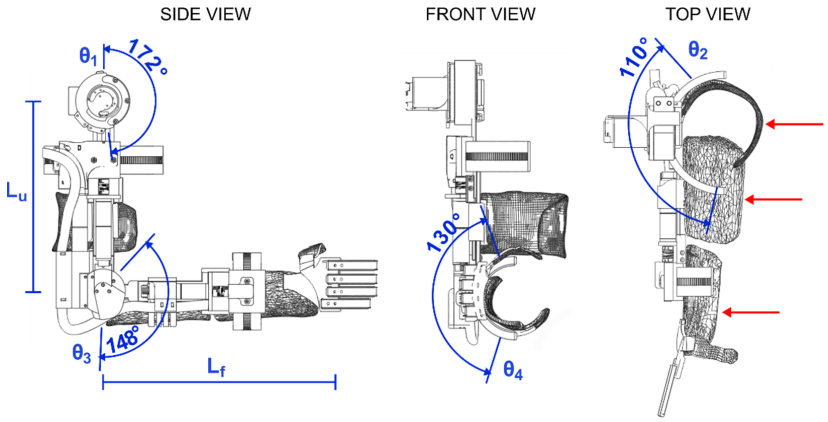


Fig. B.3: Detail drawings of the EXOTIC exoskeleton and the range of motion of the four main degrees of freedom contributing to gross motion. From left: side view with shoulder (θ_1) and elbow (θ_3) flexion/extension and indications of link lengths (L_u and L_f), which are adjustable; front view with wrist (θ_4) supination/pronation; top view with upper arm internal/external rotation (θ_2) and arrows that indicate the three braces used to support the arm. Notation corresponds to the Denavit-Hartenberg parameters given in Table B.2.

The joints were practically non-backdrivable except for the shoulder flexion/extension. However, this could only be backdriven under heavy load. The exoskeleton could be mounted onto wheelchairs and was designed to be powered by regular wheelchair batteries (24V). Perspective 3D renders of

the exoskeleton and a simulated workspace are shown in Figure B.4. The Denavit-Hartenberg parameters of the four main DoFs of the exoskeleton are shown in Table B.2 and visualized in Figure B.3. While the focus of this paper is primarily on the overall user-based system design and particularly on the experimental side of our five DoF intelligently tongue controlled exoskeleton system as in [20, 22], kinematics and modelling such as in [47, 48], are available for a similar exoskeleton in Gull et al. [46].

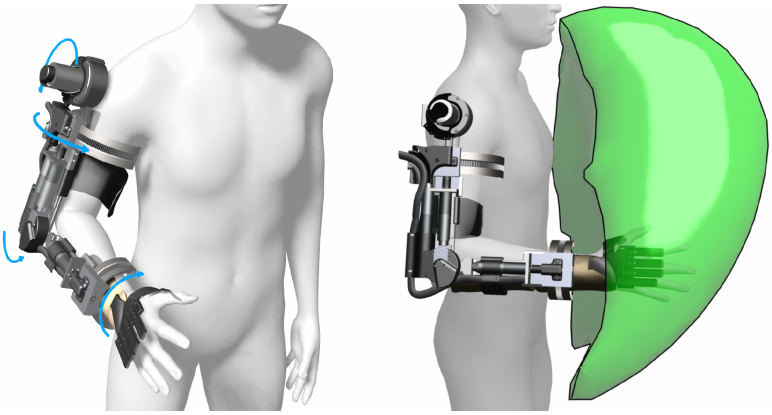


Fig. B.4: Threedimensional renders of the exoskeleton with light blue arrows indicating the rotational joints and the available workspace indicated by the green volume on the right. Simulation was performed by permutating the joint angles over the range of motion of each joint and recording the hand location. The visualization was created from an approximation of the bounding volume of the resulting point cloud.

Each motor was fitted with incremental encoders (>500 counts per rotation) for accurate motor control with sinusoidal commutation. Joint-level angular control was provided through miniature magnetic absolute encoders (RLS Merilna tehnika d.o.o., Sl) added directly to the shoulder and elbow flexion/extension joint axes. As direct attachment of encoders to the joint axis of the internal/external shoulder rotation and the wrist pronation/supination was not possible due to the joint axis being internal to the arm of the wearer, a custom gearing was used on these encoders to extract the joint angles. Detail drawings of the exoskeleton frame are shown in Figure B.3.

The tendon-based soft-exoskeleton glove, CarbonHand (Carbonhand®, Bioservo Technologies AB, SE), was used to provide one DoF for grasping objects. As the CarbonHand glove only enabled grasping and not actively released the grasp, a set of fabric elastic bands were used to counteract the grasping of the glove such that when the tendons were relaxed, the elastic bands would pull the hand back into an open pose.

2.3 Control interface

Of the available interfaces for individuals with tetraplegia, chin sticks are reliable control inputs but are limited in the number of available commands and unattractive due to their indiscretion; if the users are even able to command them. Similarly, sip and puff systems are reliable but have a limited number of command inputs and are indiscrete. Conversely, eye-tracking can accommodate many simultaneous commands and can be useful when no other options are available, but it occupies the users gaze and attention and moreover the reliability can be a challenge [25]. Lastly, BMIs are particularly important for users who are paralyzed throughout the body as seen in the late stages of ALS. BMIs have recently shown promising results for exoskeleton control in laboratory settings [27]. However, BMIs often require substantial calibration, may be invasive [27], and are still too complicated to render applicable in everyday domestic settings for the continuous control of many DoFs. However, state-based control of exoskeletons using BMIs has recently shown promising results in assistive applications [22, 28], even outside the laboratory [28].

To address the problem of providing a reliable interface for individuals with tetraplegia, research into user interface technology is ongoing and has produced the intraoral tongue control interface, which was originally developed in our group [49, 50], see Figure B.5.

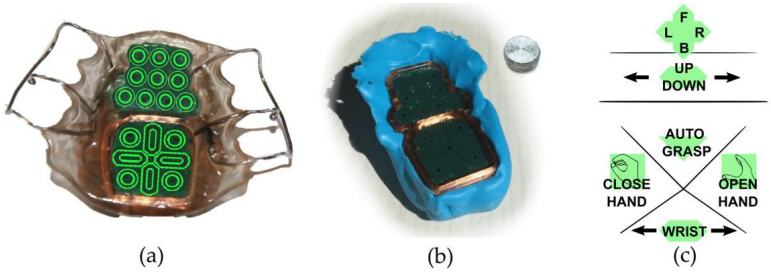


Fig. B.5: The iTongue tongue interface. (a) A standard commercial iTongue mouthpiece with sensor placements superimposed, (b) a temporary silicon palate brace as used in the current study together with an activation unit, and finally, (c) the control layout developed in this study to control the exoskeleton with shared computer vision based semiautomation. The topmost part acts as a joystick to control the position of the exoskeleton in a horizontal plane, while the up/down slider controls the vertical axis. The wrist supination/pronation is controlled through the bottommost slider. The remaining three icons on the interface act as simple pushandhold buttons. Note: the text size in the layout has been increased in the figure to increase readability.

As the name suggests, this interface allows the user to control technology through movements of the tongue. In particular, the tongue control interface has proven to be powerful allowing individuals with tetraplegia to control a wide range of assistive solutions, including robots [51–55]. The intraoral tongue control system consists of an inductive sensing device embedded in a

palate brace that works in conjunction with an activation unit attached to the tongue. The inductive sensing area detects the position of the activation unit, which corresponds to the position of the tongue. This allows the interface to have several simultaneously available command inputs. In effect, this can be used for direct control of several DoFs. The control takes place without the need for visual feedback in the form of, e.g., a screen as the user is able to feel where the tongue is, given that the user is trained in using the system.

The intraoral control interface used in this study was an adapted version of the commercially available CE-certified wireless iTongue device (iTongue®, TKS A/S, Nibe, DK) [49, 50] as it enables an invisible and unintrusive way to control the exoskeleton, see Figure B.6.

The mouthpiece of the iTongue is self-contained with a battery, onboard processing, and a wireless radio allowing it to be completely concealed in the mouth. It features 18 sensor areas that can be programmed to act as individual buttons or be interpolated to create larger virtual buttons and joysticks [37], see Figure B.5 for a visual representation of the interface and its sensors and layout. The wireless signals are picked up using a central unit which conveys the signals to a computer for further processing.

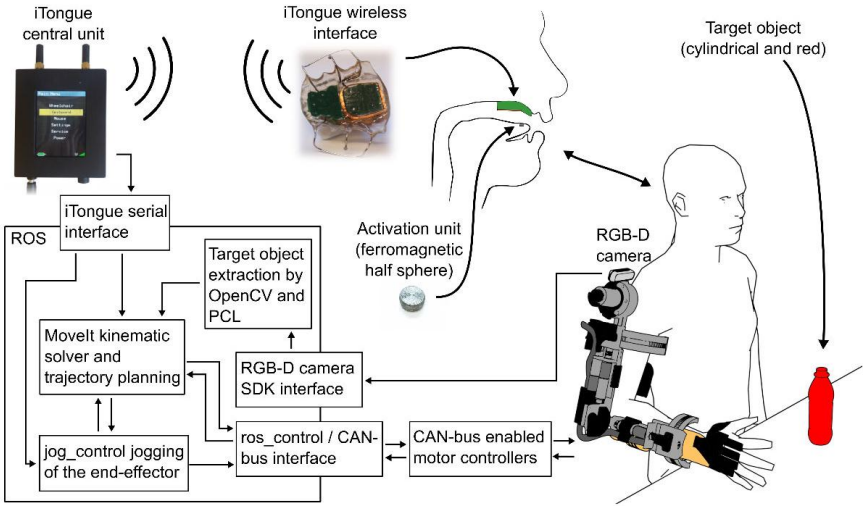


Fig. B.6: System overview. From the right bottom: The exoskeleton communicates via CANbus with a computer. From here all sensors of the exoskeleton are read and all motors are controlled. The RGBD camera on the shoulder feeds its images through a computer vision pipeline resulting in a relative position and orientation of a target object (red bottle) enabling autonomous control. The user commands the exoskeleton through the adapted iTongue mouth unit. The wireless signals are received by the iTongue central unit, which in turn sends the received commands through a serial connection to a PC running the ROS ecosystem. Linear and square arrows indicate communication, double arrows indicate bidirectional communication. Note: the exoskeleton glove used in this study is not shown in this overview, but it is controlled through a USB serial interface.

2.4 Computer vision-based shared control system

Whereas the iTongue tongue control system is a versatile control method, it is a two-dimensional control interface meaning that a maximum of two degrees of freedom (DoF) can be controlled intuitively at the same time, for example for controlling an end-effector in a 2D plane. By adding intelligence to the control through intention prediction and spatial awareness, the same interface can be used to enable simultaneous control of more degrees of freedom through shared semi-automation. This approach has previously been applied for ARMs [30–32] and recently in an exoskeleton glove using different interfacing methods [22]. In effect, this allows a single button to control an exoskeleton with an arbitrary number of DoFs to guide it towards an object of interest.

In some previous systems [22, 30, 31], autonomous functionality was integrated with a "point and click" approach, e.g., pointing to an object of interest, clicking, and letting the robot/exoskeleton perform the desired action without any user opportunity to stop the system, potentially causing harm [56, 57]. However, this approach has two caveats: the individual operating the device may feel that it is acting on its own accord and thus may feel distanced from it rather than identifying with the device [31]. In the case of the EXOTIC, this point is particularly important as the sense of empowerment gained from performing and achieving with the exoskeleton may in turn be paramount to avoid abandonment of the technology [38]. Secondly, if the control algorithms command the exoskeleton to perform unintended movements or the movement is not entirely correct, the "point and click" method may mean that the device cannot be stopped before reaching its destination. The optimal strategy seems to be a shared control, in which the user is always in control while autonomy assists in performing the movements [30, 31].

To overcome these caveats, the intelligent control system presented here acted as a "push and hold" function, such the user was to command the exoskeleton to move continuously until the user deemed the motion complete or wanted to stop the motion in case the exoskeleton was not moving as intended. If the latter was the case, the user could manually correct the motion using direct manual control. Not only did this ensure that the user was always in control, but it also increased the safety considerably [56, 57].

To enable spatial awareness and semi-autonomous control, a color and depth camera (Intel RealSense D415, Intel Corporation, CA, USA) was added to the EXOTIC exoskeleton system. The camera was placed above the shoulder joint, see Figure B.6. The camera pointed towards the workspace in front of the exoskeleton enabling the use of computer vision algorithms to locate the position and orientation of objects of interest. An overview of the system is shown in Figure B.6.

3 Methods

3.1 Exoskeleton control

The actuators located at each joint on the exoskeleton were controlled through CAN-bus enabled motor controllers (EPOS4, Maxon motor ag, CH). Each motor controller was connected via a CAN-bus USB interface (USB-CAN-SI-M, TITAN Electronics Inc., TW) allowing the exoskeleton to be controlled from a PC, see Figure B.6.

The encoders on the joints were connected to the motor controller modules allowing the use of a built-in PID control (autotuned). Communications were handled through the CAN-bus network management (NMT) service setting the motor controller modules as NMT Slaves, while a computer with the USB to the CAN-bus dongle acted as NMT Master. This setup allowed the system to update motor targets and read sensors at 100 Hz simultaneously for all motors.

To command the exoskeleton efficiently and continuously, a custom software interface was created to bridge the CAN-bus USB interface to the Robot Operating System (ROS), which was used as the main ecosystem for control. A forward kinematic model of the exoskeleton was defined and the MoveIt package [58], which relies on the Orocos Kinematics and Dynamics Library [59], was used to manage inverse kinematics and trajectory planning, thus enabling end-effector control. Live end-effector control of the exoskeleton was facilitated via the jog_control package [60]. See Figure B.6 for a system overview.

3.2 Intelligent control

For the intelligent control to work, the essential task was to detect and determine the position and orientation of an object of interest in relation to the exoskeleton. To find the objects of interest, a computer vision algorithm was used to process the combined color and depth camera feed. The algorithm started by performing an initial thresholding operation in the HSV color space to extract red hues in the image, as we deliberately choose red objects as the objects of interest. By having the depth and color information aligned, the resulting mask of the red objects obtained from the thresholding operation could be applied to the depth map, and thus the depth maps of red objects were isolated. Subsequently, the masked depth map was converted into point clouds, thereby forming point clouds for each of the red objects in the original image. Each of these point clouds were evaluated using a random sample consensus (RANSAC) approach for their resemblance to a cylinder shape. The computer vision algorithm is illustrated in Figure B.7.

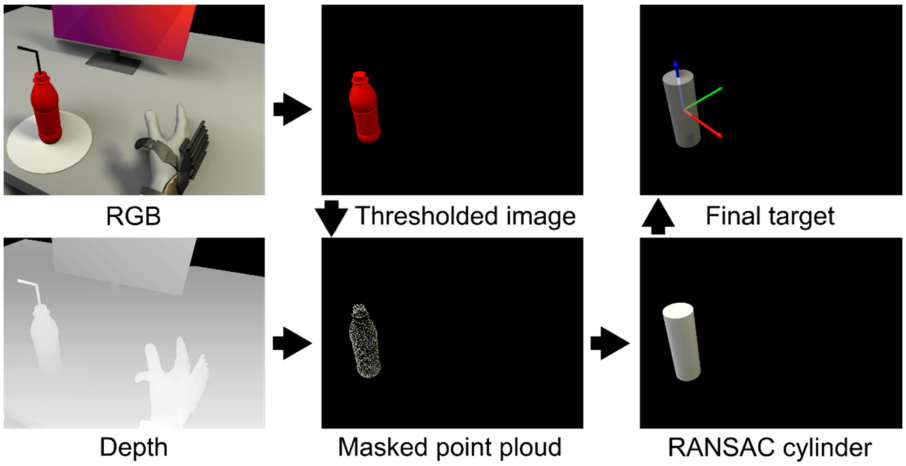


Fig. B.7: Image processing pipeline. A threshold is applied to the RGB image (**top left**), which results in a mask of all the red objects in the image (in this case the bottle, **top middle**). This mask is used to isolate the same objects in the point cloud (**bottom middle**), which is obtained from the depth image (**bottom left**). A random sample consensus method is performed to find the best fitting cylinder shape on the extracted point cloud object(s) (**bottom right**). The center of the found cylinder(s) is determined to be the target position (**top right**).

Once the position and orientation of the object were known, the trajectory could be calculated and the motion to go from the current pose of the exoskeleton to a grasp pose around the object of interest could be performed. The trajectory from the current pose of the exoskeleton was planned continuously through the jog_package. The software packages used to extract the target object from the image and depth feed were OpenCV for thresholding and reading the incoming camera data and the Point Cloud Library [61] was used to perform RANSAC [62]. The computer vision algorithm and control method is described in greater detail in Bengtson et al. [63] (referred to as the "Fixed Semi-autonomous Control" scheme).

3.3 Tongue control interface adaptations

Based on previous experiences and studies on optimizing the layout of the tongue interface [37, 64, 65], a weighted average of neighboring sensor approach was used to create a continuous sensing surface similar to a touchpad on a laptop. In a previous study from our group, this approach was found to achieve a throughput of 0.73 bits per second [37]. Additionally, a dwell time was used to prevent accidental activation during, e.g., speaking as found in a previous study [66].

Similar to Mohammadi et al. [65], a virtual joystick was implemented on the front surface of the mouth piece to control the position of the hand in a

2D plane (forwards/backwards and left/right), see Figure B.5c. Beneath the virtual joystick, a 1D virtual joystick controlled up/down and the rear surface had buttons for: opening and closing the hand; activating the autonomous control; and yet another 1D joystick to control the wrist rotation. This provided the user with direct, manual, and continuous control of the movements of the exoskeleton. The auto grasp button activated the intelligent control for approaching an object as long as the button was activated.

3.4 Test of the EXOTIC exoskeleton

To test how the exoskeleton worked in activities of daily living, two studies were conducted. One five-day study was conducted with able-bodied individuals, and another three-day study was performed with users with tetraplegia.

Participants

Able-bodied individuals

Ten able-bodied individuals (one female, mean age: 26.3 $\pm$ 4.8) were recruited for this study. The main exclusion criteria were severe right arm injuries, cognitive impairments, and drug addiction. Able-bodied participants were reimbursed for their time with DKK 100 per hour (equivalent to EUR 13.45 per hour) subject to income tax.

Individuals with tetraplegia (Users)

Three individuals with tetraplegia (all male, mean age: 44.7 $\pm$ 19.1) were recruited for the study. Main inclusion criteria for individuals with tetraplegia were: 1) between 18 and 75 years of age; 2) reduced or absent motor function in the right arm (tetraplegia) caused by a spinal cord injury (ASIA Impairment scale score of grade A to D) or ALS; 3) not able to repeatedly use the right hand and arm to grab a bottle with a straw (300g) from a table and drink from it while in a seated position, and finally; 4) at least have some tongue functionality.

User 1 was able to flex and extend the elbow against gravity but had little to no function in wrist and fingers (ASIA: D, C4 injury), user 2 had a good arm function but little to no function of the fingers and wrist (ASIA: A, C5 injury). User 3 had no functional use of the upper limbs and depended on a ventilator (ASIA: C, C2 injury). Thus, this user had complete functional tetraplegia. INCSCI assessment results for each individual are available in Table B.3.

Experiment description and setup

Able-bodied individuals

The able-bodied individuals attended a five-day study comprising a series of training sessions and experiments with the EXOTIC exoskeleton system. The data shown in this paper were assessed at the end of the five-day study. The five-day study was split into two segments: three consecutive days followed by two consecutive follow-up days approximately one month later. Each session consisted of approximately two hours of focused use of the EXOTIC exoskeleton system. The initial three days tested the manual tongue control of the exoskeleton with two different tongue control methods [67], whereas the fourth and fifth sessions added semi-autonomy. The data presented in this article represent the performance of the able-bodied participants at the end of the final session (session 5).

The able-bodied participants were instructed to relax their arm and hand as much as possible during the experiments to simulate a paralyzed limb. To verify their cooperation with this request, eight surface electromyography (sEMG) electrodes (Myo Armband, Thalmic Labs Inc., CA, 20132018) were mounted on the arm on the transverse line between the medial acromion and the fossa cubit at 1/3 of the distance from the fossa cubit (approximately the peak of the biceps muscle). An sEMG recording of the max contraction of the participant was collected before mounting the exoskeleton. From this recording, a threshold was determined as 1/5 of the maximum contraction, which, if passed, would give the experimenter a warning during the tests such that the experimenter could remind the participant to relax the muscles.

Individuals with tetraplegia (Users)

The experiment with the users comprised of three sessions in total. The first session consisted of tongue controlling a simulation of the exoskeleton running on a computer. The extent of the spinal cord injuries for each user was determined through an ISNCSCI assessment performed by a trained medical doctor. In the second session, the users trained using the exoskeleton. Finally, semi-autonomous control was added to the experiment in the third

Table B.3: ISNCSCI assessment of users. Summary ratings of the ISNCSCI assessments for each user who participated in the study. * Indicates that the user used a ventilator.

	Age (Years since Injury)	Neurological Levels				Neurological Level of Injury	Complete/ Incomplete	ASIA Impairment Scale	Zone of Partial Preservation			
		Sensory		Motor					Sensory		Motor	
User		L	R	L	R				L	R	L	R
1	59 (0.6)	C4	C4	C5	C4	C4	I	D	NA	NA	NA	NA
2	52 (32)	C5	C5	C6	C7	C5	C	A	T11	L3	S1	S1
3 *	23 (0.7)	C2	C3	C2	C3	C2	I	C	NA	NA	L3	L3

session. The data presented here are from the last part of the final session. The first and second session consisted of approximately 1½ hours and the last session approximately 2½ hours of focused exoskeleton control. Following the tests, a short semi-structured interview was conducted with the users to obtain feedback from and opinions on their experience with the EXOTIC exoskeleton system. The interviews were recorded and later transcribed.

Experiment setup

The commercial version of the iTongue tongue control interface has the active elements embedded in an acrylic palate brace with custom fitted prongs that secure it to the teeth. However, in this study a dental two-component A-silicone putty (Top Dent ImpressA Putty Soft, DAB Dental AB, SE) was used to create a temporary palate brace for the users to accommodate reuse of the system, see Figure B.5b. Furthermore, a temporary activation unit was glued to the tongue of the participants as opposed to the medically inserted activation unit used for the iTongue commercial interface. The temporary activation unit was a 5mm titanium sphere with flattened top and bottom, see Figure B.5b. The unit was glued near the tip of the tongue using a surgical skin adhesive (Histoacryl® B. Braun Surgical S.A., Rubí (Barcelona), Spain).

For both users and able-bodied individuals, the custom silicone mouthpiece was molded during the first session. It was created by pressing the two-component silicone putty containing the tongue interface up against the palate of the individuals until the putty solidified (approximately 2 mins.). Before donning the exoskeleton, the exoskeleton links were adjusted to match the body size of the wearer. Likewise, an appropriate CarbonHand exoskeleton glove size was used for each individual. Additionally, to avoid errors related to slightly different hand mounting on the exoskeleton, the hand position was calibrated by moving it to a grasping pose on an already tracked object of interest. Knowing the position of the object, the position of the hand could be corrected in the kinematic chain such that when activating the autonomous function there would not be any offset issues. The activation unit was glued to the tongue as the last step before commencing the experiment.

Each participant was seated in a chair, wheelchair, or powered wheelchair in front of a height adjusted table in such way that when the exoskeleton was mounted correctly, the wrist of the exoskeleton would be directly above the table edge, see Figure B.8. A face shield was mounted to the head of the participant due to safety concerns.

Experiment tasks

The four ADL tasks used to evaluate the performance of the EXOTIC exoskeleton system for both able-bodied individuals and individuals with tetraple-

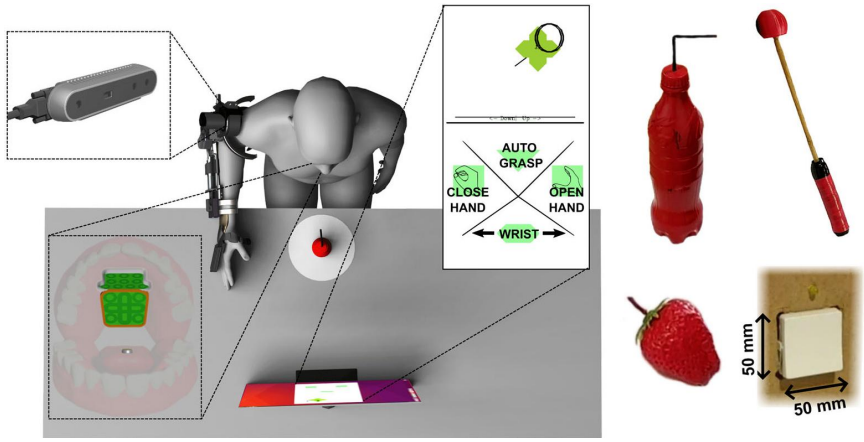


Fig. B.8: Experimental setup overview. The participant was positioned in front of a table with a bottle positioned 10 cm away from the table front. The iTongue system was mounted at the palate of the participant and the activation unit was glued to the tongue. A screen on the table showed dynamic visual feedback of the control layout and the position of the activation unit on the control layout. The objects used for ADL tasks are pictured on the right. From the top left: the bottle, the scratch stick, and the strawberry.

gia were: (1) The bottle task: grabbing a bottle with a straw from a table and moving it towards the face to make the straw touch the face shield (straw was 10 cm long above the lid, straight for able-bodied individuals and with a 90°, bend at 5 cm for individuals with tetraplegia to accommodate a more slanted sitting posture); (2) The strawberry task: grabbing a strawberry from a table and moving it to the face shield; (3) The scratch stick task: picking up a mock-up scratch stick from the table and moving it to make the end touch the side of the face or the face shield, and finally; (4) The switch task: depressing a standard Danish wall outlet switch (LK Fuga®, wall outlet switch 542D6001, 50 x 50 mm rocker switch surface) mounted to the right of the participant. As shown in Figure B.8, the bottle was upright, whereas the strawberry and scratch stick were lying on the table. The end of the scratch stick was a ball made from a soft material in order not to cause any harm in case of errors. The strawberry was an artificial plastic strawberry. The three grasp-and-transfer task objects were placed at a marked spot 10 cm from the table front centered in front of the user. A screen showing dynamic visual feedback from the tongue control interface was placed approximately 50 cm from the table front, see Figure B.8. The switch was located 35 cm above the table, 10 cm from the table front and 5 cm to the right of the exoskeleton shoulder.

Each task was performed three times and the averages of the three trials were used to obtain the results presented here. During the trials the exoskeleton was configured to move at 4.5 cm/s and was configured to start at a pre-

defined rest position, see Figure B.8. If an object was dropped during a trial performed by able-bodied individuals, the trial was restarted. However, during trials performed by users, the objects were picked up by an experimenter and replaced in the hand while the user continued as if the object was still held in the hand. The reason for this discrepancy between the able-bodied individuals and users was more frequent drops and a higher risk of fatiguing the users in case of restarting the tasks too many times. The time for these tasks was recorded for the grasping part of the task as measured from the first command issued on the iTongue until the "close hand" command. The transferring to the face part of the task was measured from the "close hand" command until the task was completed. For the three grasp-and-transfer tasks, all participants were instructed to use the semi-autonomous control.

4 Results

All experimental participants were able to control the EXOTIC exoskeleton system directly and continuously to perform the desired ADLs; even in the case of complete functional tetraplegia and the use of a ventilator (user 3).

The able-bodied individuals were able to grab the bottle and move it to the face shield in 38.7 $\pm$ 6.1 seconds on average. The three users managed to perform the same task in 55.4 $\pm$ 8.0 seconds on average. When performing the increasingly difficult strawberry task requiring the wrist to be pronated while grabbing the strawberry and requiring the wrist to be supinated when reaching the face shield, the able-bodied individuals managed to perform the entire task in 62.7 $\pm$ 8.54 seconds on average. The same task was completed in 92.3 $\pm$ 11.6 seconds on average by the three users. The able-bodied individuals were able to perform the scratch stick task in 70.3 $\pm$ 12.0 seconds on average. The three users performed the same task in 106.7 $\pm$ 16.9 seconds on average. Finally, the switch task took 34.30 $\pm$ 10.78 seconds on average for the able-bodied individuals, while it took individuals with tetraplegia 41.39 $\pm$ 9.47 seconds on average. These results along with the number of issued commands are shown in Table B.4 and as scatter and box plots in Figure B.9 and Figure B.10.

During the four ADL tasks for the users, the soft exoskeleton glove could sometimes not apply sufficient force due to the added elastic bands which led to dropped objects. In these cases, the users were asked to continue as if the object was still in their hand. The objects were then placed back in the hand by an experimenter while continuing the motion of the exoskeleton, and if needed the object was supported by the experimenter until the task was completed. This was only observed during the user tests and not during the tests with the able-bodied individuals. As the exoskeleton glove was not the focus of this work, this was not considered for performance metrics.

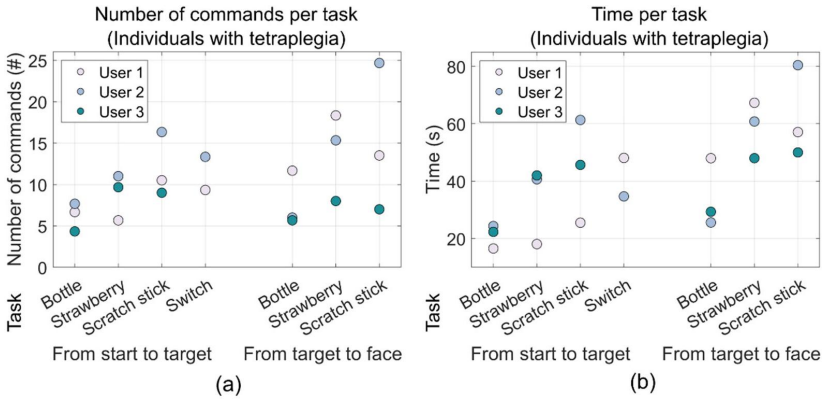


Fig. B.9: Scatter plots of the mean performance metrics for the individuals with tetraplegia (the users). The two groupings in each plot show the results from the reaching phase and the moving object phase, respectively, and each column corresponds to each task: bottle, strawberry, scratch stick, and switch, respectively. (**a**) Shows the number of commands used during each phase of each task while (**b**) shows the time it took to complete each phase of each task. The data for the switch task for one of the users (user 3) was unfortunately lost.

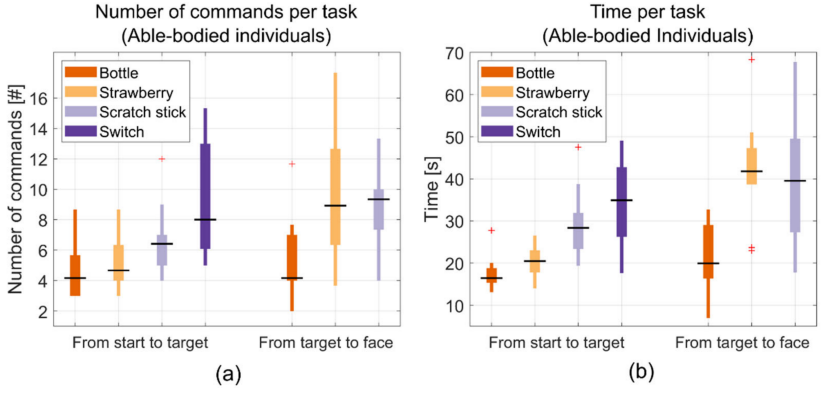


Fig. B.10: Boxplots of the mean performance metrics for ablebodied individuals. The two groupings in each plot show the results the reaching phase and the moving object phase, respectively, and each column corresponds to each task: bottle, strawberry, scratch stick, and switch, respectively. (**a**) Shows the number of commands used during each phase of the task while (**b**) shows the time it took to complete each phase of each task.

Table B.4: Metrics for the tests with ablebodied individuals and users. Mean and standard deviations of each metric of the described tests for both individuals with tetraplegia and ablebodied individuals.

Measure	Able-Bodied Individuals				Individuals with Tetraplegia			
	Bottle	Scratch Stick	Strawberry	Switch	Bottle	Scratch Stick	Strawberry	Switch
Time to object [s]	17.66 ± 4.13	29.66 ± 8.25	20.69 ± 4.09	34.30 ± 10.78	21.08 ± 4.05	44.17 ± 17.94	33.58 ± 13.44	41.39 ± 9.47
Time to reach mouth/ face shield [s]	21.02 ± 8.04	40.62 ± 15.76	42.05 ± 12.99		34.30 ± 12.00	62.50 ± 15.91	58.69 ± 9.82	
Number of commands until object [#]	4.60 ± 1.78	6.62 ± 2.41	5.10 ± 1.86	9.31 ± 3.95	6.22 ± 1.71	11.94 ± 3.87	8.78 ± 2.78	11.33 ± 2.83
Number of commands to face [#]	5.37 ± 2.82	9.10 ± 2.82	9.35 ± 4.33		7.78 ± 3.37	15.06 ± 8.94	13.89 ± 5.32	

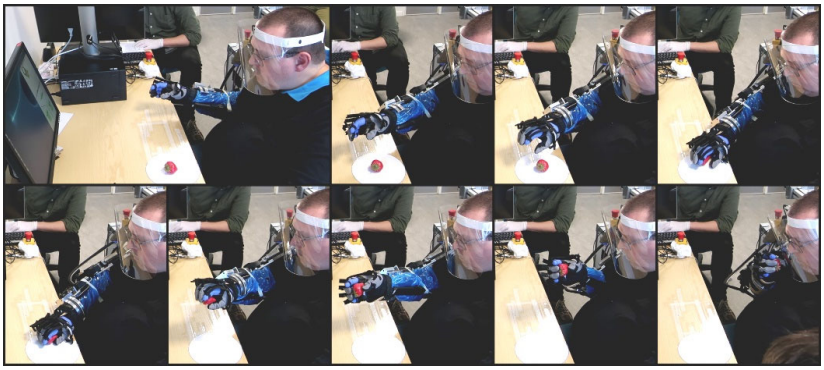


Fig. B.11: Image series of the strawberry task being performed by a user. The images show the initial approach, turn of the wrist, grasping and transfer to the mouth. Note: the blue plastic sleeve on the arm and the purple glove underneath the exoskeleton glove are only present due to the Covid19 outbreak happening during the experiments.

There were no occurrences of AD during the experiments. For an image series showing the strawberry task with a user, see Figure B.11. For video sequences of the ADL tasks, see Movie M1.

4.1 Interviews

After ending the experiments, a semi-structured individual interview was conducted with each of the three users with tetraplegia. When questioned about how they experienced the functioning of the exoskeleton, the users indicated that the functioning was good. However, one of the users, user 1, who had some remaining arm function noted, “There is one function that I think I would like to have eventually. Currently, it can turn the wrist, but it cannot tip the wrist [abduction/adduction]. I think that will be missing if

it is not there in the long run." User 2 answered, "I think there is so much potential in this project. The freedom it would be to be able to pick up a bottle, drink from it yourself, and decide yourself. It would mean a massive difference. Function-wise, I think it is good, about where it should be." User 3, who had the highest cervical damage and no arm function, agreed, "It has been great, great to be able to move the arm again - it was delightful."

When asked whether they could see themselves use it in their everyday, the consensus was that the appearance should be more attractive, but they could all imagine using it. User 1 noted, "With respect to functioning and sound, I wouldnt have second thoughts about using it, [...] but I think it is unattractive." When asked about the appearance, user 3 noted, "It looked big, but it would probably be smaller when attached to the arm [if it was attached directly to the powered wheelchair], so I think it is good at this point" and added that function and appearance are equally important. When questioned about the soft exoskeleton glove functionality, all users agreed that it needed substantial improvements as it simply could not supply enough grip force and spread between the fingers and the thumb.

The final questions revolved around the control method. The users, except user 3, agreed that the tongue control was tricky at first but noted that, "It will likely be easier with more time and training." This fits well with prior experiences with learning curves [65] and the short time for training in this study.

5 Discussion

Through this work, we present the first shared, semi-autonomous, tongue-based upper limb exoskeleton system capable of fully assisting individuals with severe tetraplegia to perform multiple ADLs. In particular, this work shows the prospect for users with complete functional tetraplegia to be empowered through high DoF robotic devices by using a tongue control interface in combination with computer vision assistance as any possible motions of the compact arm and hand exoskeleton could be controlled directly and continuously by the users, thus facilitating a large variety of ADLs.

The EXOTIC exoskeleton was created to solve several challenges regarding full upper limb assistance, compact and mobile design, calibration and mounting time, ergonomic mounting to mitigate the risk of provoking AD, incorporation of user desires to perform certain ADLs, and finally to provide an effective, intelligent, and invisible control interface. These challenges were resolved by creating a novel five DoF exoskeleton with an ergonomic and loose mounting to the user and a workspace that included the space in which common ADLs could be performed. The control interface, the adapted iTongue, provided a concealed, flexible, and effective control of the

exoskeleton shared with an optional semi-autonomous functionality to assist in grasping. Despite using a screen for visual feedback in this study, the tongue interface can be used without it in trained users who memorize the control layout, akin to touch-typing on regular computer keyboards. The results from the tests performed in this study showed that it was possible to grab a bottle and move it to the face in less than one minute on average for both groups, whereas other more complicated tasks could be performed within two minutes in most cases. The users, who were interviewed after the experiment, endorsed the exoskeleton and attested the use of exoskeletons as assistive devices for individuals with tetraplegia.

During the user tests, it became apparent that the simple implementation of the soft exoskeleton glove chosen for this experiment exhibited some shortcomings that were not observed during tests on able-bodied individuals. The explanation for this discrepancy is most likely that the able-bodied individuals have flexible finger joints compared with the users, who exhibited contractures and spasticity in the finger joints. Thus, more advanced [22,28] or mechanically rigid solutions for grasping are recommended in future studies.

Further, the temporary silicone fitting used for the tongue interface in this study has a significantly larger size than the standard interface, which may have reduced the mobility of the tongue and may have affected the control efficiency and learnability as compared with the standard iTongue interface. The commercial iTongue unit would be a drop-in replacement for the modified device used in this study as they are technically identical, except from the software defining the layout of the commands (Figure B.5c). The commercial version would be a safer, more comfortable, and likely an even easier device to use.

When comparing the able-bodied individuals with the actual users, a discrepancy between the two groups is clear. This discrepancy is probably due to the differences in demographics, prior training with the iTongue control interface, and the difficulties with he soft exoskeleton glove. Finally, the able-bodied individuals had a longer training period with the system and the age of the two groups was considerably different. These factors may well have caused the discrepancy. The differences observed would most likely diminish with more training, matched groups, and a better hand opening and closing mechanism. Further, a larger sample group of individuals with tetraplegia would better represent the actual performance of this group, though the number of participants is comparable to similar studies on individuals with complicated medical conditions [15, 22, 23]. An increase in the number of participants is often accompanied by a shorter experiment [22].

Nann et al. [22] demonstrated the use of state-based BMI control of a four DoF upper limb exoskeleton [23] with a five DoF wrist and hand exoskeleton. Their results showed that individuals with hemiparesis from stroke were

able to execute a drinking task using their system. However, the users in that study still had arm functionality and the exoskeleton used was of a considerable size compared with the exoskeleton presented here. As indicated in the user interviews in our study, both functionality and appearance should be of high priority. BMIs have the advantage that they can be used by individuals without tongue function, but while the deployed state-based control for the BMI [23] may ensure better compliance in performing certain tasks, it reduces the flexibility considerably compared to the direct continuous tongue-based control employed in this work. All the users with tetraplegia participating in this study could imagine using the presented EXOTIC exoskeleton system in their everyday life.

6 Conclusion

Until recently, options for individuals with complete functional tetraplegia to control all motions of high DoF arm/hand exoskeletons continuously and directly have been practically non-existing. The results of and the user feedback on the presented combination of an adapted available tongue control system, optional shared autonomous function, and a full compact and mobile arm/hand exoskeleton with a user-driven design indicate that this may be a viable solution to regain some independence and significantly increase the quality of life, even for individuals with complete functional tetraplegia.

Supplementary Materials: The following supporting information can be downloaded at: `https://www.mdpi.com/article/10.3390/s22186919/s1`, Video S1: Demonstration of the experiment tasks using the EXOTIC exoskeleton.

Author Contributions: Conceptualization: M.B.T., M.M., M.A.G., S.H.B., F.V.K., S.B., T.B., T.B.M., A.M.K and L.N.S.A.S; Data curation: M.B.T., M.M., M.A.G., S.H.B., F.V.K., B.Y.A., K.E.S. and L.N.S.A.S; Formal analysis: M.B.T.; Funding acquisition: S.B., T.B., T.B.M., A.M.K. and L.N.S.A.S; Investigation: M.B.T., M.M., S.H.B., F.V.K., B.B., B.Y.A., K.E.S. and L.N.S.A.S; Methodology: M.B.T., M.M., M.A.G., S.H.B., F.V.K., B.B., K.E.S. and L.N.S.A.S; Project administration: S.B., T.B., T.B.M., A.M.K. and L.N.S.A.S; Resources: M.M., M.A.G., B.B., B.Y.A., K.E.S., S.B., T.B. and T.B.M.; Software: M.B.T., M.M. and S.H.B.; Supervision: S.B., T.B., T.B.M., A.M.K. and L.N.S.A.S; Validation: M.B.T., M.M. and L.N.S.A.S; Visualization: M.B.T., M.M. and L.N.S.A.S; Writing original draft: M.B.T. and L.N.S.A.S; Writing - review & editing: M.B.T., M.M., M.A.G., S.H.B., F.V.K., B.B., B.Y.A., K.E.S., S.B., T.B., T.B.M., A.M.K. and L.N.S.A.S; All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted as part of the EXOTIC project funded by Aalborg University, Denmark.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Science Ethics Committee for the North Denmark Region (reg. no.: VN-20190030 and VN-20210016, approved 17 August 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not publicly available due to privacy concerns.

Acknowledgments: We would like to thank the participants in the studies for their effort and patience. Additionally, we would like to thank Rasmus Leck Kæseler for always being ready to step in and help when needed.

Conflict of Interest: The authors declare no conflicts of interest.

References

[1] J. Bickenbach, A. Officer, T. Shakespeare, P. von Groote, W. H. Organization, and T. I. S. C. Society, *International perspectives on spinal cord injury / edited by Jerome Bickenbach ... [et al]*. World Health Organization, 2013.

[2] M. Wyndaele and J. J. Wyndaele, "Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey?" *Journal of the International Spinal Cord Society (ISCoS)*, no. 44, pp. 523–529, 2006.

[3] A. B. Jackson, M. Dijkers, M. J. DeVivo, and R. B. Poczatek, "A demographic profile of new traumatic spinal cord injuries: Change and stability over 30 years," *Archives of Physical Medicine and Rehabilitation*, vol. 85, no. 11, pp. 1740–1748, Nov. 2004. [Online]. Available: https://doi.org/10.1016/j.apmr.2004.04.035

[4] P. J. Manns and K. E. Chad, "Components of quality of life for persons with a quadriplegic and paraplegic spinal cord injury," *Qualitative Health Research*, vol. 11, no. 6, pp. 795–811, Nov. 2001. [Online]. Available: https://doi.org/10.1177/104973201129119541

[5] R. Williams and A. Murray, "Prevalence of depression after spinal cord injury: A meta-analysis," *Archives of Physical Medicine and*

Rehabilitation, vol. 96, no. 1, pp. 133–140, Jan. 2015. [Online]. Available: https://doi.org/10.1016/j.apmr.2014.08.016

[6] United Nations Department of Economic and Social Affairs Population Division 2015 (ST/ESA/SER.A/390), *World Population Ageing 2015*. United Nations, 2015.

[7] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, "Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics*, June 2011, pp. 1–5.

[8] C.-S. Chung, H. Wang, and R. A. Cooper, "Functional assessment and performance evaluation for assistive robotic manipulators: Literature review," *The Journal of Spinal Cord Medicine*, vol. 36, no. 4, pp. 273–289, 2013.

[9] M. A. Gull, S. Bai, and T. Bak, "A review on design of upper limb exoskeletons," *Robotics*, vol. 9, no. 1, p. 16, Mar. 2020.

[10] M. R. Islam, C. Spiewak, M. H. Rahman, and R. Fareh, "A brief review on robotic exoskeletons for upper extremity rehabilitation to find the gap between research porotype and commercial type," *Advances in Robotics & Automation*, vol. 06, no. 03, 2017. [Online]. Available: https://doi.org/10.4172/2168-9695.1000177

[11] T. Nef, M. Mihelj, G. Kiefer, C. Perndl, R. Muller, and R. Riener, "ARMin - exoskeleton for arm therapy in stroke patients," in *2007 IEEE 10th International Conference on Rehabilitation Robotics*. IEEE, Jun. 2007, pp. 68–74. [Online]. Available: https://doi.org/10.1109/icorr.2007.4428408

[12] B. Kim and A. D. Deshpande, "An upper-body rehabilitation exoskeleton harmony with an anatomical shoulder mechanism: Design, modeling, control, and performance evaluation," *The International Journal of Robotics Research*, vol. 36, no. 4, pp. 414–435, Apr. 2017. [Online]. Available: https://doi.org/10.1177/0278364917706743

[13] R. A. R. C. Gopura, K. Kiguchi, and Y. Li, "SUEFUL-7: A 7dof upper-limb exoskeleton robot with muscle-model-oriented EMG-based control," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2009, pp. 1126–1131. [Online]. Available: https://doi.org/10.1109/iros.2009.5353935

[14] J. Huang, X. Tu, and J. He, "Design and evaluation of the RUPERT wearable upper extremity exoskeleton robot for clinical and in-home therapies," *IEEE Transactions on Systems, Man, and Cybernetics:*

Systems, vol. 46, no. 7, pp. 926–935, Jul. 2016. [Online]. Available: https://doi.org/10.1109/tsmc.2015.2497205

[15] E. Kirchner, N. Will, M. Simnofske, L. Benitez, B. Bongardt, M. M. Krell, S. Kumar, M. Mallwitz, A. Seeland, M. Tabie, H. Woehrle, M. Yüksel, A. HeSS, R. Buschfort, and F. Kirchner, "Recupera-reha: Exoskeleton technology with integrated biosignal analysis for sensorimotor rehabilitation," *Technische Unterstützungssysteme, die die Menschen wirklich wollen*, pp. 504–517, 12 2016.

[16] S. Kumar, H. Wöhrle, M. Trampler, M. Simnofske, H. Peters, M. Mallwitz, E. Kirchner, and F. Kirchner, "Modular design and decentralized control of the recupera exoskeleton for stroke rehabilitation," *Applied Sciences*, vol. 9, no. 4, p. 626, Feb. 2019. [Online]. Available: https://doi.org/10.3390/app9040626

[17] P. Garrec, J. Friconneau, Y. Measson, and Y. Perrot, "ABLE, an innovative transparent exoskeleton for the upper-limb," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Sep. 2008, pp. 1483–1488. [Online]. Available: https://doi.org/10.1109/iros.2008.4651012

[18] A. Kapsalyamov, S. Hussain, and P. K. Jamwal, "State-of-the-art assistive powered upper limb exoskeletons for elderly," *IEEE Access*, vol. 8, pp. 178 991–179 001, 2020. [Online]. Available: https://doi.org/10.1109/access.2020.3026641

[19] S. Bai, S. Christensen, M. Islam, S. Rafique, N. Masud, P. Mattsson, L. O'Sullivan, and V. Power, "Development and testing of full-body exoskeleton AXO-SUIT for physical assistance of the elderly," in *Biosystems & Biorobotics*. Springer International Publishing, Oct. 2018, pp. 180–184. [Online]. Available: https://doi.org/10.1007/978-3-030-01887-0_35

[20] T. Otsuka, K. Kawaguchi, H. Kawamoto, and Y. Sankai, "Development of upper-limb type HAL and reaching movement for meal-assistance," in *2011 IEEE International Conference on Robotics and Biomimetics*. IEEE, Dec. 2011, pp. 883–888. [Online]. Available: https://doi.org/10.1109/robio.2011.6181399

[21] D. Sui, J. Fan, H. Jin, X. Cai, J. Zhao, and Y. Zhu, "Design of a wearable upper-limb exoskeleton for activities assistance of daily living," in *2017 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, Jul. 2017, pp. 845–850. [Online]. Available: https://doi.org/10.1109/aim.2017.8014123

[22] M. Nann, F. Cordella, E. Trigili, C. Lauretti, M. Bravi, S. Miccinilli, J. M. Catalan, F. J. Badesa, S. Crea, F. Bressi, N. Garcia-Aracil, N. Vitiello, L. Zollo, and S. R. Soekadar, "Restoring activities of daily living using an eeg/eog-controlled semiautonomous and mobile whole-arm exoskeleton in chronic stroke," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2314–2321, 2021.

[23] S. Crea, M. Nann, E. Trigili, F. Cordella, A. Baldoni, F. J. Badesa, J. M. Catalán, L. Zollo, N. Vitiello, N. G. Aracil, and S. R. Soekadar, "Feasibility and safety of shared eeg/eog and vision-guided autonomous whole-arm exoskeleton control to perform activities of daily living," *Scientific Reports*, vol. 8, no. 1, p. 10823, Jul 2018.

[24] M. Barsotti, D. Leonardis, C. Loconsole, M. Solazzi, E. Sotgiu, C. Procopio, C. Chisari, M. Bergamasco, and A. Frisoli, "A full upper limb robotic exoskeleton for reaching and grasping rehabilitation triggered by MI-BCI," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*. IEEE, Aug. 2015, pp. 49–54.

[25] H. A. Caltenco, B. Breidegard, B. Jönsson, and L. N. A. Struijk, "Understanding computer users with tetraplegia: Survey of assistive technology users," *International Journal of Human-Computer Interaction*, vol. 28, no. 4, pp. 258–268, Mar. 2012. [Online]. Available: https://doi.org/10.1080/10447318.2011.586305

[26] S. Li, X. Zhang, and J. D. Webb, "3-d-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2824–2835, Dec. 2017. [Online]. Available: https://doi.org/10.1109/tbme.2017.2677902

[27] C. E. Bouton, A. Shaikhouni, N. V. Annetta, M. A. Bockbrader, D. A. Friedenberg, D. M. Nielson, G. Sharma, P. B. Sederberg, B. C. Glenn, W. J. Mysiw, A. G. Morgan, M. Deogaonkar, and A. R. Rezai, "Restoring cortical control of functional movement in a human with quadriplegia," *Nature*, vol. 533, no. 7602, pp. 247–250, Apr. 2016. [Online]. Available: https://doi.org/10.1038/nature17435

[28] S. R. Soekadar, M. Witkowski, C. Gómez, E. Opisso, J. Medina, M. Cortese, M. Cempini, M. C. Carrozza, L. G. Cohen, N. Birbaumer, and N. Vitiello, "Hybrid EEG/EOG-based brain/neural hand exoskeleton restores fully independent daily living activities after quadriplegia," *Science Robotics*, vol. 1, no. 1, Dec. 2016. [Online]. Available: https://doi.org/10.1126/scirobotics.aag3296

[29] R. Readioff, Z. K. Siddiqui, C. Stewart, L. Fulbrook, R. J. O'Connor, and E. K. Chadwick, "Use and evaluation of assistive technologies for upper limb function in tetraplegia," *The Journal of Spinal Cord Medicine*, pp. 1–12, Feb. 2021. [Online]. Available: https://doi.org/10.1080/10790268.2021.1878342

[30] S. H. Bengtson, T. Bak, L. N. S. A. Struijk, and T. B. Moeslund, "A review of computer vision for semi-autonomous control of assistive robotic manipulators (arms)," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731–745, 2020.

[31] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[32] M. Hildebrand, F. Bonde, R. Kobborg, C. Andersen, A. Norman, M. Thøgersen, S. Bengtson, S. Dosen, and L. Struijk, "Semi-autonomous tongue-control of an assistive robotic arm for individuals with quadriplegia," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, ser. IEEE International Conference on Rehabilitation Robotics. Proceedings, vol. 2019. United States: IEEE, Jun. 2019, pp. 157–162.

[33] D. Hill, C. S. Holloway, D. Z. M. Ramirez, P. Smitham, and Y. Pappas, "WHAT ARE USER PERSPECTIVES OF EXOSKELETON TECHNOLOGY? a LITERATURE REVIEW," *International Journal of Technology Assessment in Health Care*, vol. 33, no. 2, pp. 160–167, 2017. [Online]. Available: https://doi.org/10.1017/s0266462317000460

[34] K. Shinohara and J. O. Wobbrock, "In the shadow of misperception: Assistive technology use and social interactions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2011. [Online]. Available: https://doi.org/10.1145/1978942.1979044

[35] F. V. Kobbelgaard, S. Bødker, and A. M. Kanstrup, "Designing a game to explore human artefact ecologies for assistive robotics," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, Oct. 2020.

[36] F. V. Kobbelgaard, A. M. Kanstrup, and L. N. S. A. Struijk, "Exploring user requirements for an exoskeleton arm insights from a user-centered

study with people living with severe paralysis," in *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 2021, pp. 312–320.

[37] M. Mohammadi, H. Knoche, M. Gaihede, B. Bentsen, and L. N. S. Andreasen Struijk, "A high-resolution tongue-based joystick to enable robot control for individuals with severe disabilities," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, 2019, pp. 1043–1048.

[38] A. Kintsch and R. Depaula, "A framework for the adoption of assistive technology," *Proceedings of SWAAAC 2002: Supporting Learning Through Assistive Technology*, 01 2002.

[39] K. C. Eldahan and A. G. Rabchevsky, "Autonomic dysreflexia after spinal cord injury: Systemic pathophysiology and methods of management," *Autonomic Neuroscience*, vol. 209, pp. 59–70, Jan. 2018. [Online]. Available: https://doi.org/10.1016/j.autneu.2017.05.002

[40] M. Dezman, T. Asfour, A. Ude, and A. Gams, "Exoskeleton arm pronation/supination assistance mechanism with a guided double rod system," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, Oct. 2019, pp. 559–564. [Online]. Available: https://doi.org/10.1109/humanoids43949.2019.9034992

[41] U. A. T. Hofmann, T. Butzer, O. Lambercy, and R. Gassert, "Design and evaluation of a bowden-cable-based remote actuation system for wearable robotics," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2101–2108, Jul. 2018. [Online]. Available: https://doi.org/10.1109/lra.2018.2809625

[42] D. Sasaki, T. Noritsugu, and M. Takaiwa, "Development of active support splint driven by pneumatic soft actuator (ASSIST)," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005, pp. 520–525. [Online]. Available: https://doi.org/10.1109/robot.2005.1570171

[43] S. Lessard, P. Pansodtee, A. Robbins, L. B. Baltaxe-Admony, J. M. Trombadore, M. Teodorescu, A. Agogino, and S. Kurniawan, "CRUX: A compliant robotic upper-extremity exosuit for lightweight, portable, multi-joint muscular augmentation," in *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, Jul. 2017. [Online]. Available: https://doi.org/10.1109/icorr.2017.8009482

[44] E. Casanova-Batlle, M. de Zee, M. Thøgersen, Y. Tillier, and L. N. Andreasen Struijk, "The impact of an underactuated arm exoskeleton on

wrist and elbow kinematics during prioritized activities of daily living," *Journal of Biomechanics*, vol. 139, p. 111137, 2022.

[45] L. A. V. der Heide, B. van Ninhuijs, A. Bergsma, G. J. Gelderblom, D. J. van der Pijl, and L. P. de Witte, "An overview and categorization of dynamic arm supports for people with decreased arm function," *Prosthetics & Orthotics International*, vol. 38, no. 4, pp. 287–302, Aug. 2014. [Online]. Available: https://doi.org/10.1177/0309364613498538

[46] M. Gull, M. Thøgersen, S. Bengtson, M. Mohammadi, L. Struijk, T. Moeslund, T. Bak, and S. Bai, "A 4-dof upper limb exoskeleton for physical assistance: Design, modeling, control and performance evaluation," *Applied Sciences*, vol. 11, no. 13, Jun. 2021.

[47] B. D. M. Chaparro-Rico, D. Cafolla, M. Ceccarelli, and E. Castillo-Castaneda, "NURSE-2 DoF device for arm motion guidance: Kinematic, dynamic, and FEM analysis," *Applied Sciences*, vol. 10, no. 6, p. 2139, Mar. 2020. [Online]. Available: https://doi.org/10.3390/app10062139

[48] J. F. Rodríguez-León, B. D. M. Chaparro-Rico, M. Russo, and D. Cafolla, "An autotuning cable-driven device for home rehabilitation," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–15, Feb. 2021. [Online]. Available: https://doi.org/10.1155/2021/6680762

[49] L. Struijk, E. Lontis, M. Gaihede, H. Caltenco, M. Lund, H. Schiøler, and B. Bentsen, "Development and functional demonstration of a wireless intraoral inductive tongue computer interface for severely disabled persons," *Disability and Rehabilitation: Assistive Technology*, vol. 12, no. 6, pp. 631–640, 2017.

[50] L. Struijk, "An inductive tongue computer interface for control of computers and assistive devices," *I E E E Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2594–2597, 2006.

[51] E. Lontis, B. Bentsen, M. Gaihede, and L. Struijk, "Sensor activation for wheelchair driving in confined spaces with a tongue controlled oral interface," in *Proceedings of the International Convention on Rehabilitation Engineering & Assistive Technology. i-CREATe 2016, 25-28 July 2016, Bangkok, Thailand*. Singapore Therapeutic, Assistive & Rehabilitative Technologies (START), 2016.

[52] L. N. S. A. Struijk, L. L. Egsgaard, R. Lontis, M. Gaihede, and B. Bentsen, "Wireless intraoral tongue control of an assistive robotic arm for individuals with tetraplegia," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, Nov. 2017.

[53] L. N. S. A. Struijk, B. Bentsen, M. Gaihede, and E. R. Lontis, "Error-free text typing performance of an inductive intra-oral tongue computer interface for severely disabled individuals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 2094–2104, Nov. 2017. [Online]. Available: https://doi.org/10.1109/tnsre.2017.2706524

[54] M. E. Lund, H. V. Christiensen, H. A. Caltenco, E. R. Lontis, B. Bentsen, and L. N. S. A. Struijk, "Inductive tongue control of powered wheelchairs," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, Aug. 2010, pp. 3361–3364. [Online]. Available: https://doi.org/10.1109/iembs.2010.5627923

[55] D. Johansen, C. Cipriani, D. B. Popovic, and L. N. S. A. Struijk, "Control of a robotic hand using a tongue control system—a prosthesis application," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1368–1376, Jul. 2016. [Online]. Available: https://doi.org/10.1109/tbme.2016.2517742

[56] J. Clausen, E. Fetz, J. Donoghue, J. Ushiba, U. Spörhase, J. Chandler, N. Birbaumer, and S. R. Soekadar, "Help, hope, and hype: Ethical dimensions ofneuroprosthetics," *Science*, vol. 356, no. 6345, pp. 1338–1339, Jun. 2017. [Online]. Available: https://doi.org/10.1126/science.aam7731

[57] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa, "Is more autonomy always better?" in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Mar. 2020. [Online]. Available: https://doi.org/10.1145/3319502.3374818

[58] Ioan A. Sucan and Sachin Chitta, "MoveIt." [Online]. Available: https://moveit.ros.org/

[59] Ruben Smits, "Orocos Kinematics and Dynamics - KDL Wiki." [Online]. Available: https://www.orocos.org/kdl.html

[60] R. Tajimaaaaa, "jog_control - ROS wiki." [Online]. Available: https://wiki.ros.org/jog_control

[61] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (PCL)," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, May 2011, pp. 1–4. [Online]. Available: https://doi.org/10.1109/icra.2011.5980567

[62] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and auto-

mated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381395, jun 1981.

[63] S. H. Bengtson, M. B. Thøgersen, M. Mohammadi, F. V. Kobbelgaard, M. A. Gull, L. N. S. A. Struijk, T. Bak, and T. B. Moeslund, "Computer vision-based adaptive semi-autonomous control of an upper limb exoskeleton for individuals with tetraplegia," *Applied Sciences*, vol. 12, no. 9, p. 4374, Apr. 2022.

[64] M. Mohammadi, H. Knoche, B. Bentsen, M. Gaihede, and L. N. S. A. Struijk, "A pilot study on a novel gesture-based tongue interface for robot and computer control," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, Oct. 2020.

[65] M. Mohammadi, H. Knoche, and L. N. S. A. Struijk, "Continuous tongue robot mapping for paralyzed individuals improves the functional performance of tongue-based robotic assistance," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2552–2562, Aug. 2021.

[66] E. R. Lontis, M. E. Lund, H. V. Christensen, B. Bentsen, M. Gaihede, H. A. Caltenco, and L. N. S. A. Struijk, "Clinical evaluation of wireless inductive tongue computer interface for control of computers and assistive devices," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, Aug. 2010, pp. 3365–3368. [Online]. Available: https://doi.org/10.1109/iembs.2010.5627924

[67] M. Mohammadi, H. Knoche, M. Thøgersen, S. H. Bengtson, M. A. Gull, B. Bentsen, M. Gaihede, K. E. Severinsen, and L. N. S. A. Struijk, "Eyes-free tongue gesture and tongue joystick control of a five DOF upper-limb exoskeleton for severely disabled individuals," *Frontiers in Neuroscience*, vol. 15, Dec. 2021.

Paper C

A Review of Computer Vision for Semi-Autonomous Control of Assistive Robotic Manipulators (ARMs)

Stefan Hein Bengtson, Thomas Bak, Lotte N. S. Andreasen Struijk, and Thomas Baltzer Moeslund

The paper has been published in
Disability and Rehabilitation: Assistive Technology, vol. 15, no. 7, pp. 731-745, 2019

This is an Accepted Manuscript of an article published by Taylor & Francis in *Disability and Rehabilitation: Assistive Technology* on the 03 Jul 2019, available at: http://www.tandfonline.com/10.1080/17483107.2019.1615998.
The layout has been revised.

Abstract

Purpose: *The advances in artificial intelligence have started to reach a level where autonomous systems are becoming increasingly popular as a way to aid people in their everyday life. Such intelligent systems may especially be beneficially for people struggling to complete common everyday tasks, such as individuals with movement-related disabilities. The focus of this paper is hence to review recent work in using computer vision for semi-autonomous control of assistive robotic manipulators (ARMs).*

Methods: *Four databases were searched using a block search, yielding 257 papers which were reduced to 14 papers after apply-ing various filtering criteria. Each paper was reviewed with focus on the hardware used, the autonomous behaviour achieved using computer vision and the scheme for semi-autonomous control of the system. Each of the reviewed systems were also sought characterized by grading their level of autonomy on a pre-defined scale.*

Conclusions: *A re-occurring issue in the reviewed systems was the inability to handle arbitrary objects. This makes the systems unlikely to perform well outside a controlled environment, such as a lab. This issue could be addressed by having the systems recognize good grasping points or primitive shapes instead of specific predefined objects. Most of the reviewed systems did also use a rather simple strategy for the semi-autonomous control, where they switch either between full manual control or full automatic control. An alternative could be a control scheme relying on adaptive blending which could provide a more seamless experience for the user*

1 Introduction

Machines are becoming increasingly smarter and the effort invested into research in artificial intelligence is at an all time high. This large interest in artificial intelligence is triggered by its ability to make smart decision to aid us in our everyday life.

The healthcare sector is one area which could benefit immensely from artificial intelligence by enabling assistive devices to act autonomously. Autonomous machines could for instance assist the elderly and disabled individuals in feeding, getting dressed and other activities of daily living. This is especially of interest given the increasing demand for caregivers [1, 2].

Persons suffering from quadriplegia, i.e. total or partial loss of control of all four limbs, would especially benefit from such assistance due to the severity of their disability. For instance, a study found that the use of an assistive robotic manipulator (ARM) could reduce the need for assistance with 1.25 hours per day for persons with upper-extremity disabilities [3]. Another study confirmed these findings as their results showed that the use of an

ARM could reduce the need for assistance by 41% [4]. This reduced need for assistance would not only be economically beneficial but also increase the users' quality of life by empowering them and providing them with some privacy. Furthermore, a survey on disabled persons found that 86% of the participants would consider purchasing an ARM given the possibility [5] .

Another factor making autonomous control of ARMs increasingly interesting is how readily available the necessary hardware is becoming. For instance, the commercially available ARMs, which are specifically targeted at empowering users with movement impairments, such as JACO from Kinova [6] or iARM from Exact Dynamics [7].

However, any system which is to behave autonomously must rely on some sort of input to make informed decision, for instance knowledge of its immediate environment. Computer vision is hence often a part of such autonomous systems as it enables the system to capture and understand visual information. For instance, recognizing objects in an image and figuring out how to grasp said object [8–10].

The purpose of introducing autonomous behaviour into these systems is to reduce both the time it takes to execute a task and to reduce the cognitive burden on the user. This research has been expanded to other types of ARMs as well, such as exoskeletons [11, 12]. The idea of using an exoskeleton is to provide a more integrated solution than e.g. a robotic arm mounted on a wheelchair.

However, having an ARM act autonomously is not necessarily a bliss for the user, even though it might reduce the time it takes to execute different tasks [13, 14]. An important aspect of using this technology is hence how the control is shared between the human and the machine, i.e. the design of a scheme allowing for semi-autonomous control of the ARM.

The contribution of this paper is hence a review of recent efforts in employing computer vision for semi-autonomous control of ARMs, such as exoskeletons or robotic arms. The goal of this review is to: (1) provide an overview of existing efforts in using computer vision for semi-autonomous control of ARMs; (2) highlight the current challenges associated with this area of research; and thereby (3) point out new directions of interest for this field.

2 Methods

The following outlines how the review was conducted in the terms of the literature search and subsequent sorting of found material. The extraction of data from each reviewed paper is described as well.

2.1 Data sources

The literature search was based on the following databases: *Engineering Village*, *Web of Science*, *Scopus*, and *Embase*. The search was conducted by constructing blocks of keywords related to computer vision, robotic manipulators and people with disabilities. A paper had to match at least one keyword from each of these blocks to show up when searching each database.

The keywords in each of these blocks were as follows:

- **Block 1 - computer vision**: *("computer vision" OR "robot vision" OR "robotic vision" OR "object detection" OR "image-based" OR "grasp detection" OR "vision-based" OR perception*)*
- **Block 2 - robotic manipulators**: *("robot arm" OR "robotic arm" OR "robot manipulator" OR "robotic manipulator" OR exoarm OR exoskeleton OR "personal robot")*
- **Block 3 - people with disabilities**: *(disab* OR impair* OR adl* OR "activities of daily living" OR handicap* OR "personal robot" OR rehabilitat*)*

It should be noted that the asterisk * serves as a wildcard for unknown terms and different inflections of the same word.

Only the titles, abstracts and keywords were used while searching and any results not in English were removed. Publications before 2008 and duplicates were removed as well. Only conference proceedings, reports and journals were included during the literature search and book chapters or book reviews were removed from the list of results. This initial search resulted in 257 results after applying the above filters, as illustrated in figure (C.1).

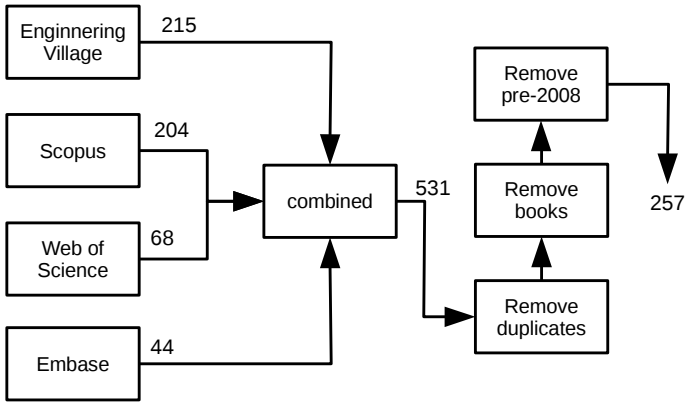


Fig. C.1: Databases and exclusion criteria used during the initial literature search.

2.2 Filtering Criteria

Additional criteria were imposed on the initial search to further narrow down the amount of relevant papers. Each paper should fulfil each of the following criteria to be considered relevant:

1. **Purpose**: The intended use of the system described in the paper should be object manipulation tasks. Papers focusing on e.g. rehabilitation and wheelchair navigation were discarded. This criterion was imposed to focus the scope of the review.
2. **Camera**: The system described in the paper should make use of a camera or a similar visual sensor, such as a laser scanner. Any papers failing this criterion are not doing computer vision and are hence outside the scope of this review.
3. **Disabled user**: The intended user of the system described in the paper should be a person suffering some kind of movement impairment, such that they would benefit from an ARM.
4. **Autonomous behaviour**: The papers should describe a system capable of exhibiting some degree of autonomous behaviour. Papers solely describing a way of directly controlling an ARM are discarded.
5. **Details**: The paper should be described in a sufficiently detailed way. A paper is considered sufficiently detailed if it is possible to identify the parameters and information described in the next section.

The initial set of 257 papers was reduced to 14 papers after applying the above criteria. Papers which were of interest, even though they failed the criteria, are included in the discussion part later.

2.3 Data Extraction

The first set of parameters extracted from each of the included papers relates to the platform and the hardware used in the system, namely: the type of sensor(s) used, the placement of the sensor(s), the type of robotic manipulator and the associated number of degrees of freedom. These parameters are of interest as they impact how both the computer vision and the semi-autonomous control functions. Furthermore, this information could reveal interesting tendencies in terms of hardware selection. It should be noted that the technologies for the user to interface with the different systems are not covered in this review.

The second set of parameters extracted from each paper focuses on the semi-autonomous behaviour of the different systems. This is done by identifying which parts of the system that acts autonomous, using computer vision, and in which part the human is still in control. These parameters are

extracted in systematic way by using the widely cited framework proposed by Parasuraman et al. [15].

This framework suggests that a semi-autonomous system can be split into four stages, as shown in figure (C.2). This model is based on a simple model of the way humans process information and act on it. The model is hence not intended to be perfect and all-encompassing but rather a simplification making it possible to impose some structure when analysing a system.

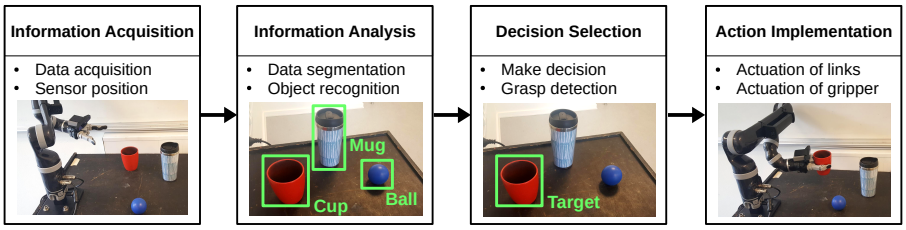


Fig. C.2: The four-stage model originally proposed by Parasuraman et al. [15], with examples of the tasks associated with each individual stage. The figure is adapted from Pitzer et al. [16].

The different stages of this four-stage model are:

- **Stage 1: Information acquisition**
  This stage contains functions related to sensing the environment such as gathering raw data from e.g. a camera. Calculations related to depth estimation can also be considered to belong in this step, for instance the registration between two cameras in a stereo vision setup. This stage can also include strategies for automatically moving the sensor(s) to better observe certain things. For instance, re-positioning the camera to get a better view of an object.

- **Stage 2: Information analysis**
  This stage is associated with the cognitive functions of the system. This is essentially the stage where the system interprets the information acquired during the previous stage. An example of such could be recognition of an object in an image, i.e. detecting an object's position and classifying the type of object.

- **Stage 3: Decision selection**
  The focus of this stage is to make a decision based on the multiple alternative options identified in the previous stage. The decision could for instance be which of the detected objects to pick-up and how to grasp said object. A system with a low level of autonomy would for instance offer the user all possible options. A system with a high level of autonomy would, on the other hand, act without user input and grab an object based on some pre-defined measure, e.g. the nearest object.

- **Stage 4: Action implementation**
  The final stage encompasses the actual execution of the necessary actions once a decision has been made. This includes sending the correct signals to the actuators, i.e. motors, of the robot to reach the desired goal such as the position of an object. It is also the stage responsible for actuating the gripper during grasping of objects.

Furthermore, Parasuraman et al. [15] also suggests a continuum when speaking of autonomous behaviour, ranging from a low-level to a high-level of autonomy. The authors specifically suggest 10 levels of autonomy, as outlined in table (C.1). This autonomy scale mainly relates to the last two stages of the four-stage model, i.e. decision selection and action implementation, and will hence only be applied in relation to these two stages.

Table C.1: The different levels of autonomy. Adapted from Parasuraman et al. [15].

Levels of autonomy
1) The system offers no assistance.
2) - offers a complete set of decisions/actions.
3) - narrows down the selection to a few.
4) - suggests one alternative.
5) - executes the suggestion if the human approves.
6) - allows the human a restricted time to veto before executing.
7) - executes automatically, then necessarily informs the human.
8) - informs the human only if asked.
9) - informs the human only if it, the system, decides to.
10) - decides everything, ignoring the human.

3 Results

The first part of this section summarizes the different hardware used in each of the reviewed papers. The second part outlines the semi-autonomous behaviour of each reviewed system by following the four-stage model presented earlier.

3.1 Hardware Selection Overview

The hardware associated with each of the reviewed systems are summarized in table (C.2). It should be noted that the stated degrees of freedom (DoF) in the table refers to the ARM only. DoFs gained from mounting on mobile platforms, such as wheelchairs, are not included.

Table C.2: Overview of the hardware used in the different reviewed papers. Hardware such as wheelchairs have been omitted from the table as the focus of this review is object manipulation using an ARM.

	Year	Sensor	Robotic platform
[17]	2008	**Sensor**: Passive stereo vision (custom). **Placement**: End-effector.	**Platform**: MANUS (Exact Dynamics) **Degrees of freedom**: 6
[8]	2009	**Sensor**: Passive stereo vision (custom). **Placement**: End-effector.	**Platform**: MANUS (Exact Dynamics) **Degrees of freedom**: 6
[16]	2011	**Sensor**: Active stereo vision (Kinect v1). **Placement**: Robot's head.	**Platform**: PR2 (Willow Garage) **Degrees of freedom**: 8
[18]	2012	**Sensor**: Passive stereo vision (custom) and force sensor. **Placement**: End-effector.	**Platform**: MANUS (Exact Dynamics) **Degrees of freedom**: 6
[19]	2013	**Sensor**: 2x Monocular RGB cameras. **Placement**: End-effector and overhead.	**Platform**: iARM (Exact Dynamics) **Degrees of freedom**: 6
[9]	2013	**Sensor**: Active stereo vision (2x Kinect v1). **Placement**: Table. Towards user and towards objects.	**Platform**: JACO (Kinova) **Degrees of freedom**: 6
[11]	2014	**Sensor**: Active stereo vision (Kinect v1). **Placement**: Table. Facing the user.	**Platform**: L-Exos, active wrist and hand orthosis (custom) **Degrees of freedom**: 8
[20]	2015	**Sensor**: Active stereo vision (Kinect v1). **Placement**: Overhead.	**Platform**: JACO (Kinova) **Degrees of freedom**: 6
[21]	2016	**Sensor**: Active stereo vision (2x Kinect v1). **Placement**: Table. Towards user and towards objects.	**Platform**: JACO (Kinova) **Degrees of freedom**: 6
[22]	2017	**Sensor**: Active stereo vision (Carmine). **Placement**: End-effector.	**Platform**: Baxter (Rethink Robotics) **Degrees of freedom**: 7
[10]	2017	**Sensor**: Passive stereo vision (Bumblebee). **Placement**: Overhead.	**Platform**: WAM Arm (Barrett Tech) **Degrees of freedom**: 7
[23]	2017	**Sensor**: Time-of-flight camera (2x Kinect v2). **Placement**: Table. Towards user and Towards objects.	**Platform**: JACO (Kinova) **Degrees of freedom**: 6
[24]	2017	**Sensor**: Time-of-flight camera (Kinect v2). **Placement**: Table. Towards user.	**Platform**: JACO (Kinova) **Degrees of freedom**: 6
[25]	2017	**Sensor**: Eye-tracking (EyeX) and RGB camera. **Placement**: Table.	**Platform**: Dobot Magician (Dobot) **Degrees of freedom**: 4

Sensor

Looking at the choice of sensor, most of the papers rely on some form of stereo vision to gather depth information with several of the papers using the Kinect v1 from Microsoft. This is a sensible choice given how easy the it is to acquire and work with, but it does impose restrictions in terms of possible mounting locations as it has a minimum distance of ≈ 0.5 m [26]. Any object closer than that is unlikely to be captured by the sensor.

The newer model, Kinect v2, is used in some of the more recent papers such as [23, 24]. This model relies on a time-of-flight (ToF) camera, instead of stereo vision, to acquire depth information but its minimum working distance is identical to the Kinect v1. These restrictions in terms of minimum distance is also clearly visible in the table, as no one mounts their Kinect sensors near the end-effector for this exact same reason.

A few of the reviewed papers, [9, 21, 23], even use two Kinects with the second one being orientated towards the user of the system. The purpose being either gesture recognition and/or detection of the user's face to move e.g. food to the mouth of the user.

The papers which mount their sensor near the end-effector primarily rely on customized stereo vision setups, likely because it allows them to control the baseline distance and hence their minimum distance. The only exception being [22] utilizing the Carmine camera from PrimeSense but this device is also marketed as having a minimum distance of ≈ 0.35 m, making it more suitable for such a mounting location than the Kinect.

Another discerning characteristics in terms of sensor choice is whether active or passive stereo vision is used. Active stereo vision relies on a light source, e.g. infrared light, to actively illuminate the scene whereas passive stereo vision relies on the ambient light only. Active stereo vision is hence more robust in terms of lacking illumination. Another benefit of active stereo vision is its ability to handle lack of texture as the active light source can be used to introduce texture in the scene. Lack of texture is a general problem in stereo vision as it makes it harder to recognize the same point in two images, which is needed to estimate the depth to said point. However, most of the reviewed systems using passive stereo vision are tested using highly textured objects, for instance [8, 17, 18], and may hence not experience this problem during the tests.

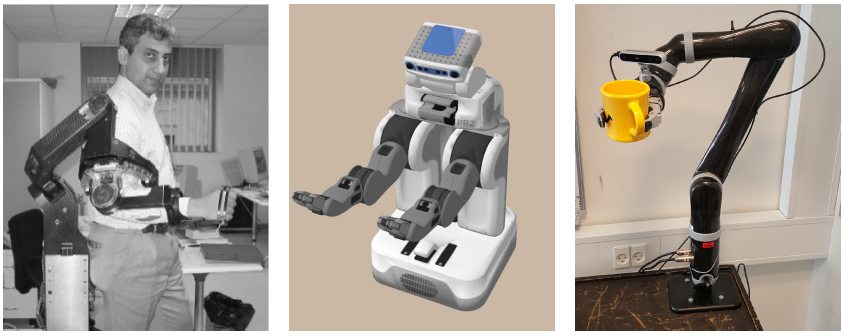
Furthermore, all the custom stereo vision setups found in the review are of the passive variety. This is not surprising as active stereo vision setups are generally more complicated to implement due to the active light source.

Robotic platform

There is a clear tendency of using robotic arms amongst the selected papers, as only one of them relies on an exoskeleton as their platform. This tendency can likely be explained by the accessibility of robotic arms, as they are generally cheaper than an exoskeleton and more readily available in the market. This observation is further emphasized by the fact that all the robotic arms listed in table (C.2) are commercially available.

The single exoskeleton, the L-Exos, is on the other hand custom made by one of the co-authors of the paper as described in [27]. This exoskeleton is also notable in the sense of its high number of degrees of freedom in comparison to the other system. It should however be noted that Loconsole et al. [11] states the redundancy of some of these DoFs.

Another outlier in terms of platform selection is the PR2 from Willow Garage, which is a full-blown robot featuring two arms and multiple sensors, such as the Kinect. This robot is intended to be used for teleoperation, i.e. being controlled remotely from a distance. Baxter from Rethink Robotics



(a) L-Exos [27]. **(b)** PR2 robot. **(c)** JACO.

Fig. C.3: Example of ARMs used in the reviewed papers.

is a full-blown robot as well, intended for industrial purposes. Gualtieri et al. [22] did however describe that they salvaged an arm from a Baxter robot, essentially reducing it to an ARM on-par with e.g. JACO from Kinova. A few of the mentioned robotic platforms are shown in figure (C.3).

3.2 Semi-autonomous Control Overview

The aspects related to the semi-autonomous behaviour of each reviewed system is summarized in table (C.3). The table follows the four-stage model from earlier and seeks to characterize the autonomous behaviour of each system by highlighting how each of them deals with certain aspects associated with each stage. Note that the system presented by Quintero et al. [20] can function in two distinct ways when considering the information analysis stage and the decision selection stage. This is signified by the notation [20]a and [20]b which is used to distinguish between these two configurations of the same system.

Information Acquisition

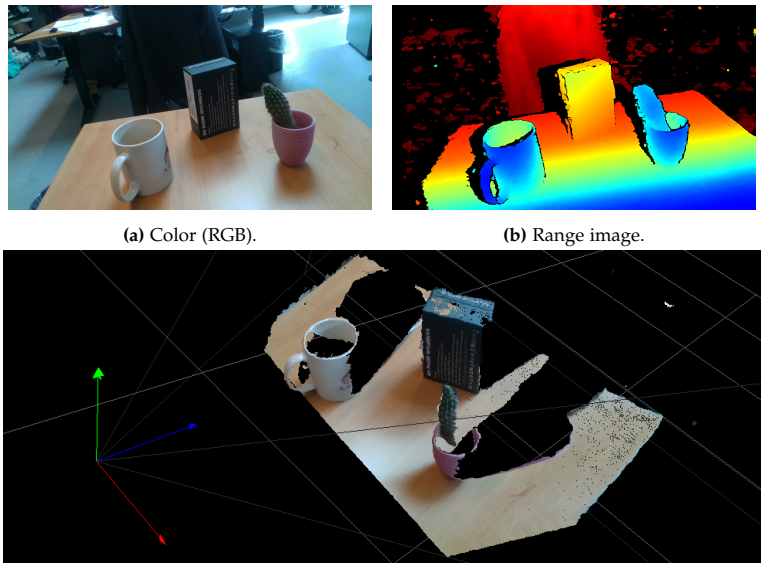
Two parameters where emphasized in this stage; 1) the type of data acquired by the system and 2) the ability to adapt the position of the sensors. Knowing the type of data gathering is important as it imposes restrictions on the system later. The ability to change the position is an important factor as well, as it influences what data that can be acquired.

All the reviewed papers collect RGB information, even though some system does not directly use it during the subsequent information analysis stage. This information is however used during the decision selection stage, to visualize different options to the user. Most of the reviewed papers do also acquire depth information, which is sensible given that the systems are ex-

Table C.3: Overview of the semi-autonomous aspects of the different reviewed papers.

	Year	Information Acquisition	Information analysis	Decision selection	Action implementation
[17]	2008	**Data**: RGB. Optical and force sensors in gripper. **Position**: User controlled.	**Segmentation**: Object bounding box from user. **Recognition**: None.	**Selection**: User draws object bounding box. **Grasping**: Estimates position but not orientation.	**Actuation**: Automatic but user can stop it at any time.
[8]	2009	**Data**: RGB and point cloud. Force sensors in gripper. **Position**: Automatic centering of user's selection.	**Segmentation**: Object point from user. **Recognition**: Template matching against database.	**Selection**: User selects single point on object. **Grasping**: Estimates position and orientation.	**Actuation**: User controls the gripper and the rate of link actuation.
[16]	2011	**Data**: RGB, point cloud and range images. **Position**: User controlled.	**Segmentation**: Object bounding box from user. **Recognition**: Matches point cloud against database.	**Selection**: User draws object bounding box. **Grasping**: Pre-computed for objects in the database.	**Actuation**: Automatic.
[18]	2012	**Data**: RGB and point cloud. Force sensors in gripper. **Position**: User controlled. Automatic centering of user's selection.	**Segmentation**: Object point from user. **Recognition**: Matches RGB image against database.	**Selection**: User selects single point on object. **Grasping**: Estimates position and orientation.	**Actuation**: Automatic.
[19]	2013	**Data**: RGB and point cloud. **Position**: Overhead camera is fixed. System controls end-effector camera.	**Segmentation**: None. **Recognition**: Matches RGB image against database.	**Selection**: Presents multiple graspable objects. User selects one object. **Grasping**: Pre-defined for objects in the database.	**Actuation**: Automatic.
[9]	2013	**Data**: RGB and range image. **Position**: Fixed.	**Segmentation**: None. **Recognition**: Matches RGB image against database.	**Selection**: Object(s) to be grasped was pre-defined. **Grasping**: Done by the user.	**Actuation**: Coarse control done automatic. Fine control done by user.
[11]	2014	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: Distance thresholding. **Recognition**: Recognizes cylinder-shaped objects.	**Selection**: Any cylinder-shaped objects detected. **Grasping**: Estimates position and orientation.	**Actuation**: Automatic.
[20]a	2015	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: Object bounding box from user. **Recognition**: None.	**Selection**: User draws object bounding box. **Grasping**: Estimates position and orientation.	**Actuation**: Either fully automatic or the user controls the rate of actuation.
[20]b	2015	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: Remove main planar surface. **Recognition**: None.	**Selection**: User Selects from list of detected objects. **Grasping**: Estimates position and orientation.	**Actuation**: Either fully automatic or the user controls the rate of actuation
[21]	2016	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: Remove main planar surface. **Recognition**: Matches image and point cloud against database.	**Selection**: User selects from a set of pre-defined actions. **Grasping**: Pre-defined for objects in the database.	**Actuation**: Automatic. User input is queued until the robot is done executing.
[22]	2017	**Data**: RGB and point cloud. **Position**: Automatic. System gathers data from multiple viewpoints.	**Segmentation**: None. **Recognition**: None.	**Selection**: User selects object using laser pointer. **Grasping**: Estimates position and orientation.	**Actuation**: Automatic.
[10]	2017	**Data**: RGB and point cloud. **Position**: User controlled.	**Segmentation**: Remove main planar surface. **Recognition**: Matches point cloud against database.	**Selection**: Intent infered using end-effectors position and orientation. **Grasping**: Pre-defined for objects in the database.	**Actuation**: Blending between system and user based on confidence of infered intent.
[23]	2017	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: Region growing using normals. **Recognition**: Matches image against database.	**Selection**: User selects from a set of pre-defined actions. **Grasping**: Estimates position but not orientation.	**Actuation**: Automatic.
[24]	2017	**Data**: RGB and point cloud. **Position**: Fixed.	**Segmentation**: None. **Recognition**: Marker-based.	**Selection**: User selects from a set of pre-defined actions. **Grasping**: Estimates position. Orientation is pre-defined.	**Actuation**: Automatic.
[25]	2017	**Data**: RGB and gaze points. **Position**: Fixed.	**Segmentation**: Color thresholding. **Recognition**: None. All objects are cuboids.	**Selection**: User selects from detected objects using gaze. **Grasping**: Estimates position and orientation.	**Actuation**: Coarse control done automatic. Fine control done by user.

pected to navigate in three dimensions to complete their task. The depth information are either represented as a point cloud or as a range image, as shown in figure (C.4).



(a) Color (RGB). **(b)** Range image.

(c) Point cloud with RGB data.

Fig. C.4: Different types of data from the same scene.

However, a few systems do not gather any depth information at all, which should in theory complicate tasks such as grasping objects. Remazeilles et al. [17] solves this issue by employing an optical sensor to detect when an object is inside the gripper and force sensors to ensure sufficient force when picking up the object. Another approach, used by Zeng et al. [25], is to make the simplifying assumption that all objects are cuboids and placed on a table.

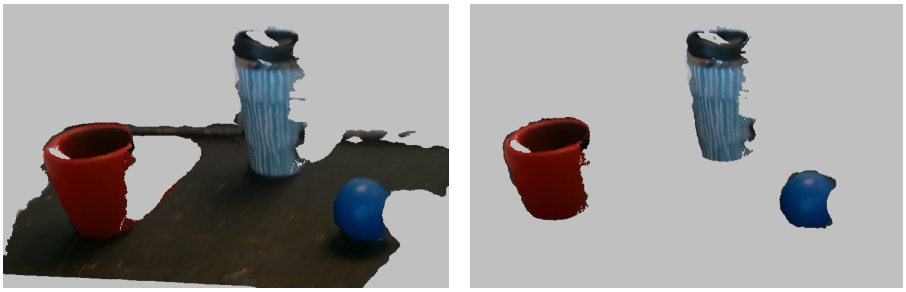
Looking at the positioning of the sensors, over half of the reviewed systems employ a fixed position, which could make them susceptible to blind spots. For instance, the ARM occluding the view of the camera. A strategy to avoid such issue is to enable the user to influence the position of the sensor by mounting it on the end-effector as done in several of the reviewed systems (see table (C.2) from earlier). A third option is to have the system automatically position the camera [8, 18, 19, 22].

Information Analysis

This stage can generally be characterized by two important tasks; 1) segmentation of data into what is of interest and what is not and 2) recognizing patterns in the data to recognize e.g. an object. Computer vision-systems

will often have well-defined strategies for these tasks which is why it was chosen to focus on these two aspects for this stage.

Looking at segmentation, a quite popular strategy is to remove the main planar surface in the scene, leaving behind objects placed on e.g. a table. An example of such is shown in figure (C.5). This planar surface is often found by using RANSAC (Random Sample Consensus) [28] to fit a plane to the point cloud data. The main drawback of this approach is the underlying assumption that the objects of interest are placed on a single planar surface without much else in the scene.



(a) Before segmentation. **(b)** After segmentation.

Fig. C.5: Example of segmentation of objects in a scene.

Another often used strategy amongst the reviewed papers is to rely on the user to manually perform the segmentation task. This is done either by drawing a bounding box around the object of interest or by selecting a single point on said object. In case of the latter, the point is used as the initial seed for segmentation algorithms such as the system describes by Pitzer et al. [16]. However, some of the reviewed papers, such as [9, 19], downright skip the segmentation step and are therefore processing information from the entire scene during the subsequent recognition step. This is possible as these systems relies on the SURF keypoint extractor and feature descriptor [29] which are optimized to be fast.

In terms of recognition, many of the reviewed papers relies on matching against a database of known objects. This is commonly done by extracting a set of features from the input data and then applying machine learning to match the features against the database of known objects. The majority of the papers either rely on point cloud or image data during this process, with only Jiang et al. [21] making use of both information sources in this step. Zhang et al. [23] combines both the feature extraction and matching process by training a CNN (Convolutional Neural Network) to distinguish between four pre-defined objects.

The drawback of relying on a pre-defined set of known objects is the inability to deal with objects which are not present in the database. Especially

the approach by Arrichiello et al. [24] suffers from this issue as it also requires the objects to be physically marked using pre-defined tags. A much more general approach is to match against primitive shapes, for instance cylinders [11]. In doing so, the system should be able to handle anything cylinder-shaped.

An even more general approach is used by Gualtieri et al. [22] which relies on identifying good grasping points instead of detecting objects in the scene. This essentially negates the need for both the segmentation and recognition steps for this system. The main problem of this approach is the time it takes to detect the grasping points in the scene, with the authors stating a processing time of two minutes on average.

Decision Selection

In this stage, each reviewed system was sought characterized based on; 1) how the system selects what to do, for instance what object to grasp and 2) its approach when deciding how to grasp an object. I.e. how to position and orient the end-effector of the ARM for the grasping procedure.

In relation to the decision selection of each reviewed system it is quite natural to also consider the associated level of autonomy. The scale presented earlier, see table (C.1), have hence been used to determine the level of autonomy for each system. The result is shown in table (C.4).

Table C.4: The reviewed papers and their level of autonomy based on their decision selection behaviour (stage 3). The indicators (a,b) signifies different configurations of the same system, as outlined in Table C.3.

Decision selection (level of autonomy)	
Level 1	[8, 16–18, 22] and [20]a
Level 2	[25] and [20]b
Level 3	[9, 19, 21, 23, 24]
Level 6	[10]
Level 10	[11]

Many of the reviewed systems rely on the user directly selecting the object to interact with. This is either done by having the user draw a bounding box around the object or selecting a point on it, as mentioned above. Such approaches rely entirely on the user and can hence be associated with the lowest level of autonomy. A few systems are a bit more restrictive, as they narrow the user's options down, either based on the objects detected in the scene by the system or a pre-defined set of options. These systems are given a rating of 2 and 3, respectively.

The system by Loconsole et al. [11] is however characterized by a high level of autonomy, as it will try to grasp any cylinder-shaped object presented to it. This system has been given a rating of 10, as the user has no say in the matter. Another outlier is the system described by Mülling et al. [10], as it automatically infers the intention of the user based on the end-effectors proximity to objects and how well the end-effectors orientation aligns with these objects. The system will automatically start to act on this estimated intention, but the user can still veto this decision by moving the end-effector in another direction, hence the rating of 6.

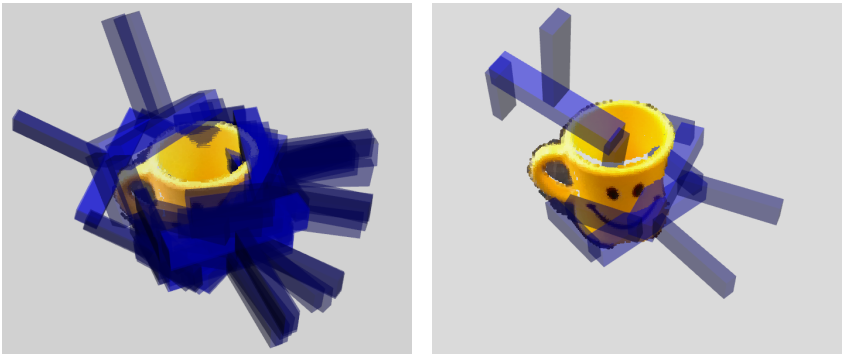
An important part of the decision selection stage is to figure out how to grasp an object to manipulate it. This entails figuring out how to position and orient the end-effector for the best grasp. How much to close the gripper is an important step in the grasping procedure as well but this part is not included in this review.

Several of the reviewed systems rely on both grasp positions and poses being pre-defined for a set of known objects. The drawback of relying on pre-defined information for a small set of objects is the inability to handle unknown objects, as stated earlier. Other papers ignore the problem of identifying a proper orientation of the end-effector and only estimates where to position the end-effector for grasping. This approach is possible as these systems makes assumptions like the objects always being placed such that their major axis is aligned vertically with the ARM. Such assumptions restrict the system's ability to function in an uncontrolled environment, where the objects are likely to be placed arbitrarily.

A few of the reviewed systems, like [8, 18], uses a PCA-based approach (Principal Component Analysis) in order to estimate the major axis of each object. This approach relies on point cloud data for each object and are hence dependent on proper segmentation of the object. Another drawback is that the estimated axis of the object may easily be miscalculated in cases where parts of the object are not present in the point cloud.

The approach used by Loconsole et al. [11] avoids this issue as it is quite straightforward to extract the major axis of a cylinder, which the system identified during the information analysis step. The disadvantage of this approach is the underlying assumption that every object is cylinder-shaped.

The only reviewed system which is truly able to grasp arbitrary objects is the one described by Gualtieri et al. [22] as it relies on detecting good grasping poses. The process of detecting these grasping poses is split into two stages; grasp synthesis and grasp analysis, as illustrated in figure (C.6). The synthesis stage seeks to generate a large number of grasp candidates whereas the analysis seeks to reduce the larger number of candidates to a few good ones. It should be noted that Mülling et al. [10] describes a similar extension of their system in their future works which enables them to handle arbitrary objects as well.



(a) Grasp synthesis stage. **(b)** Grasp analysis stage.

Fig. C.6: Detecting grasping poses for an arbitrary object using the algorithm by Gualtieri et al. [22].

Action Implementation

This stage is sought characterized by considering who is in control of the ARM's movement, i.e. who controls the actuation of the ARM's links and its end-effector. Factors like how trajectories are planned could have been considered as well in this stage, but it is deemed outside the scope of this review.

The action implementation stage is also a good candidate for judging a system's level of autonomy and this stage have hence been mapped using the autonomy scale as well. The result is shown in table (C.5).

Table C.5: The reviewed papers and their level of autonomy based on their action implementation behaviour (stage 4).

Action implementation (level of autonomy)	
Level 5	[8, 20]
Level 6	[17]
Level 7	[9, 25]
Level 10	[11, 16, 18, 19, 21–24]
Adaptive	[10]

The majority of the reviewed systems is assigned a score of 10, as the actuation of both the ARM's links and end-effector is fully automatic once a decision have been made. Jiang et al. [9] and Zeng et al. [25] are assigned a lower score of 7, as these two systems relies on the idea of dividing the control of the ARM into fine and coarse control. Coarse control entails moving the end-effector to the general position of the object to manipulate and is done

automatically. Fine control deals with grasping the object and is initiated by the system, which then asks the user to take over and perform the grasping. It should be noted that Zeng et al. [25] does estimate the orientation of the object to be grasped but this information is only used to guide the user during fine control.

The system by Remazeilles et al. [17] is assigned a rating of 6 as it essentially allows the user to veto the automatic actuation of the ARM. Kim et al. [8] and Quintero et al. [20] employs a scheme where the user continuously have to allow the system to operate automatically, for instance by holding down a button. This results in a score of 5 as the system is essentially limited to only executing actions if the user approves. Finally, the system by Mülling et al. [10] is not assigned a score as the automation level changes depending of the system's confidence in inferring the intention of the user. For instance, the user will be completely in control if the system has no idea about the intention of the user. Furthermore, it should be noted that authors of Parasuraman et al. [15] do point out that their framework fails to encompass such adaptive automation well.

3.3 Level of Autonomy Summary

The purpose of this section is to summarize the results related to the level of autonomy of the reviewed systems to highlight tendencies. This is done by grouping each of the reviewed papers, as shown in table (C.6).

These groups are created by grouping systems where the level of autonomy is identical for both the decision selection and action implementation stage. The resulting groups are then plotted, as shown in figure (C.7), with respect to their level of autonomy for the decision selection and action implementation stage.

Table C.6: Grouping of the reviewed systems based on their level of autonomy for the decision selection and action implementation stage. The indicators (a,b) signifies different configurations of the same system, as outlined in Table C.3

Group	Paper(s)
A	[19, 21, 23, 24]
B	[16, 18, 22]
C	[8] and [20]a
D	[17]
E	[20]b
F	[25]
G	[10]
H	[9]
I	[11]

Looking at the plot in figure (C.7) it is quite clear that most of the reviewed systems are placed in the upper left quadrant. Such systems are characterized by having a quite clear-cut strategy for sharing control, as the user decides what to grasp whereas the system performs the actual grasping. These approaches are hence quite similar to e.g. the claw machines found at arcades; the user points the machine towards the object of interest, the user presses a button and the machine takes over. A few of the reviewed systems did however allow the user some control in such scenarios, for instance the systems found in group *C* and *E*. These systems rely on constant confirmation from the user, e.g. holding down a button, to continue executing the planned action.

An entirely different approach for semi-autonomous control can be seen in group *G* consisting of only the system by Mülling et al. [10]. This group differs from the others due to its adaptive nature which is also sought illustrated in figure (C.7) by having this group span the entire action implementation continuum.

Another outlier is group *I*, consisting of the system by Loconsole et al. [11], which have the highest possible level of autonomy for both its decision selection and action implementation stage. It can hence be argued that this system is fully autonomous and hence of no interest when discussing semi-autonomous systems. To be fair, it should be noted that the focus of Loconsole et al. [11] is skewed towards computer vision and not semi-autonomous control.

4 Discussion

The purpose of this section is to expand upon the findings in the previous sections by pointing out challenges in relation to reviewed systems and suggest further potential avenues to explore. Three challenges will be discussed; ensuring optimal semi-autonomous control, handling arbitrary objects and sensing the environment.

4.1 Challenge: Optimal Semi-Autonomous Control

Most of the reviewed systems tend to rely on pre-defined roles for respectively the human and the system, as shown earlier in figure (C.7). The user decides what to do and the system takes over control, thereby creating this claw machine-like behaviour. The benefit of such schemes is that the user is never in doubt as to who is in control at any time.

However, this behaviour could be problematic as the user has no or very limited control once the system is in charge. This issue was outlined in a study by Chung et al. [13] which found that the users felt less accomplished

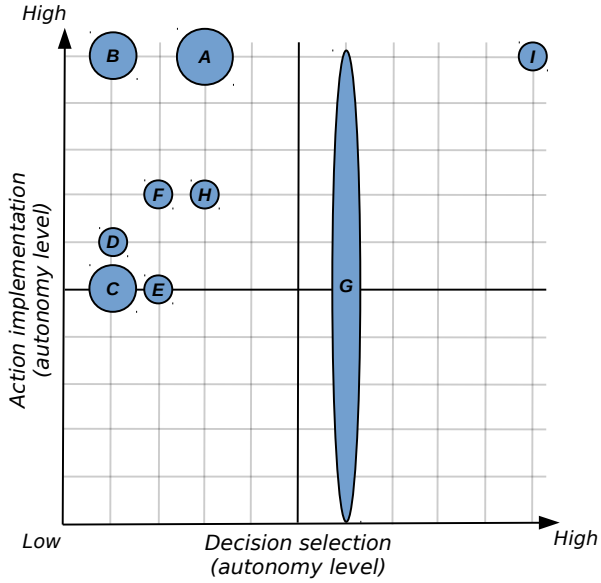


Fig. C.7: Plot of the groups from table (C.6). The size of each circle increases with the number of members in the group. The large span of group *G* signifies the adaptive nature of the action implementation stage for the system by Mülling et al. [10].

when relying entirely on the system to complete the task automatically. The participants did in fact experience a lower level of satisfaction, despite completing the task faster, due to this lack of accomplishment. A similar observation was made by Kim et al. [14] where individuals with movement impairment appeared less inclined to relinquish control of the ARM than able-bodied persons.

A way to address the above issue could be to rely on adaptive semi-autonomous control, as seen in the system by Mülling et al. [10]. Such a scheme will allow the user some control throughout the entire process, thereby providing the user with some sense of accomplishment when finishing a task, while still aiding the user to some extent.

This form of semi-autonomous control can be viewed as an arbitration of control between the system and the user which can essentially be reduced to a linear blending, controlled by the arbitration factor α, as shown in figure (C.8). This is also the approach used by Mülling et al. [10] where α is computed using a sigmoid function dependent on the confidence of the goal predicted by the system.

The idea of viewing the arbitration as a blending problem is based on the work by Dragan et al. [30], which also uses the plots of different arbitration factors α to characterize the behaviour of semi-autonomous systems. An example of different arbitration behaviours is shown in figure (C.9).

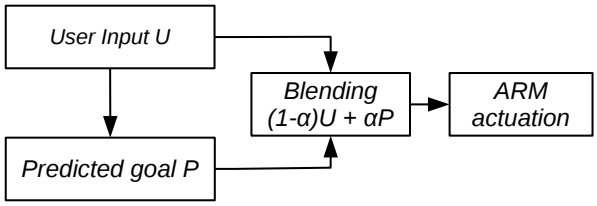
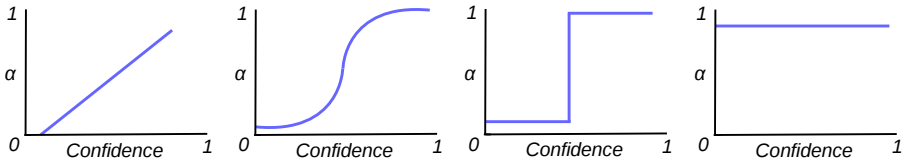


Fig. C.8: Arbitration between the user U and the goal P predicted by system using linear blending. The figure is adapted from [30].

The idea of defining the behaviour of the system using arbitration functions may also make it easier to customize the behaviour of the system to the preference of the user. A lot of different behaviours can be achieved by simply changing the function governing the arbitration factor used during the blending. For instance, the behaviour of the system by Mülling et al. [10] can be characterized by figure (C.9b) whereas the behaviour of e.g. [16, 18] can be characterized by figure (C.9d) as the action implementation stage is always fully automatic for these systems. Customization through these arbitration curves may also be beneficial due to their visual nature making it easier to understand for people with a non-technical background.

However, an important prerequisite for adaptive semi-autonomous control is for the system to be able to gauge its confidence. For instance, how confident the system is that the user is reaching for object B and not object A. One way of inferring confidence in this scenario could be to rely on proximity, i.e. how close is the end-effector of the ARM to each object. Such proximity-based approaches is used in the work of [10, 30, 31]. A possible downside of proximity-based approaches is that they are memory-less, i.e. they only consider the system in its current state. An example of why this lack of memory can be problematic is shown in figure (C.10), where the user is reaching for object A but the system misinterprets the user's goal as being object B due to the proximity-based approach.

A way to introduce memory into the process of inferring the intention of the user is to consider the trajectory of the ARM, as done by [30, 32]. Looking at figure (C.10) again, it is possible to see that considering the trajectories it



(a) Linear function. **(b)** Sigmoid function. **(c)** Unit step function. **(d)** Constant function.

Fig. C.9: Examples of different functions which can be used to control the arbitration factor.

would have been possible to correctly infer that the user was reaching for object A.

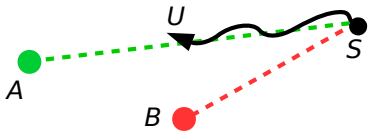


Fig. C.10: A scenario with two objects, A and B, with an ARM denoted by the current position U and initial position S of its end-effector. The figure is adapted from Dragan et al. [30].

The different measures of confidence, mentioned above, are all related to stage 3, i.e. decision selection. However, it is possible to expand the model to make use of confidences derived from the other stages as well. For instance, the confidence of the sensor data gathered during the information acquisition stage. Another idea could be extract a confidence measure from the information analysis stage based on how certain objects are commonly used.

This idea could be achieved through affordance detection, with affordance being the notion that objects "invite" the user to interact with them in certain ways. A handle on a mug would for instance be an obvious affordance for grasping. The idea of grasp detection, as discussed earlier, could hence be considered a limited form of affordance detection, which only focuses on the affordance related to grasping. However, multiple other affordances exist, for instance; cutting, scooping, containing, pounding and supporting. These affordances are the focus in the work by for instance Myers et al. [33], which proposes a way of detecting different affordances using RGB and depth information. An example of their results is shown in figure (C.11).

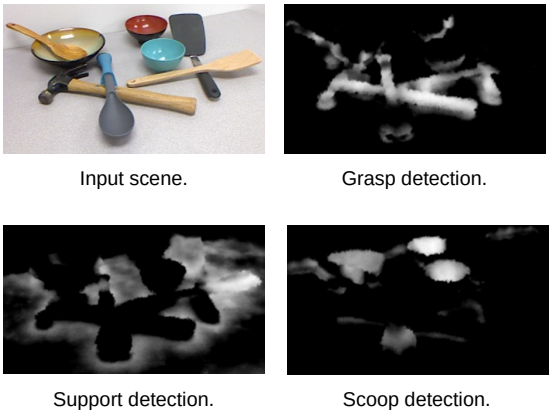


Fig. C.11: Example of affordances detected in a scene. The input scene is from the dataset published by Myers et al. [33].

Affordance detection could hence be useful in scenario where the user is trying to accomplish a task involving multiple objects. For instance, using a spoon to scoop something or when the user wants to pour a liquid into a container.

Yet another possibility is to incorporate the system's confidence in the user, which is suggested by Dragan et al. [30] as well. An interesting addition to this idea could be for the system to provide a level of assistance which keeps the user in a state of flow or "being in the zone". The idea of flow is described as a mental state where the user would feel a sense of mastery and satisfaction by ensuring that the difficulty of the tasks matches the skill of the user [34].

This idea is often illustrated as shown in figure (C.12), where a person is kept in a state of mental flow by matching the difficulty of the task with the skill level of the person. Failure to match these two parameters could cause a person to enter either a state of boredom or anxiety, which is not desirable. The presence of the flow state can hence influence a person's sense of accomplishment and satisfaction which is why it could be interesting to consider it in relation to semi-autonomous control.

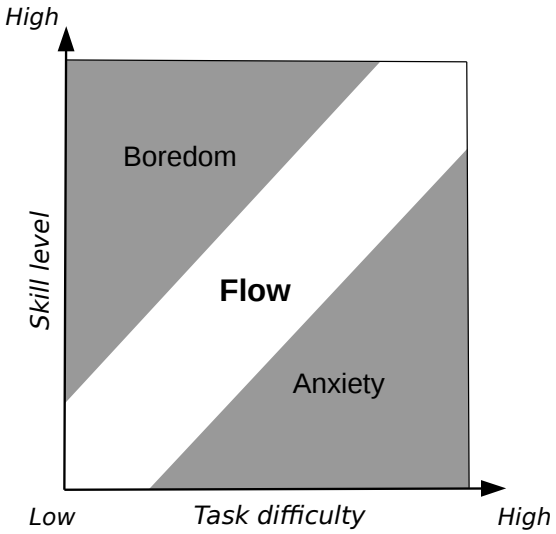


Fig. C.12: Illustration of how flow can be achieved by matching task difficulty and skill level. The figure is adapted from Csikszentmihalyi et al. [34].

4.2 Challenge: Handling Arbitrary Objects

Another challenging aspect of using computer vision for semi-autonomous control of ARMs is to be able to handle arbitrary objects, i.e. objects never

encountered by the system before. Most of the reviewed papers seem to agree that this is an important issue but only a few of them actually address it.

Looking at the reviewed systems, both Gualtieri et al. [22] and Mülling et al. [10] address this issue by discarding the notion of detecting separate objects and instead rely on detecting good grasping points on arbitrary object. However, detecting good grasping points can be rather slow [22] due to the synthesis stage which is time-consuming because of the large search space (six variables; three for grasp position and three for grasp orientation).

A way to speed up this process could be to look into approaches such as [35, 36], which rely on a range image instead of a point cloud data during grasp synthesis. Domae et al. [35] reports a processing time of 0.31 seconds or less, making it significantly faster than Gualtieri et al. [22]. The work by [37, 38] relies on both RGB and range images to infer grasping points using CNNs, with Redmon et al. [37] reporting processing times of $\approx$ 77 milliseconds. The low processing time is likely because a GPU is used to accelerate the computations by taking advantage of the highly parallelizable nature of CNNs.

Another way of handling arbitrary objects is to decompose them into primitive shapes like cylinders, cuboids and spheres, as illustrated in figure (C.13). This is somewhat similar to the idea used by Loconsole et al. [11], which focused on cylinder-shaped objects only. The idea is to expand this approach to encompass any object by including more shapes than just cylinders and by allowing these shapes to be combined [39, 40].

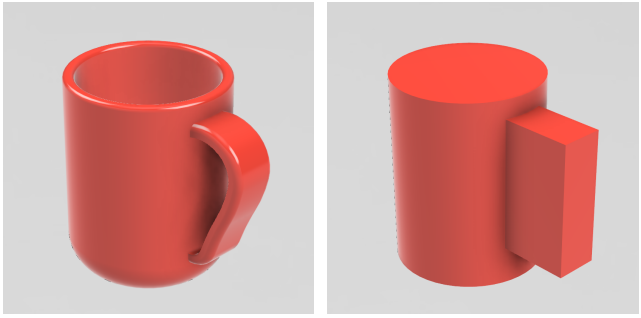


Fig. C.13: Example of an object and its decomposition into primitive shapes. This approach was used by Milleret al. [39].

How to handle arbitrary objects is not necessarily limited to one of the approaches mentioned above. In fact, combining multiple approaches could be a viable solution. An example of such is the work by Ciocarlie et al. [41], which defines a grasping procedure for known objects and a procedure for unknown objects not encountered before. This idea is especially interesting in the scope of semi-autonomous control as an unknown object could be added

to the set of known objects by the user showing the system how to grasp said object. This is somewhat similar to the approach by Herzog et al. [42], where the system learns grasp poses through demonstration by the user. The work of Krainin et al. [43] could also help expand this idea as it describes an approach for creating 3D models of objects once they have been grasped by a robotic manipulator.

4.3 Challenge: Sensing the Environment

The last challenge which this paper will touch upon is how to acquire complete and precise data about the environment that the ARM is to operate in. These aspects are important as the subsequent stages in any system will suffer if the information acquisition stage is not up to par.

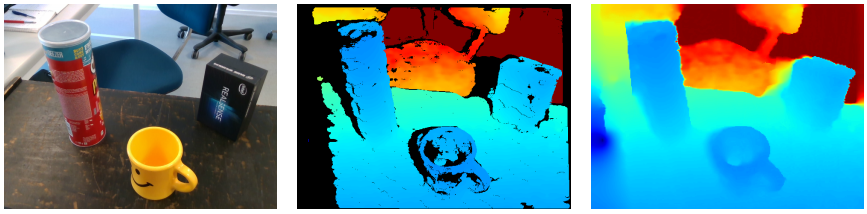
Roughly half of the reviewed papers decided to mount their sensors near or on the end-effector, a configuration sometimes called eye-in-hand. Such as configuration is advantageous as it is near impossible for the ARM to occlude the view of the sensor and it offers some flexibility, as the sensor can be re-positioned using the ARM. Allowing the user to re-position the sensor, by controlling the end-effector, could also make it easier to infer the intention of the users as they would likely orient the end effector towards the object they are interested in. A few of the reviewed papers, see [8, 18], utilize this option by having the system automatically re-positioning the end-effector such that the user's selection is centred in the view of the camera. The idea is to get a better view of the object to interact with.

The work by Gualtieri et al. [22] takes this approach a step further by re-positioning the sensor to gather information from multiple viewpoints in order to increase the quality of the gathered point cloud. The authors specifically states that doing so have shown an improvement in grasp detection according to their prior work [44]. This idea is somewhat similar to the work by Klingensmith et al. [45] where an end-effector mounted depth sensor is used to map the nearby environment using a SLAM-like approach (Simultaneous Localization And Mapping). The authors demonstrate that their approach improves the quality of the data gathered while continuously estimating the position of the end-effector, i.e. the localization part of SLAM. Employing some strategy for accumulating data from multiple viewpoints may hence be beneficial when dealing with an eye-in-hand configuration [22, 45].

Another way to improve upon the depth information acquired by the system could be to use techniques for depth completion [46]. The idea is to use information from a colour image to estimate the missing depth information, as shown in figure (C.14). The colour image is used to estimate surface normals for the entire scene which are then combined with the sparse set of depth measurements to infer depth for the entire scene.

The main drawback of this work is the processing time, as the authors

states a processing time of between 0.3 and 1.5 seconds, depending on the hardware used. It can be argued that the processing time is not a big issue as it may not be necessary to use depth completion on every single frame received from the sensor. However, a benefit of this approach is that it will work with sensors mounted in a fixed position, as opposed to the SLAM-like approaches mentioned earlier.



(a) Color information. **(b)** Depth information. **(c)** Completed depth.

Fig. C.14: Example of depth completion using the algorithm from Zhang and Funkhouser [46].

An area which could also improve the system's ability to sense its immediate environment is the actual sensors employed by each system. Most of the reviewed systems rely on stereo vision to gather depth information and it may hence be interesting to explore other options such a ToF cameras like the Kinect v2 used by [23, 24]. The absence of ToF cameras amongst the other reviewed paper can likely be attributed to the available ToF cameras at the time, which were likely expensive and bulky.

A general difference between ToF cameras and stereo vision cameras is that the former does not rely on a baseline to estimate depth. ToF cameras can hence be made more compact, making it possible to mount them in location not possible for stereo vision cameras. For instance, inside the gripper of the ARM. An example of such is the CamBoard Pico flexx camera from PMD Technologies [47] which is significantly smaller than the Kinect sensors while featuring a minimum working distance of 0.1 m.

5 Conclusion

The focus of this review paper was on computer vision systems enabling movement impaired individuals to do object manipulation using an assistive robotic manipulator (ARM). The initial literature search yielded 257 results which were narrowed down to 14 relevant papers. These papers where reviewed in relation to their selection of hardware and their use of computer vision for the semi-autonomous behaviour of the system. Different schemes for the semi-autonomous control were reviewed as well. A four-stage model was used during the review of each system to characterize their behaviours

in terms of; information acquisition, information analysis, decision selection and action implementation. A scale, consisting of 10 levels [15], was used to rate the autonomy of each system as well.

The reviewed papers mainly made use of stereo vision-based sensors to capture depth information. Many of the papers used the Kinect from Microsoft which was often mounted near the shoulder or head of the user, viewing the scene from a distance. The second most popular sensor placement was at the end-effector, making it possible for both the user and the system to re-position the sensor. However, only a few of the reviewed systems fully utilized this option by mapping the immediate environment using data from multiple viewpoints. Furthermore, exploring other options in terms of sensor choice may be interesting as well. For instance, a small ToF camera which could be mounted inside the gripper of the ARM.

Handling of arbitrary objects was found to be a general issue with only a few of the reviewed systems being able to do so. The majority made simplifying assumptions such as all objects having a certain shape or all object being in a database of pre-defined objects. A way to approach the issue of handling arbitrary objects could be to reduce it to a problem of detecting good grasping points or decomposing objects into primitive shapes.

Most of the reviewed papers rely on a clear switch between the user and the system for the semi-autonomous control of said system. Adaptable automation, in the form of linear blending, is used in one of the reviewed papers but should be explored further. Such a scheme could be beneficial as it allows the user some control at all times which is especially important for movement impaired users. A scheme based on linear blending may also allow for easy customization of the semi-autonomous control. Such an adaptive approach may also benefit from the concept of flow, known from psychology, to adjust the level of assistance based on the skill level of the user and the difficulty of the task at hand.

To summarize; there is a substantial amount of on-going research focusing on using computer vision for semi-autonomous control of ARMs. Several working prototypes have demonstrated that this idea can work in a controlled environment, such as a lab. The next big step is to advance the technology to a point where it is possible to move beyond the labs and into the home of the actual user. The benefit of doing so would be priceless for the individual user, and society in general may benefit as well due to less demand for caregivers.

References

[1] K. M. Marasinghe, "Assistive technologies in reducing caregiver burden among informal caregivers of older adults: a systematic review," *Dis-*

ability and Rehabilitation: Assistive Technology, vol. 11, no. 5, pp. 353–360, 2016.

[2] S. Bedaf, P. Marti, F. Amirabdollahian, and L. de Witte, "A multi-perspective evaluation of a service robot for seniors: the voice of different stakeholders," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 6, pp. 592–599, 2018.

[3] G. Romer, H. Stuyt, and A. Peters, "Cost-savings and economic benefits due to the assistive robotic manipulator (arm)," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 201–204.

[4] V. Maheu, P. S. Archambault, J. Frappier, and F. Routhier, "Evaluation of the jaco robotic arm: Clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics*, June 2011, pp. 1–5.

[5] S. D. Prior, "An electric wheelchair mounted robotic arm a survey of potential users," *Journal of Medical Engineering & Technology*, vol. 14, no. 4, pp. 143–154, 1990.

[6] KINOVA, "Robotic arms series," visited on 05/20/2019. [Online]. Available: https://www.kinovarobotics.com/en/products/robotic-arm-series

[7] Exact Dynamics, "iARM," visited on 05/20/2019. [Online]. Available: http://www.exactdynamics.nl/site/?page=iarm

[8] D. J. Kim, R. Lovelett, and A. Behal, "An empirical study with simulated adl tasks using a vision-guided assistive robot arm," in *2009 IEEE International Conference on Rehabilitation Robotics*, June 2009, pp. 504–509.

[9] H. Jiang, J. P. Wachs, and B. S. Duerstock, "Integrated vision-based robotic arm interface for operators with upper limb mobility impairments," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, June 2013, pp. 1–6.

[10] K. Muelling, A. Venkatraman, J.-S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to brain computer interface controlled manipulation," *Autonomous robots*, vol. 41, no. 6, pp. 1401–1422, 2017.

[11] C. Loconsole, F. Stroppa, V. Bevilacqua, and A. Frisoli, "A robust real-time 3d tracking approach for assisted object grasping," in *Haptics: Neuroscience, Devices, Modeling, and Applications*, M. Auvray and C. Duriez, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 400–408.

[12] V. W. Oguntosin, Y. Mori, H. Kim, S. J. Nasuto, S. Kawamura, and Y. Hayashi, "Design and validation of exoskeleton actuated by soft modules toward neurorehabilitation-vision-based control for precise reaching motion of upper limb," *Frontiers in neuroscience*, vol. 11, pp. 352–352, Jul 2017.

[13] C.-S. Chung, H. Wang, and R. A. Cooper, "Functional assessment and performance evaluation for assistive robotic manipulators: Literature review," *The Journal of Spinal Cord Medicine*, vol. 36, no. 4, pp. 273–289, 2013.

[14] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[15] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[16] B. Pitzer, M. Styer, C. Bersch, C. DuHadway, and J. Becker, "Towards perceptual shared autonomy for robotic mobile manipulation," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 6245–6251.

[17] A. Remazeilles, C. Leroux, and G. Chalubert, "Sam: A robotic butler for handicapped people," in *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, Aug 2008, pp. 315–321.

[18] D. J. Kim, Z. Wang, and A. Behal, "Motion segmentation and control design for ucf-manus;an intelligent assistive robotic manipulator," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 5, pp. 936–948, Oct 2012.

[19] M. Elarbi-Boudihir and K. A. Al-Shalfan, "Eye-in-hand/eye-to-hand configuration for a wmra control based on visual servoing," in *2013 IEEE 11th International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics*, June 2013, pp. 1–6.

[20] C. P. Quintero, O. Ramirez, and M. Jägersand, "Vibi: Assistive vision-based interface for robot manipulation," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 4458–4463.

[21] H. Jiang, T. Zhang, J. P. Wachs, and B. S. Duerstock, "Enhanced control of a wheelchair-mounted robotic manipulator using 3-d vision and multimodal interaction," *Computer Vision and Image Understanding*, vol. 149, pp. 21–31, Aug. 2016.

[22] M. Gualtieri, J. Kuczynski, A. M. Shultz, A. Ten Pas, R. Platt, and H. Yanco, "Open world assistive grasping using laser selection," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4052–4057.

[23] Z. Zhang, Y. Huang, S. Chen, J. Qu, X. Pan, T. Yu, and Y. Li, "An intention-driven semi-autonomous intelligent robotic system for drinking," *Frontiers in Neurorobotics*, vol. 11, p. 48, 2017.

[24] F. Arrichiello, P. D. Lillo, D. D. Vito, G. Antonelli, and S. Chiaverini, "Assistive robot operated via p300-based brain computer interface," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 6032–6037.

[25] H. Zeng, Y. Wang, C. Wu, A. Song, J. Liu, P. Ji, B. Xu, L. Zhu, H. Li, and P. Wen, "Closed-loop hybrid gaze brain-machine interface based robotic arm control with augmented reality feedback," *Frontiers in Neurorobotics*, vol. 11, Oct. 2017.

[26] Microsoft, "Kinect Sensor," visited on 05/20/2019. [Online]. Available: https://msdn.microsoft.com/en-us/library/hh438998.aspx

[27] A. Frisoli, F. Salsedo, M. Bergamasco, B. Rossi, and M. C. Carboncini, "A force-feedback exoskeleton for upper-limb rehabilitation in virtual reality," *Applied Bionics and Biomechanics*, vol. 6, no. 2, pp. 115–126, 2009.

[28] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381395, jun 1981.

[29] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[30] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[31] J. E. Downey, J. M. Weiss, K. Muelling, A. Venkatraman, J.-S. Valois, M. Hebert, J. A. Bagnell, A. B. Schwartz, and J. L. Collinger, "Blending of brain-machine interface and vision-guided autonomous robotics

improves neuroprosthetic arm performance during grasping," *Journal of NeuroEngineering and Rehabilitation*, vol. 13, no. 1, p. 28, 2016.

[32] C. Schultz, S. Gaurav, M. Monfort, L. Zhang, and B. D. Ziebart, "Goal-predictive robotic teleoperation from noisy sensors," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[33] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features." in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1374–1381.

[34] M. Csikszentmihalyi, *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Springer, 8 2014.

[35] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1997–2004.

[36] E. Klingbeil, D. Rao, B. Carpenter, V. Ganapathi, A. Y. Ng, and O. Khatib, "Grasping with application to an autonomous checkout robot," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 2837–2844.

[37] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[38] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[39] A. Miller, S. Knoop, H. Christensen, and P. Allen, "Automatic grasp planning using shape primitives," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, 2003, pp. 1824–1829 vol.2.

[40] K. Huebner, S. Ruthotto, and D. Kragic, "Minimum volume bounding box decomposition for shape approximation in robot grasping," in *2008 IEEE International Conference on Robotics and Automation*, May 2008, pp. 1628–1633.

[41] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan, "Towards reliable grasping and manipulation in household environments," in *Experimental Robotics*. Springer Berlin Heidelberg, 2014, pp. 241–252.

[42] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal, "Template-based learning of grasp selection," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 2379–2384.

[43] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and object tracking for in-hand 3d object modeling," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1311–1327, 2011.

[44] M. Gualtieri, A. ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 598–605.

[45] M. Klingensmith, S. S. Sirinivasa, and M. Kaess, "Articulated robot motion for simultaneous localization and mapping (ARM-SLAM)," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 1156–1163, 2016.

[46] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-d image," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 175–185.

[47] PMDTec, "CamBoard Pico Flexx," visited on 05/20/2019. [Online]. Available: https://pmdtec.com/picofamily

Paper D

Computer Vision-Based Adaptive Semi-Autonomous Control of an Upper Limb Exoskeleton for Individuals with Tetraplegia

Stefan Hein Bengtson, Mikkel Berg Thøgersen, Mostafa Mohammadi, Frederik Victor Kobbelgaard, Muhammad Ahsan Gull, Lotte N. S. Andreasen Struijk, Thomas Bak, and Thomas B. Moeslund

The paper has been published in
Applied Sciences, vol. 12, no. 9, 4374, 2022.

The layout has been revised.

Abstract

We propose the use of computer vision for adaptive semi-autonomous control of an upper limb exoskeleton for assisting users with severe tetraplegia to increase independence and quality of life. A tongue-based interface was used together with the semi-autonomous control such that individuals with complete tetraplegia were able to use it despite being paralyzed from the neck down. The semi-autonomous control uses computer vision to detect nearby objects and estimate how to grasp them to assist the user in controlling the exoskeleton. Three control schemes were tested: non-autonomous (i.e., manual control using the tongue) control, semi-autonomous control with a fixed level of autonomy, and a semi-autonomous control with a confidence-based adaptive level of autonomy. Studies on experimental participants with and without tetraplegia were carried out. The control schemes were evaluated both in terms of their performance, such as the time and number of commands needed to complete a given task, as well as ratings from the users. The studies showed a clear and significant improvement in both performance and user ratings when using either of the semi-autonomous control schemes. The adaptive semi-autonomous control outperformed the fixed version in some scenarios, namely, in the more complex tasks and with users with more training in using the system.

1 Introduction

The use of robotics for aiding people in everyday life is becoming increasingly more common and assistive robotics is a promising aspect for individuals who would otherwise be dependent on a caregiver for daily activities [1, 2]. This potential for assistive robots is especially high for individuals who are exceedingly dependent on caregivers in everyday life such as individuals with tetraplegia, i.e., partial or complete loss of control in both arms and legs. A typical cause of tetraplegia is SCI (spinal cord injury), and it is estimated that 250,000 to 500,000 people suffer from SCI every year [3], with roughly one-third of these cases resulting in tetraplegia [4].

The mean age of individuals with tetraplegia was reported to be 33 years old when sustaining the injury [4], but especially males aged 20–29 years and females aged 15–19 years have seen spikes in the incident rate of traumatic spinal cord injuries [3]. The motivation for focusing on tetraplegia is further supported by a high life expectancy after the injury, especially for young individuals [3]. An individual with tetraplegia caused by SCI at the age of 25 years can often expect to live for another 40 years. One of the most important tools for individuals with tetraplegia is often a powered wheelchair, offering both mobility and independence [3]. However, tasks requiring interaction with an object, such as drinking and eating, still require assistance from a caregiver. The frustration of not being able to perform these things inde-

pendently becomes even greater for prolonged sessions, for instance, during recreational activities, such as watching a movie while enjoying a beverage or snacking [5].

To mitigate the challenges above and increase the independence of individuals with tetraplegia, we propose the novel combination of using a tongue-based interface combined with computer vision in a semi-autonomous control scheme to enable individuals with tetraplegia to effectively control an upper limb exoskeleton, as shown in Figure D.1.

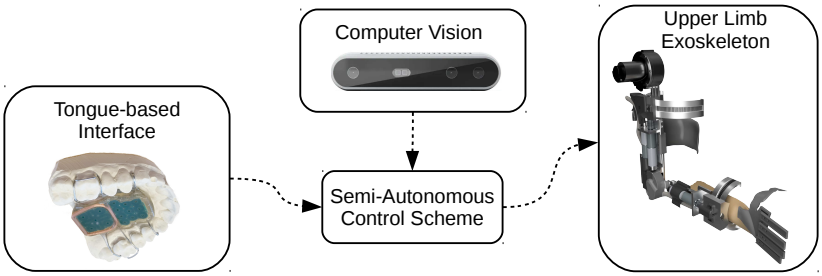


Fig. D.1: Overview of the proposed system. A semi-autonomous control scheme combines input from a tongue-based interface and a computer vision module to assist a user in controlling an upper limb exoskeleton.

An upper limb exoskeleton is used, as several studies have successfully demonstrated how it can help restore some of the lost functionality for individuals with movement impairments in the arms [6, 7]. Furthermore, the tongue-based interface has previously been shown to be an efficient and suitable interface for individuals with tetraplegia to control an exoskeleton [8, 9]. Finally, a semi-autonomous control scheme assists the user in controlling the exoskeleton based on input from a computer vision module, which detects and analyzes nearby objects. This is performed as several studies suggest that such an approach can be beneficial for controlling assistive robotic manipulators [10].

Hence, the main contributions of the paper are the following:

- We design and implement an adaptive semi-autonomous control scheme based on computer vision and evaluate it in the context of controlling an upper limb exoskeleton through a tongue-based interface.
- We evaluate the effectiveness and intuitiveness of various control schemes for performing semi-autonomous tongue-control of an upper limb exoskeleton, through studies including both participants with and without tetraplegia.

2 Related Work

The proposed system consists of an upper limb exoskeleton as it would enable individuals with tetraplegia to regain some of the lost functionality in their arms. The idea of upper limb exoskeletons for users with movement impairments have, in several previous studies, been shown to be useful for both assistive purposes [6, 11] and for rehabilitation [7, 12, 13].

However, a challenging aspect of using an exoskeleton for people with tetraplegia is how to interface with it. A common approach for controlling upper limb exoskeletons is with EMG (electromyography) [7, 14] to detect muscle activity, which is not feasible in case of severe tetraplegia. Other approaches require the user to control the exoskeleton using a joystick, operated by a single finger [6], which is not possible for complete functional tetraplegia either. Others have explored the idea of using eye movements [13] or voice commands [6] to allow individuals with movement impairments to interface with an upper limb exoskeleton. Eye movements or voice commands are plausible options but can be very tedious to use in the long run and are easily susceptible to noise such as accidental eye movements or nearby sounds. In one study, the majority of the participants preferred other options over the voice-based control [6]. Another possibility is BCI-based control (brain–computer interface), where signals are measured from the brain of the user controlling the upper limb exoskeleton [12]. However, weak points of a BCI-based interface is the low signal-to-noise ratio, the need for substantial calibration, and the low throughput, both in terms of the number of different commands and also how fast one can issue them. This often restricts the use of BCI-based interfaces to rely on predefined movements which are completely automated [15], forcing the user to relinquish control completely for periods of time.

A tongue-based interface does not suffer from many of the issues highlighted above. It offers high throughput, both in terms of the number of possible commands and also in terms of how fast they can be issued [16]. Furthermore, several studies have demonstrated that a tongue-based interface can be used by individuals with tetraplegia to control various assistive devices, such as an upper limb exoskeleton [8] or a robotic arm [17]. These considerations have led to the choice of a tongue-based interface.

A semi-autonomous control scheme, where parts of the control are automated using computer vision, is included in the system to further enhance the tongue-based control of the exoskeleton. Several studies have reported increased performance when employing computer vision for semi-autonomous control of assistive robotic manipulators [10], such as completing tasks faster [18, 19] or being more precise in the movement of the manipulator [20, 21]. Furthermore, there are several examples of computer vision

either improving the fine control of an upper limb exoskeleton [13, 21] or automating entire parts of a task [11, 22, 23] for users with paralysis.

Many of the approaches relying on computer vision employ a clear-cut strategy for arbitrating control between the user and the system, where certain parts of the process are completely automated [10]. For example, reaching for and grasping an object once the user triggers this predefined task, either from a tongue-based interface [20], through voice commands [18], or through eye movements [11, 22]. The user would, in these cases, relinquish complete control until the task is completed, i.e., the object is reached and grasped. This fixed level of autonomy, where the automated process is clearly defined, is likely common because it is easy to implement, easy to understand for the user, and it improves performance in many cases. Sometimes it is also the only option due to the limitations of the interface used for the control [11, 22].

However, using a fixed level of autonomy introduces the problem of finding an optimal balance in the arbitrating control between the user and the system [10]. If the user is primarily in control at all times, without any automating, it defeats the purpose of having semi-autonomous control in the first place. On the other hand, a high level of autonomy where nearly everything is automated may not be a satisfying experience for the user either [24]. Automating may even counteract what the user is trying to achieve in cases where the automating performs the wrong action, e.g., reaching for the wrong object, and it may impose a safety risk. Even cases where the automating acts as intended may result in lower satisfaction for the user as they may no longer feel in control, especially for individuals with movement impairments [19].

One way to avoid or minimize many of these issues is to rely on an adaptive level of autonomy instead of a fixed one. Several studies on teleportation of robots have successfully demonstrated semi-autonomous control with an adaptive level of autonomy based on a confidence-measure [25, 26]. This confidence-measure is an expression of how certain the system is in its own prediction of the intent of the user, such as interacting with a certain object. The system will hence offer a lot of assistance in scenarios where it has a high confidence of being able to correctly assist the user. The opposite is also true; the system will offer no or little assistance in cases of low confidence where it is unclear what the user is trying to accomplish. A benefit is hence that it can adapt its level of autonomy to fit different scenarios.

Hence, in the current study, three different control schemes were implemented: a non-autonomous (i.e., manual) control, a semi-autonomous control with a fixed level of autonomy, and a semi-autonomous control using a confidence-based adaptive level of autonomy. The implementation of each is described in further detail in Section 3.4. These three different control schemes are evaluated and compared against each other, as described in

Section 4. The purpose is to determine whether using computer vision for tongue-based control of an upper limb exoskeleton is beneficial or not and whether semi-autonomous control with a fixed or adaptive level of autonomy is preferable in this context.

3 Method

The following describes the different main components of the proposed system, as also shown in Figure D.1. The upper limb exoskeleton is controlled by mixing input from the user and from the computer vision module. The user provides input to the system through a tongue-based interface as the exoskeleton is designed for individuals with tetraplegia. The computer vision module is designed to detect objects in front of the user and infer how to grasp them. The computer vision module is also designed to predict the intention of the user, i.e., what object to grasp, to assist the user in controlling the exoskeleton. Finally, the control scheme module combines input from the user and the computer vision module to actuate the exoskeleton.

3.1 Upper Limb Exoskeleton

The exoskeleton used in this study is the EXOTIC upper limb exoskeleton, as shown in Figure D.2a. It has four degrees of freedom (DoFs): two in the shoulder, one in the elbow, and one in the wrist. The number of DoFs is kept at a minimum to reduce the bulkiness of the exoskeleton. The DoFs included have been carefully selected to support tasks such as picking up objects from a table and bringing them to the mouth of the user. The number of DoFs in the exoskeleton also means that it must use three of its four DoFs to reach an arbitrary position, leaving only rotation around the wrist as the free DoF for altering the orientation of the end effector, i.e., the hand. A more thorough description of the upper limb exoskeleton and its capabilities can be found in [27, 28].

For the end effector, a Carbonhand from Bioservo Technologies AB is used and provides active actuation when closing the hand of the user. Only the thumb, middle finger, and ring finger are actuated in the Carbonhand. Opening of the hand is passive and is performed using an elastic fabric on the back of the hand.

3.2 Tongue-Based Interface

The proposed system makes use of an inductive intra-oral tongue interface (ITCI) [9, 29], as shown in Figure D.2b. The ITCI sits in the roof of the user's mouth and is held in place similar to a dental brace. The unit contains a

(a) The EXOTIC upper limb exoskeleton with a camera at the shoulder joint.

(b) The inductive tongue interface (ITCI).

(c) The control layout for the ITCI.

Fig. D.2: An overview of the hardware used in the proposed system. (**a**) The EXOTIC upper limb exoskeleton along with the Carbonhand for the end effector. An RGB-D camera is mounted at the shoulder joint. (**b**) The part of the inductive tongue interface (ITCI) placed in the roof of the user's mouth. (**c**) The layout of the ITCI used to control the exoskeleton.

small battery and can hence operate wirelessly while sitting in the mouth of the user. The entire area of the ITCI is covered by 18 small inductive sensors which can be activated using a tongue piercing made of metal. The tongue piercing and dental braces are for long-term usage and not for temporary usage. In the studies, surgical glue was used to attach a small piercing-like metal cylinder on the tongue of the participants instead. The ITCI was held in place in the roof of the participant's mouth using dental putty instead of a custom dental brace.

The inductive sensors of the ITCI were mapped to the layout shown in Figure D.2c to control the upper limb exoskeleton [28]. The upper part of the ITCI functions as a joystick where the user can control the forward (F), backward (B), left (L), and right (R) motion of the exoskeleton. A slider is used to control the exoskeleton either down or up and is placed right below the joystick-like control. All these movements happen in relation to the current position of the end-effector, i.e., the hand of the user. The lower part of the ITCI contains controls for opening and closing the Carbonhand along with a slider for controlling the wrist rotation of the exoskeleton. Finally, an "auto grasp" button is located slightly below the middle of the ITCI layout. When pressed, the button will activate the fixed semi-autonomous control of the exoskeleton which is described in more detail later in Section 3.4. The exact same layout is used for both the manual control and the adaptive semi-autonomous control. The only exception is that the auto grasp button does nothing while using these two control schemes.

The layout of the ITCI, along with the current location of the tongue pierc-

ing in the mouth, is shown to the user only when training to use the tongue interface. All the results presented later are hence gathered without any visual feedback from the tongue interface. This is performed as the main idea of the system is to be able to use it without any visual feedback besides the actual movement of the exoskeleton.

3.3 Computer Vision Module

The computer vision part of the system is mainly responsible for performing object detection, intent prediction, and grasp detection, as outlined in Figure D.3. The input for the computer vision module is a small RGB-D camera (Intel RealSense D415), providing both color and depth information, mounted at the shoulder joint of the exoskeleton (see Figure D.2a) and pointing towards the area in front of the user. This specific camera was chosen for its small baseline, i.e., the distance between the two sensors used for depth measurements, making it suitable for capturing depth data at close range (minimum operating range ≈30 cm). The small baseline also results in a small camera footprint, making it easier to mount on the exoskeleton discreetly and without the camera getting in the way. The depth information from the camera is not used during object detection but it is used for both the intent prediction and grasp detection, as described later.

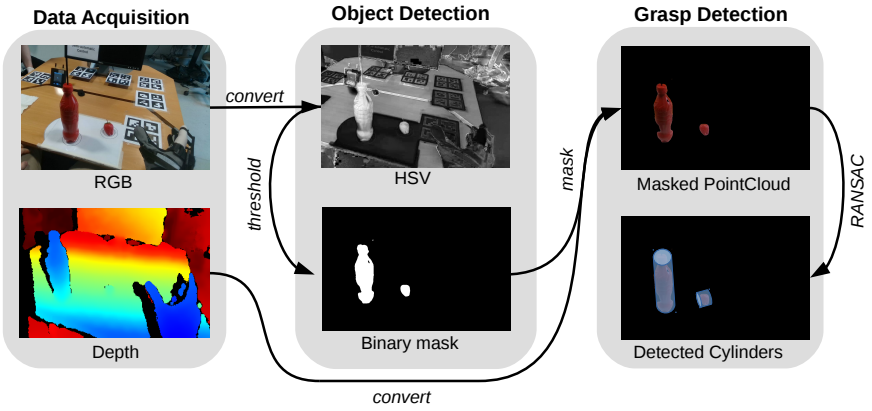


Fig. D.3: Overview of the pipeline for the computer vision module. An RGB-D camera (Intel RealSense D415) is mounted at the shoulder joint of the exoskeleton and captures both RGB and depth information from the area in front of the user. The object detection relies on the RGB data where objects are detected using color thresholding. The depth information is masked based on the detected objects and then converted to a point cloud. Cylinder-like shapes are then detected in the resulting masked point cloud using an RANSAC-based algorithm. Finally, the detected cylinders are converted to grasp poses for the exoskeleton using a rule-based approach.

Object Detection

An important part of the computer vision module is to be able to detect any objects of interest in front of the user that the exoskeleton might be able to reach. The current state-of-the-art approaches for object detection are often based on deep learning [30, 31], where neural networks are trained on huge amounts of labeled data [32]. These huge amounts of training data are required for the deep learning-based object detectors to learn a wide range of different objects and to generalize well to different environments. However, performing object detection in this way adds another layer of complexity, and thereby uncertainty, on top of an already complex system.

Instead, a classic approach of relying on color for segmentation of the objects is used, where thresholding is applied to the HSV (hue, saturation, value) color space, such that all bright red objects in the RGB image from the camera are detected. This approach is characterized by producing stable object detections in a controlled environment for a few objects, which is what is needed for the experiment. The decision to use this classic approach based on colors for object detection is hence an attempt at minimizing any uncertainty during the experiment related to object detection. This decision was deemed acceptable as the focus is not the computer vision part but rather on testing the different control schemes.

Intent Prediction

Once any objects are detected, the system must predict the intent of the user. The intent prediction is based on the direction of the user's palm, the intuition being that people generally have their palm pointed towards an object when grasping it. Using this assumption also had the benefit of making it easy to explain how the system works to the users of the exoskeleton.

The intent prediction works by projecting a ray from the palm of the user, as depicted in Figure D.4. The orthogonal Euclidean distance between this ray $\vec{R}$ and any object in the scene is then calculated and the object resulting in the shortest distance is then considered the predicted intention of the user, i.e., object A. Additionally, only objects facing the palm of the user's hand are considered during the intent prediction and any objects facing the back of the hand are ignored, such as object C.

An alternative to this ray-based method could have been an approach based purely on distance [26], i.e., finding the nearest object. The obvious drawback of such an approach is that it would only consider the nearest object, i.e., B, and disregard all other objects in the scene. Furthermore, using solely the shortest distance may also cause the intent prediction to gravitate towards the nearest object to the point where it might become difficult for the user to break away from that object. This effect is less pronounced with the

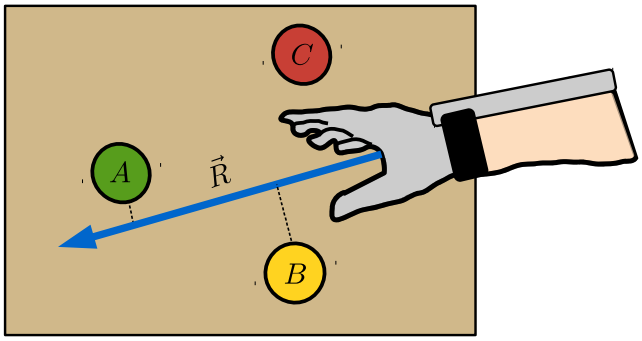


Fig. D.4: Example of the intent prediction using the Euclidean distance from the detected objects (A, B, C) to a ray $\vec{R}$ projected from the palm of the user's hand in the exoskeleton. The example is shown in a plane (2D) for simplification.

ray-based method as only slight adjustments of the exoskeleton are required to point it towards the intended target. Finally, the ray-based approach for intent prediction was found to be quite stable when moving towards an object. This behavior avoids problematic scenarios where the system might suddenly change the predicted intention while moving towards an object; something that could easily occur if using a nearest object approach for the intention prediction.

Grasp Detection

The last part of the computer vision pipeline detects how to grasp the target object, i.e., the object that the user is interested in interacting with. The experimental setup included two objects for the user to interact with: a strawberry and a bottle. A simple rule-based strategy relying on fitting cylinders [23] was hence used for the grasp detection. First, the detection from earlier (Section 3.3) was used to mask the depth information from the RGB-D camera such that the result was a point cloud of the target object, as shown in Figure D.3. An RANSAC-based algorithm [33] was then used to fit a cylinder to the masked point cloud [34], resulting in both the position and orientation of the object (the central axis of the cylinder) along with its approximate size (the cylinder diameter and height).

The fitted cylinders were then converted into a grasp pose for the exoskeleton, as illustrated in Figure D.5, which depicts how the coordinate frame of end-effector, i.e., the Carbonhand, should be positioned and oriented in order to grasp the two objects on the table. The frame of the end-effector is placed in the palm of the Carbonhand and oriented such that the z-axis is pointing out from the palm, the y-axis points upwards, and the x-axis points towards the thumb of the Carbonhand.

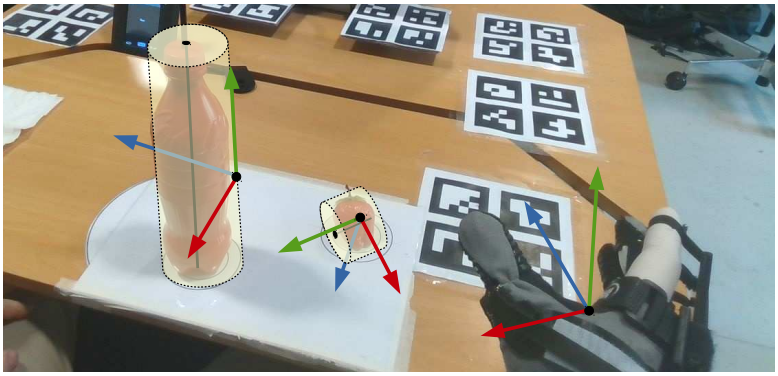


Fig. D.5: The grasp detection identifies how the coordinate frame of the Carbonhand (i.e., the end-effector) should be positioned and oriented in order to grasp the two objects on the table.

A rule-based approach is used for grasping the objects such that:

- **Position**—The coordinate frame of the end-effector should be positioned halfway along the height of the cylinder, such that the object is grasped in the middle for stability. Furthermore, the position for grasping the object should be on the outer perimeter of the cylinder to avoid pushing the object away. An offset, equal to the radius of the detected cylinder, is applied in the direction towards the end-effector to avoid this.
- **Orientation**—The coordinate frame of the end-effector should be oriented such that the y-axis is parallel with the axis of the cylinder, while also pointing upwards to avoid infeasible grasping orientations (such as trying to grasp the object with the palm of the hand facing away from the person in the exoskeleton). Furthermore, the z-axis should be orthogonal to the axis of the cylinder to avoid grasping the object at a skewed angle.

Furthermore, smaller objects (height less than $\approx$3 cm) are difficult to grasp using an upright orientation (such as the orientation of the Carbonhand depicted in Figure D.5). This is partly due to the design of the Carbonhand, where only the thumb, middle finger, and ring finger are actuated. An additional check is therefore implemented in the grasp detection, such that all objects fewer than 3 cm in height will be approached as a cylinder laying flat on the table, as also illustrated for the small strawberry in Figure D.5. Finally, it should be noted that a cylinder is a poor fit for a strawberry. Nevertheless, the above approach was found to produce an acceptable grasp for both the bottle and the strawberry.

3.4 Control Schemes

The purpose of the control schemes is to arbitrate input from the user, received through the ITCI, with the information from the computer vision module, i.e., what object to grasp and how to grasp it, in order to actuate the upper limb exoskeleton. Parts of the control of the exoskeleton will hence be automated, which is why some of the control schemes are referred to as semi-autonomous control.

Three different control schemes are implemented and tested against each other:

- **Non-Autonomous Control** —The system offers no assistance at any point and the input from the computer vision module is ignored. The exoskeleton is manually controlled by the user at all times.
- **Fixed Semi-Autonomous Control**—A fixed level of autonomy is used where the system will take over control of the exoskeleton when the user presses and holds the "auto grasp" button in the ITCI layout (Figure D.2c). While doing so, input from the computer vision module guides the hand of the exoskeleton towards the most likely object to grasp.
- **Adaptive Semi-Autonomous Control**—The system will at all times assist the user in controlling the exoskeleton. The level of autonomy is adapted based on a confidence measure related to the certainty of the intent prediction from the computer vision module. A high certainty of the predicted intention being correct will result in a high confidence and the system will provide more assistance. In low-confidence scenarios, the opposite is true, and the system will provide little to no assistance. If the user does not activate the tongue interface the system does not move. The "auto grasp" button does nothing in this control scheme.

Fixed Semi-Autonomous Control

The fixed scheme for the semi-autonomous control switches from manual to automatic control as long as the user presses and holds the "auto grasp" button on the ITCI layout, as shown earlier in Figure D.2c. Having to press and hold the button instead of only pressing the button once is a safety measure as it provides an intuitive and easy way to stop the automatic control of the exoskeleton by simply letting go of that button. It also reduces the impact of random noise activating the "auto grasp" button or the user activating it by mistake. Both scenarios can easily occur, especially when learning to use the ITCI.

Once the automatic control is activated, the exoskeleton will move towards the target object as detected by the computer vision module. The ex-

oskeleton will move the hand towards the detected grasp pose linearly while controlling both the position and orientation. While using the automatic control, the exoskeleton will avoid collisions with the table as an added safety measure. The actuation of the hand open/close of the Carbonhand is not part of the automatic control and will have to be activated manually by the user.

Adaptive Semi-Autonomous Control

The adaptive scheme for the semi-autonomous control relies on continuously blending input from the user and input from the computer vision module based on a confidence measure. This confidence measure is based on calculating the similarity of the command received from the user with the intention predicted by the computer vision module.

All commands from the user for manually moving the exoskeleton (up, right, forward, and so on) can be described in 3D using the vector $\vec{U}_{x,y,z}$. As only the direction of the user input is considered for calculating the confidence, the normalized vector, $\hat{U}$, is used. A similar vector can be formulated for the computer vision module, $\vec{C}_{x,y,z}$, which describes how the position of the exoskeleton's end-effector should change in order to reach the grasp pose for the predicted target. The direction of this vector from the computer vision module is expressed as the normalized vector, $\hat{C}$. The confidence is then measured using the scalar product between these two normalized direction vectors:

$$\rho(\hat{U}, \hat{C}) = \frac{(\hat{C} \cdot \hat{U}) + 1}{2}. \tag{D.1}$$

The addition and division of the scalar product between the two direction vectors serves to normalize the resulting scalar to a value between 0 and 1. A high confidence indicates that the user and the computer vision want to move in the same direction and vice versa. A confidence measure of 0.5 or below corresponds to an angle of 90^o or more between the direction from the user input, $\hat{U}$, and the direction from the computer vision module, $\hat{C}$.

The adaptive semi-autonomous control relies on the above confidence measure to arbitrate the control of the exoskeleton between the user and the computer vision module. This arbitration is performed using a linear blending [26] between the input from the user and the input from the computer vision module:

$$\vec{E}_{x,y,z,\theta} = (1 - \alpha)\vec{U}_{x,y,z,\theta} + \alpha\vec{C}_{x,y,z,\theta} \tag{D.2}$$

where α is the arbitration factor controlling the blending, such that the user is completely in control of the exoskeleton when $\alpha = 0$, and vice versa. $\vec{E}_{x,y,z,\theta}$ is the output used to actuate the exoskeleton in terms of a linear velocity x, y, z and an angular velocity θ for the wrist rotation. Only the angular movement around the wrist is considered in the above blending because the user is

limited to wrist rotation only, as explained earlier in Section 3.1. Lastly, $\vec{E}$, $\vec{U}$, and $\vec{C}$ are velocity vectors and hence not normalized.

Finally, the arbitration factor α for the adaptive semi-autonomous control is dependent on the confidence measure ρ, as shown in Figure D.6. The main property of the selected arbitration curve was to ensure that the confidence reached an acceptable level before providing any assistance. No assistance is provided at all until the confidence measure $\rho > 0.5$, which corresponds to an angular difference of less than 90° between the direction vectors from the user and the computer vision module.

The selected arbitration curve is inspired by another study [26] that tested a very aggressive arbitration curve with a sudden jump in the arbitration factor against a more timid one with a gradual change. Their results indicated that the aggressive one worked well in scenarios where the task was difficult and the intent prediction was correct. However, for all other scenarios the timid arbitration curve was to be preferred in terms of task completion time and user preference. The arbitration curve for the adaptive semi-autonomous control was hence designed to be more timid with a gradual change in the arbitration factor.

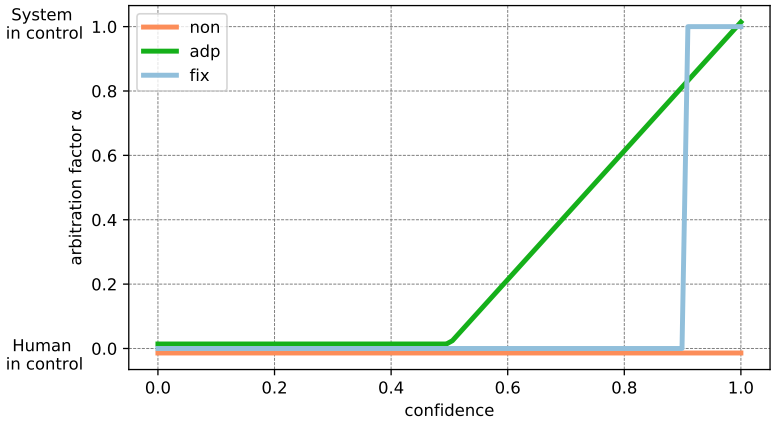


Fig. D.6: The behavior of the different control schemes is illustrated using an arbitration curve. The non-autonomous control (non) is fixed at $\alpha = 0$ as the human is always in control. The curve for the adaptive semi-autonomous control (adp) is given by the function $\alpha = \max(0, 2\rho - 1)$. The fixed semi-autonomous control (fix) is characterized by a sudden jump from the human being in control to the system being in control, which is triggered when the user presses the "auto grasp" button.

The behavior of the two other control schemes can also be illustrated using an arbitration curve, as also shown in Figure D.6. For the non-autonomous control, the user is always in full control and the computer vision provides no assistance. The arbitration factor is hence fixed at $\alpha = 0$, i.e., the human is in control, no matter what the confidence of the system is for this control

scheme. The arbitration curve for the fixed semi-autonomous control is also fixed at $\alpha = 0$ with the exception of a sudden jump to $\alpha = 1$. This jump illustrates the behavior of the fixed semi-autonomous control which takes complete control of the exoskeleton while the user presses and holds the "auto grasp" button. The confidence measure for the fixed semi-autonomous control can hence also be viewed as a step function, where $\rho = 1$ when the user is pressing and holding the "auto grasp" button, and $\rho = 0$ otherwise.

4 Evaluations

To test the developed system, two studies were conducted; study A included 10 participants without tetraplegia and study B included 7 participants with tetraplegia. The overall structure of both studies is described in this section and the points where the two studies differ are described in more detail later.

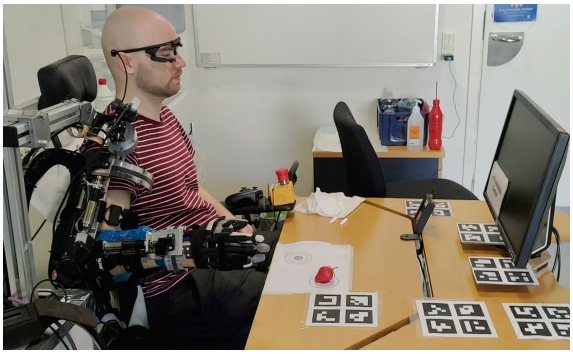
The purpose of the study is to test the following hypotheses:

- **H1:** *"The adaptive semi-autonomous control is better than the non-autonomous control"*.
- **H2:** *"The fixed semi-autonomous control is better than the non-autonomous control"*.
- **H3:** *"The adaptive semi-autonomous control is better than the fixed semi-autonomous control"*.

Whether a respective control scheme can be considered better than the others is assessed using different performance metrics and questionnaires, described later in Sections 4.2 and 4.3.

4.1 Setup

An example of the setup using during both studies can be seen in Figure D.7a, where the participant is placed in a wheelchair while the exoskeleton is attached to their right arm. The length of each link in the exoskeleton was adjusted to fit the participant and the kinematic model of the exoskeleton was updated accordingly. Once the exoskeleton was attached, the participant was asked to complete a small calibration procedure to find the center point of their hand. This was necessary as the soft nature of the Carbonhand meant that the participant's hand did not always end up in the same location. The calibration procedure consisted of grasping a bottle with a known position and orientation and was repeated each time the participant wore the exoskeleton.



(a) The test setup.

(b) The objects for grasping.

Fig. D.7: An overview of the test setup. (**a**) The participant is placed in a wheelchair with the exoskeleton attached to the right arm in the starting position. In front of the participant is a table with different objects to grasp. (**b**) The two objects used in the studies for the task of grasping: a plastic strawberry and a plastic bottle. The objects can be placed on the two predefined positions marked on the table below the objects.

The participant was seated at a table with a computer screen, the wireless receiver for the tongue-based interface, several ArUco markers, and the objects to interact with. The computer screen is used to provide visual feedback to the user during the experiment. The screen displayed visual feedback for the tongue-based interface while the user learned to control it during the first days of the study. When testing the control schemes, the screen is only used to display the current control scheme to the participant. The ArUco markers were added to make it possible to cross-reference recorded images and videos from multiple sources.

Finally, the objects on the table included either a plastic bottle or a plastic strawberry, as shown in Figure D.7b. Both objects are bright red to make detection easier. These two objects could be placed in two predefined positions, as illustrated by the markers under each object. The choice of using a strawberry and a bottle was to have a larger object which was easy to grasp and to have a smaller object which would be more difficult to grasp. Grasping the strawberry would, in most cases, require the participant to rotate the wrist of the exoskeleton. This was not necessary in the case of the bottle.

Four possible test scenarios were constructed from these two objects:

- **Bottle—Single:** Only the bottle is present in one of the two predefined positions. The user must grasp and lift the bottle.
- **Strawberry—Single:** Only the strawberry is present in one of the two predefined positions. The user must grasp and lift the strawberry.
- **Bottle—Multi:** Both the bottle and the strawberry are present and

placed in the two predefined positions. The user must grasp and lift the bottle.

- **Strawberry—Multi:** Both the bottle and the strawberry are present and placed in the two predefined positions. The user must grasp and lift the strawberry.

The test scenarios were constructed to create both easy and difficult situations. In easy scenarios, with only a single object, the intention prediction would always be correct. In the more challenging scenarios, with two objects, the intent prediction could possibly be wrong, and the participant would also have to avoid collision with the object not to grasp. Furthermore, the test scenarios involving the strawberry are anticipated to be more difficult due to its smaller size and as it requires using the wrist rotation of the exoskeleton in order to grasp it.

Each trial started with the exoskeleton being in a predefined home position, as shown in Figure D.7a. The participant was then told what control scheme was active, and what object to grasp and lift; afterwards, they could start moving the exoskeleton. The start of each trial is marked by the participant starting to move the exoskeleton, and the trial ends once the participant has grasped and lifted the object from the table for a few seconds. The trial is restarted if it is deemed impossible to finish the trial successfully. The most common occurrence was situations where an object was accidentally pushed outside the reach of the exoskeleton by the participant.

4.2 Performance Metrics

During each trial, the following metrics were measured to assess the performance of participants in controlling the exoskeleton:

- **Time**—How long it takes the participant to finish the task, measured from when the exoskeleton is first actuated until the participant has grasped and lifted the target object.
- **Commands**—The number of changes in issued commands during the different tasks. Repeatedly pressing the same button on the tongue-controlled interface would hence not count towards this number. Only commands different from the previous command are counted.
- **Cartesian Travel**—The length of the path traveled by the end-effector, i.e., the Carbonhand, during the tasks as measured in Cartesian space. The Cartesian position of the end-effector at each time instant is found using the forward kinematics of the exoskeleton.

4.3 Questionnaires

Two questionnaires were used to assess the intuitiveness and performance of the three tested control schemes. The first questionnaire is the INTUI [35] questionnaire to assess the intuitiveness of completing the tasks when using the different control schemes. It consists of 16 questions where the participant is asked to rate opposite statements on a 7-point scale. The second questionnaire is the raw NASA-TLX (NASA Task Load Index) [36] to assess the workload as perceived by the participant when controlling the exoskeleton using the different control schemes. In the questionnaire, the participant is asked to rate workload based on five factors: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each of these factors is graded on a 21-point scale between two opposite statements, e.g., "Very High" and "Very Low".

Both questionnaires were provided after conducting the last experiment on the last day. The participant was asked to score the different control schemes simultaneously on the same question in the questionnaires, as opposed to separate and successive questionnaires for each control scheme. This was a deliberate choice as separate and successive questionnaires could make it hard for the participant to keep track of previous scoring and the main purpose was to find the difference between the control schemes.

4.4 Statistics

The following describes the post-processing of the metrics and, namely, the statistical analysis of the collected data. The performance metrics measured during the trials, i.e., time, commands, and Cartesian travel, were first grouped based on participant ID, what object to grasp (strawberry or bottle), and whether there was a single object or multiple objects in the scene. This resulted in four groups for each participant, with six samples for each of the three control schemes. Many repetitions per group were performed to avoid problems with outliers. To also avoid problems with pseudo replication, i.e., artificially inflating the number of samples and hence the power in the statistical analysis, only the mean of these samples is used for each group in the following statistical analysis.

All the measured performance metrics were found to be positively skewed and hence log transformed. Afterwards, the normality of the transformed data was then confirmed using Shapiro–Wilk's test. A one-way repeated measures ANOVA with the three control schemes were factors used to test for significance. Mauchly's test was used to test for sphericity and in the case where sphericity was violated, the Greenhouse–Geisser correction was used. Post hoc analysis was conducted for each of the metrics which showed significance in the repeated measures ANOVA test. These post hoc tests

consisted of pairwise comparisons among the different conditions, i.e., the three control schemes, using Bonferroni correction.

The data from the TLX and INTUI questionnaires were tested for significance using the nonparametric Friedman test. This was followed by a post hoc analysis using the Wilcoxon signed rank tests between each unique pairing of the three control schemes. Bonferroni correction was applied in this post hoc analysis as well. Nonparametric tests were used, as both questionnaires rely on an ordinal scale. It should be noted that a significance level of $p = 0.05$ is used throughout the discussion of the results in regard to whether a result was statistically significant or not.

5 Study A—Without Tetraplegia

A total of 10 participants without tetraplegia were recruited for study A. The recruited participants consisted of 1 female and 9 males within the age range of 19–34, with the average age being 25 years. None of them had any connections to the departments of the respective authors.

The participants in study A did not have tetraplegia and were therefore asked to relax both their hand and arm entirely when using the system. This was performed to replicate the intended use case of the system, where an individual with tetraplegia would control the exoskeleton. Furthermore, electromyography (EMG) was recorded at all times during the study to ensure that the participant did not move their hand or arm independently of the exoskeleton by accident. The EMG was recorded using a Myo armband [37] placed on the right upper arm of the participant. After mounting the Myo armband, the participants were asked to repeatedly flex their biceps. These measurements served as a reference for the maximum muscle activation that the participant was capable of. Anytime the measured EMG of a participant would exceed just 20% of the measured maximum muscle activation, the participant would be instructed to relax and possibly repeat any ongoing task.

All participants had received three days of training in using the ITCI for controlling the exoskeleton 4–5 weeks prior to the study, as previous studies on the ITCI have shown that long resting periods are beneficial when learning to use the tongue-based interface [38]. Besides the prior training in using the ITCI, study A consisted of two consecutive days where the first day was used to train using the three different control schemes and refresh how the tongue-based interface worked. In the last day, the participants used the three different control schemes to complete the four test scenarios described earlier, where they had to grasp and lift either a bottle or a strawberry. This was repeated six times to counteract outliers, as also described earlier in Section 4.4.

The ordering of the test scenarios was completely randomized, and the used control scheme was randomized such that the same control scheme could not appear more than twice in a row before using another control scheme. This was performed to avoid having large concentrations of a specific control scheme at the start or end of the study which could skew the data.

5.1 Study A—Performance Results

The results of using the one-way ANOVA test with repeated measures on the performance metrics collected during study A can be seen in Table D.1. The results show that the used control scheme has a statistically significant effect. This is true for all of the four tested scenarios and for all of the three measured performance metrics.

Table D.1: Study A—Result of running a one-way ANOVA test with repeated measures for each of the four different scenarios and the three different performance metrics.

	Time (Seconds)	**Commands (Integer)**	**Cartesian (Meters)**
Bottle Single	F(2,18) = 14.35, $\mathbf{p < 0.001}$	F(2,18) = 17.06, $\mathbf{p < 0.001}$	F(2,18) = 28.1, $\mathbf{p < 0.001}$
Strawberry Single	F(2,18) = 65.49, $\mathbf{p < 0.001}$	F(2,18) = 35.45, $\mathbf{p < 0.001}$	F(2,18) = 16.11, $\mathbf{p < 0.001}$
Bottle Multi	F(2,18) = 5.67, $\mathbf{p = 0.012}$	F(2,18) = 9.09, $\mathbf{p = 0.002}$	F(2,18) = 7.0, $\mathbf{p = 0.006}$
Strawberry Multi	F(2,18) = 30.16, $\mathbf{p < 0.001}$	F(2,18) = 17.26, $\mathbf{p < 0.001}$	F(2,18) = 5.72, $\mathbf{p = 0.012}$

The ANOVA test was hence followed up by a pairwise comparison between the three different control schemes as the selection of control schemes was found to have a statistically significant effect. The results of the pairwise comparison are shown in Table D.2, where the mean percentage-wise performance increase is reported for each pair being compared, for each metric and for each of the four scenarios used in the study.

For the comparison between the non-autonomous and the adaptive semi-automatic control, in Table D.2a, it can be seen that adaptive semi-automatic control results in an improved performance across all 12 cases, with 9 of these being significant and another being close to the threshold of $p < 0.05$. It would hence suggest that the hypothesis, **H1:** *"The adaptive semi-autonomous control is better than the non-autonomous control"*, is true with only a few exceptions in terms of the performance metrics.

Looking at non-autonomous, i.e., manual, control versus fixed semi-autonomous control, in Table D.2b the fixed semi-autonomous control results in the best

Table D.2: Study A —Pairwise comparison between the three tested control schemes. The mean percentage-wise increase in performance for each comparison is reported, where positive numbers denote an improvement (i.e., reduction) in favor of the hypothesis. The Bonferroni-corrected p-value and the 95% confidence interval are reported as well. Bold font indicates statistical significance for the p-values and an improvement in favor of the hypothesis. Significant results supporting the hypothesis are marked with green while results supporting the hypothesis but lacking significance are marked with yellow. Results marked with red do not support the hypothesis.

(**a**) Hypothesis **H1**, adaptive semi-autonomous control (adp) is better than non-autonomous control (non).

	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	**37%** $[6.0, 10]$, $\mathbf{p = 0.001}$	**37%** $[1.2, 2.3]$, $\mathbf{p = 0.007}$	**33%** $[0.14, 0.2]$, $\mathbf{p < 0.001}$
Strawberry Single	**58%** $[19, 31]$, $\mathbf{p < 0.001}$	**60%** $[3.9, 8.5]$, $\mathbf{p < 0.001}$	**38%** $[0.18, 0.3]$, $\mathbf{p = 0.001}$
Bottle Multi	**31%** $[4.9, 10]$, $p = 0.051$	**17%** $[0.52, 1.0]$, $p = 0.43$	**30%** $[0.12, 0.22]$, $\mathbf{p = 0.026}$
Strawberry Multi	**43%** $[14, 22]$, $\mathbf{p < 0.001}$	**44%** $[2.7, 5.6]$, $\mathbf{p = 0.003}$	**16%** $[0.08, 0.12]$, $p = 0.073$

(**b**) Hypothesis **H2**, fixed semi-autonomous control (fix) is better than non-autonomous control (non).

	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	**18%** $[3.1, 4.7]$, $\mathbf{p = 0.046}$	-4.1% $[-0.27, -0.14]$, $p = 1.0$	**23%** $[0.1, 0.13]$, $\mathbf{p = 0.001}$
Strawberry Single	**46%** $[14, 25]$, $\mathbf{p < 0.001}$	**37%** $[2.6, 5.1]$, $\mathbf{p = 0.009}$	**25%** $[0.11, 0.22]$, $p = 0.074$
Bottle Multi	**14%** $[2.2, 4.5]$, $p = 0.77$	-26% $[-3.8, -0.56]$, $p = 0.11$	**19%** $[0.07, 0.14]$, $p = 0.28$
Strawberry Multi	**49%** $[15, 26]$, $\mathbf{p < 0.001}$	**35%** $[2.4, 4.0]$, $\mathbf{p = 0.002}$	**28%** $[0.13, 0.23]$, $\mathbf{p = 0.034}$

(**c**) Hypothesis **H3**, adaptive semi-autonomous control (adp) is better than fixed semi-autonomous control (fix).

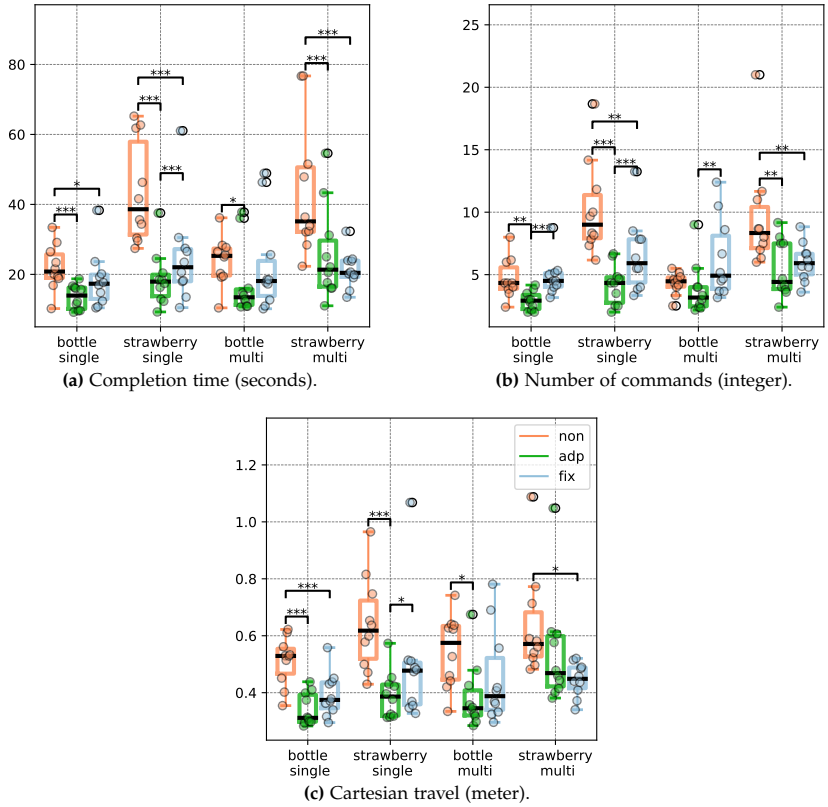
	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	**23%** $[3.0, 5.4]$, p=0.09	**40%** $[1.4, 2.4]$, **p=0.001**	**13%** $[0.04, 0.06]$, p=0.092
Strawberry Single	**22%** $[4.2, 5.5]$, **p=0.001**	**36%** $[1.8, 2.7]$, **p<0.001**	**17%** $[0.07, 0.09]$, **p=0.022**
Bottle Multi	**20%** $[3.1, 4.9]$, p=0.054	**39%** $[1.6, 3.0]$, **p=0.003**	**13%** $[0.05, 0.07]$, p=0.1
Strawberry Multi	-10% $[-3.9, -1.4]$, p=0.96	**13%** $[0.6, 1.0]$, p=0.46	-15% $[-0.14, -0.04]$, p=0.64

performance in 10 out of 12 cases, with 7 of these being significant. The hypothesis, **H2:** *"The fixed semi-autonomous control is better than the non-autonomous control"*, is hence not entirely implausible but it cannot be confirmed either. The cases where the fixed semi-autonomic control is significantly better than non-autonomous control are primarily the scenarios involving the strawberry. This scenario is also difficult as the wrist of the exoskeleton needs to be ro-

tated to grasp the strawberry. It could hence indicate that the fixed semi-autonomous control is beneficial once the task reaches a certain level of difficulty.

For the comparison between the adaptive and the fixed semi-autonomous control, shown in Table D.2c, the adaptive scheme results in the best performance in 10 out of 12 cases, with five out of these cases being significant and another being close to $p < 0.05$ significance threshold. It is hence not possible to decisively confirm or deny the hypothesis, **H3:** *"The adaptive semi-autonomous control is better than the fixed semi-autonomous control"*. However, it can be argued that this hypothesis is true for some scenarios, such as the one with a single strawberry which resulted in a significant improvement across all metrics.

Finally, the measured performance metrics from the last day of study A are shown as box plots in Figure D.8. Each plot is split based on the four different scenarios used during the study. Any pairwise significance between the three control schemes is indicated with asterisks in the plots.



(a) Completion time (seconds).

(b) Number of commands (integer).

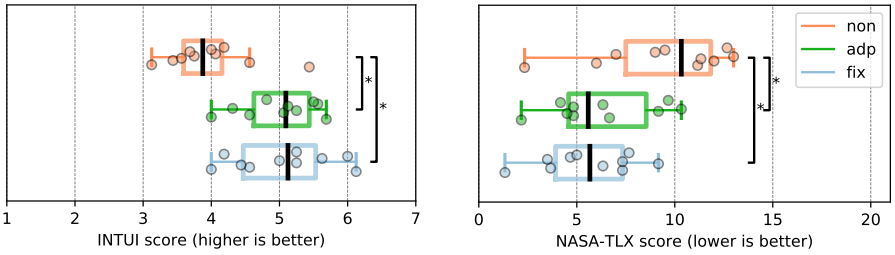
(c) Cartesian travel (meter).

Fig. D.8: Study A—Box plots for the performance metrics; (**a**) completion time, (**b**) number of commands used, and (**c**) Cartesian travel. Lower is better for all three metrics. Any significance between the different control schemes is indicated using asterisks (* = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$).

5.2 Study A—Questionnaire Results

At the end of study A, each participant had to answer the INTUI and the NASA-TLX questionnaires. These questionnaires serve to evaluate the intuitiveness of the control schemes and how demanding it was to complete the tasks when using the different control schemes. Applying the Friedman test showed a statistically significant difference in the scores depending on the used control scheme, for both the INTUI ($X^2(2) = 14.824, p < 0.001$) and NASA-TLX ($X^2(2) = 15.846, p < 0.001$).

A post hoc analysis of the results was performed using Wilcoxon, with Bonferroni correction for multiple comparisons, to identify statistical significance between the three control schemes. The results for the INTUI questionnaire show a statistically significant difference when comparing the non-autonomous control with either the adaptive semi-autonomous control ($Z = -2.67, p = 0.024$) or fixed semi-autonomous control ($Z = -2.67, p = 0.024$). This statistical significance is also indicated in the box plot of the scores shown in Figure D.9a. From the box plot it is also clear that the significant difference is an improvement, i.e., an increase in the score, in favor of both the adaptive and fixed semi-autonomous control. However, there is little to no difference when comparing the adaptive and the fixed semi-autonomous control, and no statistical significance was found ($Z = -0.341, p = 1.0$).



(a) INTUI. **(b)** NASA-TLX.

Fig. D.9: Study A—Box plots of the scores for the different control schemes from the two questionnaires; (**a**) INTUI measuring intuitiveness (higher is better) and (**b**) NASA-TLX measuring the task load (lower is better). Any significance between the different control schemes is indicated using asterisks (* = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$).

A similar trend is seen for the results of the NASA-TLX questionnaire in Figure D.9b. There is a clear and significant improvement, i.e., decrease, in the task load when using either the adaptive semi-autonomous control ($Z = -2.807, p = 0.015$) or the fixed semi-autonomous control ($Z = -2.805, p = 0.015$) in comparison to the non-autonomous control. The results of the comparison between the adaptive and fixed semi-autonomous control are similar to what was observed for the INTUI questionnaire as there is no significant difference ($Z = -1.423, p = 0.465$).

The results for the two questionnaires would have suggested that both hypotheses, **H1:** *"The adaptive semi-autonomous control is better than the non-autonomous control"*, and **H2:** *"The fixed semi-autonomous control is better than the non-autonomous control"*, are true. The last hypothesis, **H3:** *"The adaptive semi-autonomous control is better than the fixed semi-autonomous control"*, cannot be confirmed based on the results from the questionnaires.

6 Study B—With Tetraplegia

Study B included 10 individuals with varying degrees of tetraplegia, but all of them had to fullfill the following criteria:

- Reduced or no function in their upper body, especially their right arm and hand, where the exoskeleton had to be mounted.
- Tongue must be functional such that the tongue-based interface can be used.
- While seated and without assistance they must not be able to grasp and lift a bottle of water placed on a table.

However, 3 of the 10 participants had to be omitted from further data analysis due to incomplete data. The cause of the incomplete data was due to fatigue by the participants, at which point it was deemed best to cut the current session shorter. Study B was subsequently reduced to only include two out of the four scenarios previously used in study A to avoid situations such as this. The more difficult scenarios with two objects present were skipped and only the scenarios with either a single bottle or a single strawberry were tested. The data analysis of study B is hence based on the seven participants with complete data, after reducing the number of tested scenarios. These participants had a mean age of 55 years, ranging from 23 to 69, with one female.

The structure of study B consisted of three consecutive days: in the first day, the participants trained to use the ITCI on a simulation of the exoskeleton and in the second day they continued their training on the real exoskeleton. The third and final day was used to train using the different control schemes and conduct the final test of the system.

Study B was hence two days shorter than study A and with one of the days training on a simulation of the exoskeleton. Furthermore, all days were right after each other, unlike study A, which included an intermediate period of rest for several weeks. The setup for study B was not ideal and did omit many of the considerations from study A, but it was a matter of making it feasible for individuals with tetraplegia to participate. The structure of study

B was hence condensed to minimize the amount of time that the participants would have to travel and/or stay in a hotel.

6.1 Study B—Performance Results

The result of applying one-way ANOVA with repeated measures on the performance metrics collected for study B is shown in Table D.3. Statistical significance was found for both the tested scenarios and for all the three collected performance metrics.

Table D.3: Study B—Result of the one-way ANOVA test with repeated measures for each of the four different scenarios and the three different performance metrics.

	Time (Seconds)	**Commands (Integer)**	**Cartesian (Meters)**
Bottle Single	F(2,12) = 15.58, $\mathbf{p < 0.001}$	F(2,12) = 5.54, $\mathbf{p = 0.02}$	F(2,12) = 6.77, $\mathbf{p = 0.011}$
Strawberry Single	F(2,12) = 11.28, $\mathbf{p = 0.002}$	F(2,12) = 8.32, $\mathbf{p = 0.005}$	F(2,12) = 6.42, $\mathbf{p = 0.013}$

A pairwise comparison between the three control schemes was carried out to identify any statistical significance between the control schemes. The results of this comparison are shown in Table D.4, reporting the mean percentage-wise increase in performance, the associated p-values, and confidence intervals with Bonferroni correction. The same color scheme is used as previously described for the results in study A.

For the comparison between the non-autonomous control and the adaptive semi-autonomous control, the adaptive semi-autonomous improves performance across all six of the tested cases. Statistical significance is found in three out of six of these cases. The results are somewhat similar for the comparison between the non-autonomous control and the fixed semi-autonomous control, where the latter improves performance across all six of the tested cases as well. Furthermore, four out of these six cases are statistically significant. The performance metrics from study B do hence indicate the plausibility of hypothesis **H1:** *"The adaptive semi-autonomous control is better than the non-autonomous control"*, and especially hypothesis **H2:** *"The fixed semi-autonomous control is better than the non-autonomous control"*, without ultimately being able to outright confirm them. Finally, for the comparison between the adaptive and the fixed semi-autonomous control, there is an equal split between which of the two control schemes performed the best. However, none of these cases were found to have any statistical significance. It is hence not possible to support hypothesis **H3:** *"The adaptive semi-autonomous control is better than the fixed semi-autonomous control"*based on the results from study B.

Table D.4: Study B—Pairwise comparison between the three tested control schemes. The mean percentage-wise increase in performance for each comparison is reported, along with the associated p-value and 95% confidence interval (both Bonferroni-corrected). Statistically significant p-values are marked with bold. Significant results supporting the hypothesis are marked with green while results supporting the hypothesis but lacking significance are marked with yellow. Results marked with red do not support the hypothesis.

(**a**) Hypothesis **H1**, adaptive semi-autonomous control (adp) is better than non-autonomous control (non).

	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	**41%** $[12, 27]$, $\mathbf{p = 0.02}$	**43%** $[2.4, 8.3]$, $p = 0.072$	**41%** $[8.9, 35]$, $p = 0.22$
Strawberry Single	**54%** $[23, 57]$, $\mathbf{p = 0.004}$	**56%** $[4.7, 16]$, $\mathbf{p = 0.014}$	**33%** $[0.17, 0.45]$, $p = 0.099$

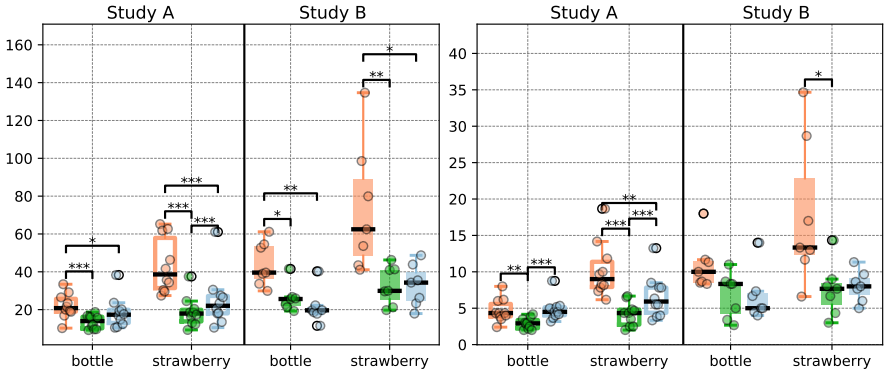
(**b**) Hypothesis **H2**, fixed semi-autonomous control (fix) is better than non-autonomous control (non).

	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	**53%** $[14, 39]$, $\mathbf{p = 0.009}$	**42%** $[2.3, 8.7]$, $p = 0.1$	**53%** $[12, 45]$, $\mathbf{p = 0.02}$
Strawberry Single	**54%** $[17, 78]$, $\mathbf{p = 0.048}$	**50%** $[3.5, 17]$, $p = 0.089$	**31%** $[0.18, 0.38]$, $\mathbf{p = 0.048}$

(**c**) Hypothesis **H3**, adaptive semi-autonomous control (adp) is better than fixed semi-autonomous control (fix).

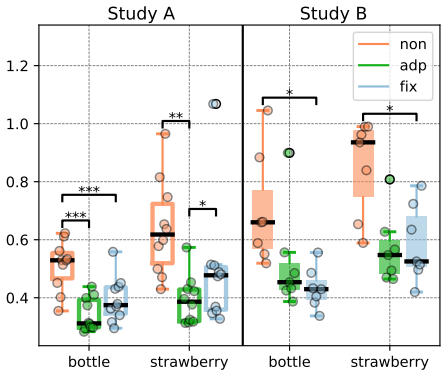
	time (seconds)	commands (integer)	Cartesian (meters)
Bottle Single	-21% $[-13, -2.2]$, $p = 0.35$	**0.37%** $[0.01, 0.04]$, $p = 1.0$	-21% $[-12, -2.4]$, $p = 0.8$
Strawberry Single	-0.29% $[-0.17, -0.05]$, $p = 1.0$	**12%** $[0.47, 2.0]$, $p = 1.0$	**2.6%** $[0.01, 0.02]$, $p = 1.0$

The measured performance metrics from the last day of study B are shown in Figure D.10, along with the results from the last day of study A for comparison. Only the two scenarios with a single bottle and a single strawberry are shown, as the scenarios with multiple objects were skipped for study B. Pairwise significance between the three control schemes are indicated with asterisks in the box plots, as previously. Looking at the performance metrics for study B, the difference between the three control schemes appears similar to the pattern observed for study A. Using either of the two semi-autonomous control schemes improves all the performance metrics in comparison with the non-autonomous control. However, most of the metrics from study B also appear to be higher than for study A, suggesting a worse performance in general for participants in study B.



(a) Completion time (seconds).

(b) Number of commands (integer).

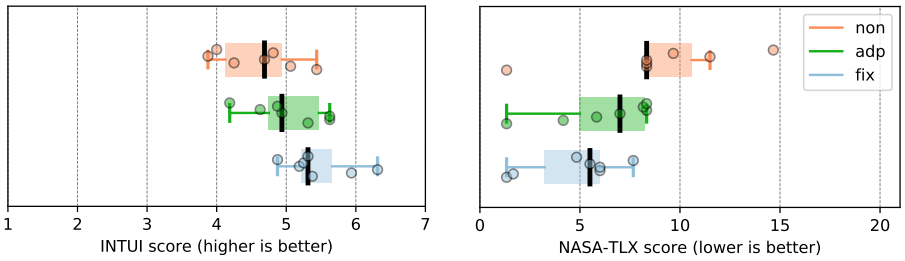


(c) Cartesian travel (meter).

Fig. D.10: Box plots for the performance metrics for both study A and B in the scenarios with only single objects; (**a**) completion time, (**b**) number of commands used, and (**c**) Cartesian travel. Lower is better. Any significance between the different control schemes is indicated using asterisks (* = $p < 0.05$, ** = $p < 0.01$ and *** = $p < 0.001$).

6.2 Study B—Questionnaire Results

The results of applying a Friedman test to the scores from the NASA-TLX and INTUI questionnaires from study B indicated statistical significance between the control schemes for both INTUI ($X^2(2) = 8.222, p = 0.016$) and NASA-TLX ($X^2(2) = 10.333, p < 0.006$). However, post hoc analysis using Wilcoxon signed rank tests showed no statistical significance between any of the control schemes once adjusting for multiple comparisons using Bonferroni correction. Despite the lack of significance, the scores do differ for the three control schemes, as shown in Figure D.11. For both INTUI and NASA-TLX, the non-autonomous control scheme appears to perform the worst while the fixed semi-autonomous control performs the best across both questionnaires.



(a) INTUI. **(b)** NASA-TLX.

Fig. D.11: Study B—Box plots of the scores for the different control schemes for the two questionnaires; (**a**) INTUI measuring intuitiveness (higher is better) and (**b**) NASA-TLX measuring the task load (lower is better).

7 Discussion

Looking at the pairwise comparison for the performance metrics from study A (Table D.2), there appears to be a trend of the scenarios with multiple objects lacking significance and vice versa. For the multi-object scenarios, there is a lack of statistical significance in 11 out of 18 cases, whereas it is 4 out of 18 cases for the single object scenarios. Another interesting observation in relation to this is how the different control schemes behave when moving from a scenario with only a single object to one with multiple objects. In the case of the non-autonomous control scheme, the introduction of multiple objects does not seem to alter any of the measured performance metrics much. The opposite is true for both the adaptive and fixed semi-autonomous control, as most of their metrics in Figure D.8 appear to increase when introducing multiple objects. This is expected, as having multiple objects in the scene requires the system to predict the intention of the user. However, these results indicate that improving the current approach for intention prediction could be beneficial.

The current method for intention prediction can be considered amnesic as it relies solely on the current state, i.e., where the object is in relation to the hand right now. It could be beneficial to use prior information in a memory-based approach for the intent prediction, such as considering the entire trajectory traveled by the hand so far [25, 26].

The idea of having two different objects in the two studies was to provide varying levels of difficulty. Looking at the results from both studies A and B, in Figure D.10, it appears that this choice was successful as some scenarios are clearly more difficult than others when using the manual control scheme. For example, the strawberry generally takes a longer time and requires more commands to pick up as compared to the bottle. This is not surprising as the small size of the strawberry often requires a pincer grasp where the wrist of

the exoskeleton needs to be rotated. The bottle can instead be grasped using a palm grasp where there is no need to rotate the wrist of the exoskeleton. In further studies, it may hence be beneficial to include even more of these difficult tasks as the benefits of using semi-autonomous control are more pronounced in these cases. This is also clear when looking at the pairwise comparisons for both study A (Table D.2) and study B (Table D.4), where the scenarios involving a strawberry account for the majority of the statistically significant results. This observation of semi-autonomous control being more beneficial for difficult tasks has been made in several other studies as well [18, 20].

However, this difference in performance between grasping the strawberry and the bottle is less apparent for the adaptive and fixed semi-autonomous controls. A possible explanation is the fact that both these two control schemes can simultaneously adjust both the position and orientation of the exoskeleton. Performing simultaneous control of both the position and orientation is not possible for the non-autonomous control as it would require issuing two commands at once, which is not possible due to the nature of the tongue-based interface. The simultaneous adjustment of both position and orientation is hence a clear benefit of using either the adaptive or fixed semi-autonomous control instead of the non-autonomous, i.e., manual, control.

In the results for the performance metrics from study B, there is no significant difference between the adaptive semi-autonomous control and the fixed version. This differs from the results found in study A, where the adaptive semi-autonomous control resulted in a significant reduction in at least a few cases, especially when considering the number of commands used. Another major difference between the results of study A and B is a consistent decrease in performance for all the different control schemes. In general, the participants in study B use longer time, more commands, and the hand of the exoskeleton travels further in comparison to the participants from study A. This is despite the task being identical and using the exact same system in terms of both hardware and software. The results from study B also appear to carry less statistical power than the results found from study A; this is likely due to the smaller sample size, i.e., number of participants, but also partly due to a higher amount of noise in the collected measurements for study B. The presence of more noise is clear when looking at the distribution of measurements in study A and B in Figure D.10.

A likely explanation for the difference in the results could be the different structure used in the two studies. In study B, the participants had less time for training to use the system, both in terms of using the ITCI, learning how the exoskeleton moves, and how the different control schemes behave. This could have impacted the performance of the participants in study B as the ITCI may have a relatively long learning curve [38], even though most learning takes place within the first 3 days. Another possible factor contributing to

the different performance between studies A and B could the age difference. The mean age of the participants in study B is over twice the age of the participants in study A (25 versus 55 years). Similar observations were made in a study on controlling a computer using neck movements, where performance decreased as the age of the participants increased [39]. This could indicate that age is indeed a factor when using the proposed system.

Looking at the NASA-TLX and INTUI questionnaires, they confirm many of the same observations made from the performance metrics for both studies. For study A, both the adaptive and fixed semi-autonomous control are significantly better than non-autonomous control in terms of both the INTUI and NASA-TLX questionnaires. However, there is no significant difference between the adaptive and fixed semi-autonomous control in the questionnaires in study A. This is despite the performance metrics indicating some significance in at least certain scenarios.

It is possible that making the questionnaires more fine-grained, e.g., one for each of the scenarios, would have yielded a significant difference between the adaptive and fixed semi-autonomous control in some cases, similarly to what was observed for the performance metrics. However, such an approach was deemed infeasible as it would require the participants to answer four times as many questionnaires.

The INTUI and NASA-TLX results from study B show a more pronounced difference between the adaptive and fixed semi-autonomous control for both questionnaires in comparison to study A. This may be altered slightly if the learning of using the ITCI had been completed as further learning takes place after the currently used 3–5 days. This difference lacks statistical significance, but it could indicate that users without much training, i.e., study B, prefer the fixed semi-autonomous control, whereas there is no clear preference between the adaptive and fixed semi-autonomous control for users with more training, i.e., study A. The difference between study A and B in terms of the questionnaires could once again be related to a combination of less training for participants in study B and the age difference between the two groups of participants. This could indicate that it would have been beneficial to run study B using the same structure as study A and preferably with a younger age group. However, at the time when study B was conducted, this was not possible, but it is something to keep in mind for future studies.

Finally, the current system relies on classic image processing techniques when performing object detection for the sake of producing reliable results in a controlled environment. This approach will hence not work well in an unconstrained environment with unknown objects, such as the home of an individual with tetraplegia. This shortcoming may be remedied by using deep learning approaches [30, 31] trained on vast datasets [32] or applying methods for object-agnostic grasp detection, which should work for any arbitrary object [40, 41].

8 Conclusions

Three control schemes with varying degrees of autonomy were implemented and used in the context of performing tongue-based control of an upper limb exoskeleton for individuals with tetraplegia. Computer vision was used to detect nearby objects to infer the intention of the user. The confidence of this prediction was used by an adaptive semi-autonomous control to continually adjust the amount of assistance provided when controlling the exoskeleton. The adaptive semi-autonomous control was tested against non-autonomous (i.e., manual) control and fixed semi-autonomous control, where the level of assistance was always the same.

The three control schemes were tested across two studies: 10 participants without tetraplegia and 7 participants with tetraplegia. Both studies showed a clear improvement when using either the adaptive or fixed semi-autonomous control instead of the non-autonomous control. The participants without tetraplegia also showed a significant improvement for several of the tested tasks when using the adaptive semi-autonomous control instead of its fixed counterpart. However, the participants with tetraplegia performed better with the fixed semi-autonomous control instead of the adaptive one in many cases. These different results and preferences across the two studies could be attributed to a much higher average age for the participants with tetraplegia along with less training in using the tongue-based control as well.

The benefits of using an adaptive versus a fixed level of autonomy for the semi-autonomous control appear to depend on the user and their amount of experience in using the system. Nevertheless, the results clearly show that both the semi-autonomous control schemes are to be preferred over manual control. Furthermore, using the adaptive semi-autonomous control instead of the manual non-autonomous control did not appear to have any drawbacks during the two studies as it was found to improve performance in all the tested cases. The fixed semi-autonomous control did, on the other hand, reduce performance in a few cases when compared to the non-autonomous control.

Author Contributions: Conceptualization, S.H.B., M.B.T., M.M., F.V.K., M.A.G., L.N.S.A.S., T.B., and T.B.M.; methodology, S.H.B., M.B.T., M.M., F.V.K., M.A.G., L.N.S.A.S., T.B., and T.B.M.; software, S.H.B., M.B.T., and M.M.; validation, S.H.B.; formal analysis, S.H.B. and M.B.T.; investigation, S.H.B., M.B.T., M.M., F.V.K., M.A.G., and L.N.S.A.S.; resources, M.B.T., M.M., F.V.K., M.A.G., and L.N.S.A.S.; data curation, S.H.B., M.B.T., and M.M.; writing—original draft preparation, S.H.B.; writing—review and editing, S.H.B., M.B.T., M.M., F.V.K., M.A.G., L.N.S.A.S., T.B., and T.B.M.; visualization, S.H.B., M.B.T., and M.M.; supervision, S.H.B., M.B.T., L.A.S, T.B., and T.B.M.; project administration,

S.H.B., M.B.T., M.M., F.V.K., and L.N.S.A.S.; funding acquisition, L.N.S.A.S., T.B., and T.B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted as part of the EXOTIC project funded by Aalborg University, Denmark.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Science Ethics Committee for the North Denmark Region (reg. no.: VN-20190030 and VN-20210016, approved 17 August 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not publicly available due to privacy concerns.

Acknowledgments: We would like to thank the participants in the studies for their effort and patience. We would also like to thank Kåre Eg Severinsen and Benjamin Yamin Ali Khan from the Spinal Cord Injury Centre of Western Denmark for their support in recruiting and conducting the studies. Bo Bentsen also deserves our gratitude for his support and guidance on how to handle the tongue-based interface. Finally, we would like to thank Rasmus Leck Kæseler for always being ready to step in and help when needed.

Conflict of Interest: The authors declare no conflicts of interest.

References

[1] K. M. Marasinghe, "Assistive technologies in reducing caregiver burden among informal caregivers of older adults: a systematic review," *Disability and Rehabilitation: Assistive Technology*, vol. 11, no. 5, pp. 353–360, 2016.

[2] G. Romer, H. Stuyt, and A. Peters, "Cost-savings and economic benefits due to the assistive robotic manipulator (arm)," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 201–204.

[3] J. Bickenbach, A. Officer, T. Shakespeare, P. von Groote, W. H. Organization, and T. I. S. C. Society, *International perspectives on spinal cord injury / edited by Jerome Bickenbach ... [et al].* World Health Organization, 2013.

[4] M. Wyndaele and J. J. Wyndaele, "Incidence, prevalence and epidemiology of spinal cord injury: what learns a worldwide literature survey?" *Journal of the International Spinal Cord Society (ISCoS)*, no. 44, pp. 523–529, 2006.

[5] F. V. Kobbelgaard, A. M. Kanstrup, and L. N. S. A. Struijk, "Exploring user requirements for an exoskeleton arm insights from a user-centered study with people living with severe paralysis," in *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 2021, pp. 312–320.

[6] M. Gandolla, S. D. Gasperina, V. Longatelli, A. Manti, L. Aquilante, M. G. D'Angelo, E. Biffi, E. Diella, F. Molteni, M. Rossini, M. Gföhler, M. Puchinger, M. Bocciolone, F. Braghin, and A. Pedrocchi, "An assistive upper-limb exoskeleton controlled by multi-modal interfaces for severely impaired patients: development and experimental assessment," *Robotics and Autonomous Systems*, vol. 143, p. 103822, Sep. 2021.

[7] M. Hosseini, R. Meattini, G. Palli, and C. Melchiorri, "A wearable robotic device based on twisted string actuation for rehabilitation and assistive applications," *Journal of Robotics*, vol. 2017, pp. 1–11, 2017.

[8] M. Mohammadi, H. Knoche, M. Thøgersen, S. H. Bengtson, M. A. Gull, B. Bentsen, M. Gaihede, K. E. Severinsen, and L. N. S. A. Struijk, "Eyes-free tongue gesture and tongue joystick control of a five DOF upper-limb exoskeleton for severely disabled individuals," *Frontiers in Neuroscience*, vol. 15, Dec. 2021.

[9] L. Struijk, E. Lontis, M. Gaihede, H. Caltenco, M. Lund, H. Schiøler, and B. Bentsen, "Development and functional demonstration of a wireless intraoral inductive tongue computer interface for severely disabled persons," *Disability and Rehabilitation: Assistive Technology*, vol. 12, no. 6, pp. 631–640, 2017.

[10] S. H. Bengtson, T. Bak, L. N. S. A. Struijk, and T. B. Moeslund, "A review of computer vision for semi-autonomous control of assistive robotic manipulators (arms)," *Disability and Rehabilitation: Assistive Technology*, vol. 15, no. 7, pp. 731–745, 2020.

[11] M. Nann, F. Cordella, E. Trigili, C. Lauretti, M. Bravi, S. Miccinilli, J. M. Catalan, F. J. Badesa, S. Crea, F. Bressi, N. Garcia-Aracil, N. Vitiello, L. Zollo, and S. R. Soekadar, "Restoring activities of daily living using an eeg/eog-controlled semiautonomous and mobile whole-arm exoskeleton in chronic stroke," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2314–2321, 2021.

[12] M. Barsotti, D. Leonardis, C. Loconsole, M. Solazzi, E. Sotgiu, C. Procopio, C. Chisari, M. Bergamasco, and A. Frisoli, "A full upper limb robotic exoskeleton for reaching and grasping rehabilitation triggered by MI-BCI," in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)*. IEEE, Aug. 2015, pp. 49–54.

[13] A. Frisoli, C. Loconsole, D. Leonardis, F. Banno, M. Barsotti, C. Chisari, and M. Bergamasco, "A new gaze-BCI-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1169–1179, Nov. 2012.

[14] Z. Tang, K. Zhang, S. Sun, Z. Gao, L. Zhang, and Z. Yang, "An upper-limb power-assist exoskeleton using proportional myoelectric control," *Sensors*, vol. 14, no. 4, pp. 6677–6694, 2014.

[15] Z. Zhang, Y. Huang, S. Chen, J. Qu, X. Pan, T. Yu, and Y. Li, "An intention-driven semi-autonomous intelligent robotic system for drinking," *Frontiers in Neurorobotics*, vol. 11, p. 48, 2017.

[16] M. Mohammadi, H. Knoche, M. Gaihede, B. Bentsen, and L. N. S. Andreasen Struijk, "A high-resolution tongue-based joystick to enable robot control for individuals with severe disabilities," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, 2019, pp. 1043–1048.

[17] L. N. S. A. Struijk, L. L. Egsgaard, R. Lontis, M. Gaihede, and B. Bentsen, "Wireless intraoral tongue control of an assistive robotic arm for individuals with tetraplegia," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, Nov. 2017.

[18] H. W. Ka, C.-S. Chung, D. Ding, K. James, and R. Cooper, "Performance evaluation of 3d vision-based semi-autonomous control method for assistive robotic manipulator," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 2, pp. 140–145, 2018.

[19] D.-J. Kim, R. Hazlett-Knudsen, H. Culver-Godfrey, G. Rucks, T. Cunningham, D. Portee, J. Bricout, Z. Wang, and A. Behal, "How autonomy impacts performance and satisfaction: Results from a study with spinal cord injured subjects using an assistive robot," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 1, pp. 2–14, Jan. 2012.

[20] M. Hildebrand, F. Bonde, R. Kobborg, C. Andersen, A. Norman, M. Thøgersen, S. Bengtson, S. Dosen, and L. Struijk, "Semi-autonomous tongue-control of an assistive robotic arm for individuals

with quadriplegia," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, ser. IEEE International Conference on Rehabilitation Robotics. Proceedings, vol. 2019. United States: IEEE, Jun. 2019, pp. 157–162.

[21] V. W. Oguntosin, Y. Mori, H. Kim, S. J. Nasuto, S. Kawamura, and Y. Hayashi, "Design and validation of exoskeleton actuated by soft modules toward neurorehabilitation-vision-based control for precise reaching motion of upper limb," *Frontiers in neuroscience*, vol. 11, pp. 352–352, Jul 2017.

[22] S. Crea, M. Nann, E. Trigili, F. Cordella, A. Baldoni, F. J. Badesa, J. M. Catalán, L. Zollo, N. Vitiello, N. G. Aracil, and S. R. Soekadar, "Feasibility and safety of shared eeg/eog and vision-guided autonomous whole-arm exoskeleton control to perform activities of daily living," *Scientific Reports*, vol. 8, no. 1, p. 10823, Jul 2018.

[23] C. Loconsole, F. Stroppa, V. Bevilacqua, and A. Frisoli, "A robust real-time 3d tracking approach for assisted object grasping," in *Haptics: Neuroscience, Devices, Modeling, and Applications*, M. Auvray and C. Duriez, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 400–408.

[24] C.-S. Chung, H. Wang, and R. A. Cooper, "Functional assessment and performance evaluation for assistive robotic manipulators: Literature review," *The Journal of Spinal Cord Medicine*, vol. 36, no. 4, pp. 273–289, 2013.

[25] K. Muelling, A. Venkatraman, J.-S. Valois, J. E. Downey, J. Weiss, S. Javdani, M. Hebert, A. B. Schwartz, J. L. Collinger, and J. A. Bagnell, "Autonomy infused teleoperation with application to brain computer interface controlled manipulation," *Autonomous robots*, vol. 41, no. 6, pp. 1401–1422, 2017.

[26] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[27] M. Gull, M. Thøgersen, S. Bengtson, M. Mohammadi, L. Struijk, T. Moeslund, T. Bak, and S. Bai, "A 4-dof upper limb exoskeleton for physical assistance: Design, modeling, control and performance evaluation," *Applied Sciences*, vol. 11, no. 13, Jun. 2021.

[28] M. Thøgersen, M. Gull, F. Kobbelgaard, M. Mohammadi, S. Bengtson, and L. Struijk, "Exotic - a discreet user-based 5 dof upper-limb exoskeleton for individuals with tetraplegia," in *2020 IEEE 3rd International Conference on Mechatronics, Robotics and Automation*. United States: IEEE, 2021, pp. 79–83.

[29] L. Struijk, "An inductive tongue computer interface for control of computers and assistive devices," *I E E E Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2594–2597, 2006.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, pp. 740–755.

[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381395, jun 1981.

[34] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1–4.

[35] D. Ullrich and S. Diefenbach, "Intui. exploring the facets of intuitive interaction," in *Mensch & Computer 2010: Interaktive Kulturen*, J. Ziegler and A. Schmidt, Eds. München: Oldenbourg Verlag, 2010, pp. 251–260.

[36] S. Hart and L. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[37] P. Visconti, F. Gaetani, G. Zappatore, and P. Primiceri, "Technical features and functionalities of myo armband: An overview on related literature and advanced applications of myoelectric armbands mainly focused on arm prostheses," *International Journal on Smart Sensing and Intelligent Systems*, vol. 11, no. 1, pp. 1–25, 2018.

[38] H. A. Caltenco, E. R. Lontis, S. A. Boudreau, B. Bentsen, J. Struijk, and L. N. S. Andreasen Struijk, "Tip of the tongue selectivity and motor learning in the palatal area," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 1, pp. 174–182, 2012.

[39] G. L. Hands and C. E. Stepp, "Effect of Age on HumanComputer Interface Control Via Neck Electromyography," *Interacting with Computers*, vol. 28, no. 1, pp. 47–54, 08 2014.

[40] M. Gualtieri, J. Kuczynski, A. M. Shultz, A. Ten Pas, R. Platt, and H. Yanco, "Open world assistive grasping using laser selection," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4052–4057.

[41] A. Miller, S. Knoop, H. Christensen, and P. Allen, "Automatic grasp planning using shape primitives," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, 2003, pp. 1824–1829 vol.2.

Paper E

Pose Estimation from RGB Images of Highly Symmetric Objects using a Novel Multi-Pose Loss and Differential Rendering

Stefan Hein Bengtson, Hampus Åström, Thomas B. Moeslund, Elin A. Topp, and Volker Krueger

The paper has been published in the
Proceedings for the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4618–4624, 2021.

Abstract

We propose a novel multi-pose loss function to train a neural network for 6D pose estimation, using synthetic data and evaluating it on real images. Our loss is inspired by the VSD (Visible Surface Discrepancy) metric and relies on a differentiable renderer and CAD models. This novel multi-pose approach produces multiple weighted pose estimates to avoid getting stuck in local minima. Our method resolves pose ambiguities without using predefined symmetries. It is trained only on synthetic data. We test on real-world RGB images from the T-LESS dataset, containing highly symmetric objects common in industrial settings. We show that our solution can be used to replace the codebook in a state-of-the-art approach. So far, the codebook approach has had the shortest inference time in the field. Our approach reduces inference time further while a) avoiding discretization, b) requiring a much smaller memory footprint and c) improving pose recall.

1 Introduction

As robotics moves towards flexible and autonomous solutions, computer vision is gradually playing a bigger role in robotic solutions, especially for 6D pose estimation. This topic has actively been researched in the robotics community [1] for many years, as the pose of an object is very useful when figuring out how to interact with it. Pose estimation is useful in other areas as well, e.g. in augmented reality.

However, it is still a challenging problem, and pose estimation has hence been the focus of many public datasets and challenges issued by the community. One challenging aspect of pose estimation is the symmetry of objects, as it complicates both the process of labeling the data and constructing methods that can adequately deal with these ambiguities in the object pose. The T-LESS dataset [2] is an industry benchmark for this problem, featuring 30 industry-like objects with multiple symmetries, examples shown in Fig. E.1. Estimating poses from the T-LESS dataset is also more challenging due to the lack of distinguishable features in the textures of the objects, which could otherwise help solve pose ambiguities caused by symmetries.

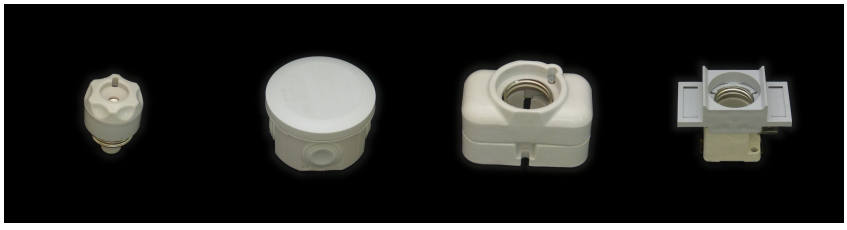


Fig. E.1: Examples from the T-LESS dataset [2]. From left to right: object 2, 30, 5 and 10. The two left-most objects exhibit continuous semi-symmetries where the two others include discrete semi-symmetries.

In this paper we propose an adaptation of the 6D pose estimation approach in [3, 4], that relies on an autoencoder for feature extraction in a codebook-based approach. By replacing their codebook with a neural network and utilizing differential rendering [5], we provide a solution that has a significantly smaller memory footprint, is faster at inference and has improved pose recall when tested on the T-LESS dataset. The solution we propose does not require discretizing poses and it is therefore more easily extendable. Like [3, 4] our method is trained on synthetic RGB images rendered from CAD models or reconstructions and requires no labelled data or predefined symmetries.

Our solution retains many of the properties making [4] interesting for use in robotics. Low inference time allows real-time execution. Only requiring RGB images imposes less restrictions on the hardware. Training on synthetic images makes the process of operating on new objects automatic, with no need for manual labor.

The main contributions of this paper are:

- We propose a depth-based loss function which inherently handles object symmetries without using predefined global symmetries. Our loss does not require a depth sensor as we leverage a differentiable renderer to produce depth maps from CAD models.
- We demonstrate that a pose regression network can be trained to do pose estimation of objects in RGB images using this new loss. We show that this network can replace the codebook used in [4], thereby avoiding discretization.
- We introduce a scheme where the network outputs multiple pose estimates and a weighting between them, and we show that this increases pose recall.
- We show that our pose regression network consumes orders of magnitude less memory, results in faster inference and improves pose estimation recall.

2 Related Work

Traditional methods for pose estimation have commonly relied on matching features, edges and templates [6, 7]. Other approaches use iterative search to find the pose of an object, such as the widely used ICP (Iterative Closest Point) algorithm. Due to their iterative nature, ICP and similar methods are slow, unless optimized for speed [8].

Lately, many of these methods have started to be replaced or complemented by machine learning methods [1, 9–12]. Supervised machine learning

relies on ground truth labels to estimate performance and generate the loss that drives the learning. This makes the way in which that loss is affected by visual traits, like symmetries, an important part of any method for pose estimation.

An object that is symmetrical or semi-symmetrical has many poses that are similar. Such poses should often be treated equally. That means that each input image might match many poses and should not be punished for predicting one of the symmetries rather than the real ground truth pose. For example, a cylinder rotated around its major axis should be treated identically independent of angle, as shown in Fig. E.2a. This problem is especially important for learning algorithms, as they need a consistent way to evaluate if a proposed pose is good or bad.

There are however some types of apparent symmetry that are more complex than others. When parts of an object are occluded, either by external objects or by other parts of the object itself, i.e. self-occlusion, an object can appear identically for many different poses even if it is not actually symmetrical, such as the example in Fig. E.2. Any method that wishes to use learning to estimate poses for these objects need to address how to resolve these ambiguities as well.

Machine learning solutions for 6D pose estimation come in many varieties. Some methods focus on comparing the pose they produce directly with the target pose and predefined global symmetries for each object [13], while other methods utilize 2D or 3D comparisons as loss metrics to train neural networks [3, 4, 14]. In the latter cases, objects with symmetries and semi-symmetries automatically avoid being penalized for miss-classifying along those symmetries. 2D image comparison methods that only consider visible parts of the object can handle apparent symmetries that arise from self-occlusions [3, 4].

Machine learning needs a large amount of data on which to train. For

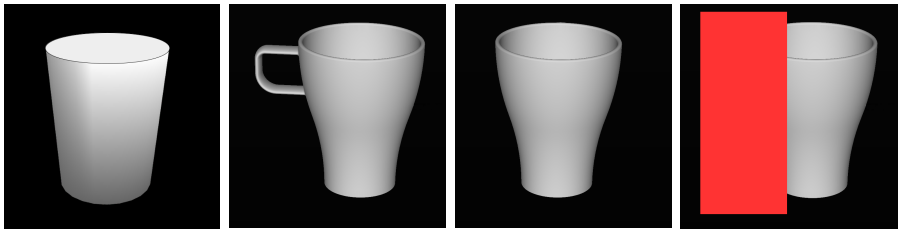


Fig. E.2: (a) Rotationally symmetric objects should be treated equally independent of angle around its major axis. Examples of how symmetries can occur for a mug with a handle. (b) Handle visible, no pose ambiguity. (c) Self-occluded due to a slight rotation and (d) occluded by another object, both of these have ambiguities in pose.

pose estimation, accurate labeling of that data is difficult and costly. To alleviate this problem synthetic data can be produced, for instance by rendering CAD models of target objects together with real images and domain randomization [3, 15].

The work in this paper is based on [3, 4] and relies on a similar synthetic data regime for training. Their work is based on training one or several autoencoders for the objects one wishes to estimate the pose of. The latent space vector produced by the encoder is in their work compared to a codebook of reference latent space vectors, the closest of which becomes the initial pose estimate. When higher accuracy is required, the estimate can be improved by producing extra temporary codebook entries of poses similar to the initial estimate.

While a codebook is in general a good solution, it has a large memory footprint, and it requires a discretization of the predictions. By replacing the codebook in [4] with a pose regression neural network and utilizing differential rendering [5] we show that these problems can be alleviated while simultaneously improving performance.

3 Method

The novel method for 6D pose estimation proposed in this paper is based on the approach initially proposed by [3, 4]. This method retains the central element of the autoencoder, but a neural network replaces the example-based codebook approach for pose regression, as shown in Fig. E.3. The encoder is a feature extractor, producing a 128-dimensional latent vector from an input image of an object. The latent vector is provided as input for the pose regression network. We use the encoder provided by [4]. During training, synthetic images are rendered based on CAD models or reconstructions, with backgrounds and augmentation in accordance with domain randomization [15]. This allows the system to be trained on new objects without manual data collection.

The benefits of replacing the codebook with a neural network are to provide a continuous pose space instead of having to discretize it into a codebook while reducing the memory consumption. Furthermore, it should be more easily extendable than the codebook. The size of a codebook grows exponentially with the number of degrees of freedom, while a pose estimation network only needs to add three more output parameters to expand a rotational representation to include e.g. translations in 3D. Our approach of using a neural network instead of a codebook can be considered a more general and scalable solution as it does not suffer from these limitations.

The pose regression network is structured loosely on the network proposed by [16]. Their network is designed to estimate the pose of an ob-

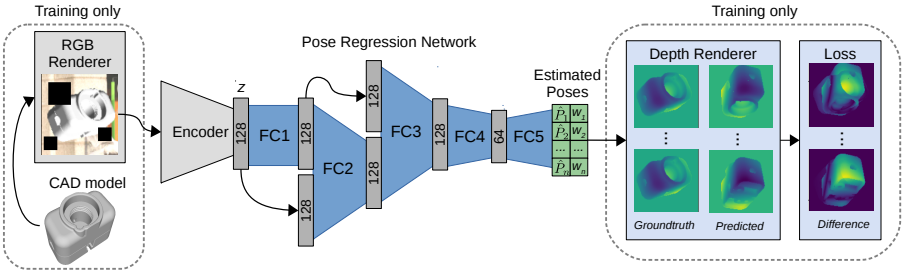


Fig. E.3: Overview of the proposed pose estimation pipeline. During training, synthetic data are continuously generated by rendering RGB images of the relevant object from a CAD model. These RGB images are augmented and fed into the encoder part [4], resulting in a latent vector z. This latent vector z is fed into a pose regression network, consisting of five fully-connected layers, which outputs n poses $\hat{P}_1, \hat{P}_2, ..., \hat{P}_n$ and associated confidences $w_1, w_2, ...w_n$. These confidences are normalized with the softmax function. Depth maps are generated from the poses using a differentiable depth renderer in order to produce the final loss. During inference, both the RGB renderer and the differentiable depth renderer are omitted. We use a pre-trained encoder provided by [4] as the first part of the pipeline.

ject from a feature vector produced from a pre-trained CNN. Their setup is thus similar to how our pose regression network estimates poses from the output of a pre-trained encoder. Our network consists of seven fully connected layers. To maintain training performance with a deeper network, skip-connections between the first three fully connected layers are added.

Note that our network, shown in Fig. E.3, produces not only one but multiple pose estimates along with a confidence for each, which is described in more detail in Sec. 3.2. The network outputs the pose in terms of the representation proposed by [17], as it performs better than regression directly on e.g. rotation matrices or quaternions. This pose representation consists of two vectors in $\mathbb{R}^3$, i.e. 6 parameters in total, which are converted into an orthonormal basis. A 3×3 rotation matrix can then be formed using this basis as column vectors. Using this method also ensures that the resulting rotation matrix is orthogonal.

During training, each predicted pose is used to render depth images using CAD models and a differentiable renderer [5]. Those images are in turn used in the network's loss function. The motivation for this depth-based loss function and how it works are described in the following section.

3.1 Single-Pose Depth Loss

The depth-based loss function used in our pose regression network is heavily inspired by the VSD (Visible Surface Discrepancy) error metric proposed by

[18, 19]. The VSD metric is defined as:

$$e_{\text{VSD}}(\hat{S}, \bar{S}, \hat{V}, \bar{V}, \tau) = \underset{p \in \hat{V} \cup \bar{V}}{\text{avg}} \begin{cases} 0, & \text{if } p \in \hat{V} \cap \bar{V} \text{ and} \\ & |\hat{S}(p) - \bar{S}(p)| < \tau \\ 1, & \text{otherwise} \end{cases} \tag{E.1}$$

where $\hat{S}$ and $\bar{S}$ are depth maps (called distance maps in [18, 19]) based on the estimated pose $\hat{P}$ and the ground-truth pose $\bar{P}$ respectively. Both poses have an associated visibility mask, $\hat{V}$ and $\bar{V}$, that contains the set of pixels actually visible in the given test image I. These visibility masks are found using the real depth map S_I for each test image I. It includes all objects in the scene and can thus be used to determine how those objects occlude each other. The union of these two visibility masks makes up the set of pixels p that are considered by the VSD metric and ensures that only visible pixels are considered. Lastly, τ defines a threshold for the tolerance when comparing the distance maps $\hat{S}$ and $\bar{S}$.

One of the main benefits of the VSD metric, and why it is often used in pose estimation benchmarks, is how it inherently can cope with symmetric objects, as it is solely based on the appearance of the object. I.e., the estimated pose and the ground truth pose may be off by 180^o but the VSD metric would still report a low error for symmetric objects where such an error will not be visible. The VSD metric is able to deal with global symmetries, such as discrete and continuous symmetries, but it is also able to cope with symmetries caused by self-occlusion of the object. Furthermore, the inclusion of the visibility masks, $\hat{V}$ and $\bar{V}$, makes it robust to symmetries caused by occlusion from other objects in the scene.

The loss function proposed in this paper relies on comparing depth maps. In an effort to achieve the same benefits as the VSD metric in Eq. E.1 each pose is evaluated by the loss function:

$$L_{\text{single}}(\hat{S}, \bar{S}) = \underset{p \in \hat{V} \cup \bar{V}}{\text{avg}} \left(\frac{\min(\delta, |\hat{S}(p) - \bar{S}(p)|)}{\delta} \right) \tag{E.2}$$

The value δ serves as a threshold such that there is an upper limit of how much each individual pixel in the distance maps can contribute to the final loss. Without this threshold the loss would be dominated by pixels where the object is present in one of the distance maps but not in the other. This is similar to comparing silhouettes and is hence not ideal as any distance/depth discrepancies on the object itself are dominated by the silhouette.

In order to train the pose regression network with backpropagation, this equation needs to be differentiable with respect to $\hat{P}$ and thereby $\hat{S}$. Regular renderers do not produce differentiable output, but some differentiable renderers have recently become available [5].

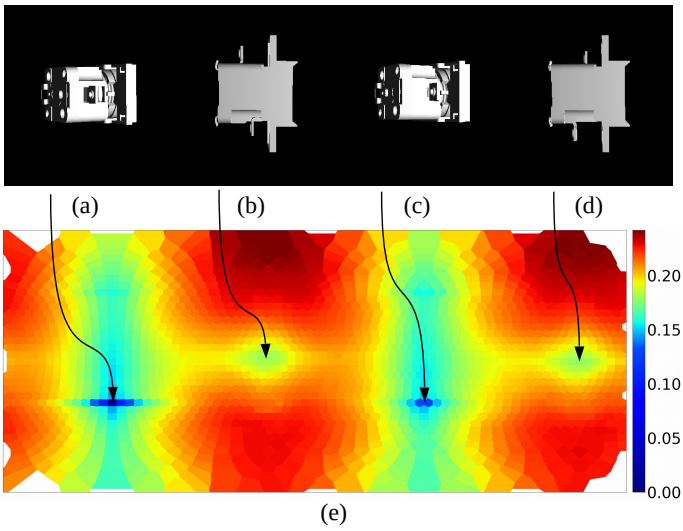


Fig. E.4: Visualization of L_{single} for different pose estimates in relation to specific ground truth pose for object 10. Rotations in the image plane are omitted to get a 2D visualisation. The global minimum (ground truth) pose is (a), and its 180 deg semi-symmetry is (c). The two most isolated non-symmetry local minima are given by (b) and (d). The loss landscape is visualized in (e), ignoring in-plane rotations.

However, one major difference from the VSD metric is how our loss is continuous instead of being limited to the binary set $\{0,1\}$ as is the case of Eq. E.1. This is necessary as the binary set is essentially a step function and thus not differentiable.

3.2 Multi-Pose Depth Loss

The loss function described in Eq. E.2 introduces many local minima, as can be seen in Fig. E.4. These new minima are problematic, as the training of the network easily gets stuck in them, owing to the output being constrained to the SO(3) space. Common training methods only consider the local environment and can therefore not easily overcome these issues. By extending the network to output multiple pose estimates, along with a confidence associated to each, this limitation can be circumvented. The final prediction for a given input is the pose with the highest confidence. As the output confidence distribution changes, the estimated pose can change drastically, if the set of predicted poses are spread over the output space. The benefits of multiple pose estimates are explored in the ablation study in Sec. 4.2.

To bring these concepts together we propose a new loss function L, that

we call "multi-pose loss". It expands on Eq. E.2, and is defined as follows:

$$L(\hat{\boldsymbol{S}}, \bar{S}, \hat{\boldsymbol{P}}) = L_{\text{pose}}(\hat{\boldsymbol{P}}) + \sum_{i=1}^{n} L_{\text{single}}(\hat{S}_i, \bar{S}) \cdot (\gamma + w_i) \tag{E.3}$$

where w_i is the confidence associated with the i'th estimated pose $\hat{P}_i$ for the set of poses $\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{S}}$ the set of depth maps associated with those poses and n is the number of poses output by the network. The confidences, $w_1, w_2, ...w_n$, predicted by the network are normalized using the softmax function. The γ parameter ensures that pose estimates with a near zero confidence also contribute to the loss. This is necessary to make sure that the network improves all pose estimates.

The L_{pose} term in Eq. E.3 forces the network to spread its pose estimates by penalizing poses that are too similar. This loss is defined as follows:

$$L_{\text{pose}}(\hat{\boldsymbol{P}}) = \frac{\sum_{R_A \in \hat{\boldsymbol{P}}} \left(\sum_{R_B \in \hat{\boldsymbol{P}}} \Delta(R_A, R_B) \right)}{n^2} \tag{E.4}$$

where R_A and R_B are rotation matrices converted from the 6D pose representation, and $\hat{\boldsymbol{P}}$ is the set of predicted poses from the network. The function $\Delta(R_A, R_B)$ is a measure of similarity between the two rotation matrices as shown in Eq. E.5.

$$\Delta(R_A, R_B) = 1 - \frac{\min(\phi, \theta)}{\phi} \tag{E.5}$$

with

$$\theta = \arccos\left(\frac{Tr(R_B R_A^T) - 1}{2} \right) \tag{E.6}$$

where $R_B R_A^T$ is the rotation matrix needed to transform R_A to R_B, and $Tr(...)$ is the trace of this matrix. This similarity measure is essentially a conversion of the rotation matrix $R_B R_A^T$ into its corresponding axis-angle representation, while ignoring the axis of rotation. The threshold ϕ serves as a boundary, with rotation matrix pairs that differ by ϕ or more not contributing any loss, while loss is maximized for rotation matrices that are identical. It ensures that the pose regression network has some leeway when predicting the multiple poses instead of just spreading them uniformly, while punishing poses that are close to one another.

3.3 Training

Our method is trained solely on synthetic data. Objects from the T-LESS dataset [2] are rendered in different poses, randomly sampled in SO(3) based on a uniform sampling of quaternions as done by [3]. CAD models of the

objects from the dataset are rendered using OpenGL and the resulting images are augmented in a similar way to [3].

A shared encoder is used for all objects as proposed by [4] and we use the publicly available pre-trained encoder they supply[1]. It is trained on 3D reconstructions of object 1-18 from the T-LESS dataset.

A separate pose regression network is trained for each individual object using a depth max of $\delta = 30$mm and a pose similarity threshold of $\phi = 0.7$ radians (i.e. $\approx 40^o$). We render $n = 10$ poses, per input image and the minimum loss weight for each pose is $\gamma = 0.01$. Each pose regression network is trained for 200 epochs of 10.000 samples each, using a learning rate cycle [20] between 0.005 and 0.0005. All these training parameters are selected through trial-and-error.

It should be noted that all weights in the pre-trained encoder are frozen when training our pose regression networks. This is done to ensure that any differences in performance are directly linked to replacing the codebook-based approach by [4] with our pose regression network, rather than additional training of the encoder.

4 Evaluation

In the following we evaluate our approach against the one proposed in [4]. We test against the publicly available codebooks and pre-trained encoder from [4], the same encoder used in our solution. The methods are compared on their ability to predict correct poses and performance in terms of memory consumption and inference time.

The pose prediction performance of our method is evaluated on real-world images from the T-LESS dataset [2] using the scripts provided as part of the BOP benchmark [21]. We report the recall of each object averaged across different thresholds for the VSD metric e_{VSD} (defined in Eq. E.1) and different tolerance thresholds τ.

In this paper we focus on the correctness of the estimated rotation of the object pose. Errors related to the translation estimate are hence ignored by using the ground truth translation at all times during evaluation for all approaches. Furthermore, the ground truth bounding boxes are used to make the results independent of any errors introduced by an object detector.

4.1 Pose Estimation Performance

In terms of pose recall, the results in Table E.1 show that our method outperforms [4] on average. Our approach has a higher performance when evaluating on object 1-18 in comparison to object 19-30. This is expected as the

[1] `github.com/DLR-RM/AugmentedAutoencoder/tree/multipath`

pre-trained encoder is only trained on object 1-18. The approach by [4] suffers similarly and to a greater extent than our method, as shown in Table E.1.

In Fig. E.5a an example of our predictions are superimposed on a test image. Here, objects 5 (yellow), 6 (magenta) and 7 (green) all match the target well, even though object 7 has a bad overall recall when compared to the two other objects, as seen in Table E.1. This discrepancy could be explained by instances as the one found in Fig. E.5b where the pose prediction for object 7 failed, likely due to the partial occlusion by the two objects in front of it.

Occlusion is of course a challenging scenario in general for pose estimation, but our approach appears to be able to handle it well for some objects. An example of such is the cylinder-shaped objects in Fig. E.6b, where the predicted poses appear correct even for heavily occluded objects.

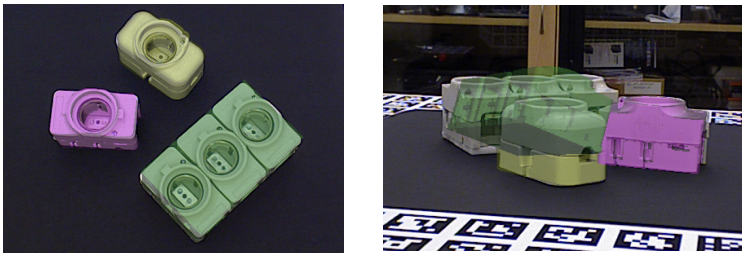
In general, our approach performs better on cylinder-shaped objects with continuous symmetries. This is exemplified in Fig. E.6a where the pose predictions for the box-shaped objects do not appear to fit as well with the

Table E.1: Average VSD recall for the T-LESS primesense test dataset, for our solution and the codebook-based solution [4]. The table is split in two parts; object 1-18 for which the encoder was trained and object 19-30 not seen before by the encoder. Finally, the average VSD recall across all objects is listed in the lower right corner. Results from our method are shown by the mean and standard deviation from three experiments.

Obj.	Codebook	Ours
01	37.82	**51.84** ± 2.8
02	51.88	**63.74** ± 1.8
03	62.87	**71.53** ± 3.3
04	56.00	**62.66** ± 3.5
05	77.18	**80.82** ± 0.3
06	**68.04**	66.71 ± 4.6
07	65.18	**65.68** ± 4.9
08	**63.11**	61.21 ± 0.8
09	**68.96**	55.66 ± 0.5
10	**58.55**	54.14 ± 2.0
11	**52.15**	51.48 ± 2.4
12	**62.19**	56.58 ± 1.6
13	63.56	**64.21** ± 5.0
14	57.29	**63.01** ± 1.2
15	64.91	**66.37** ± 3.8
16	**75.82**	73.16 ± 2.7
17	76.62	**77.72** ± 0.9
18	**71.26**	62.71 ± 2.0
mean	62.97	**63.85** ± 1.2

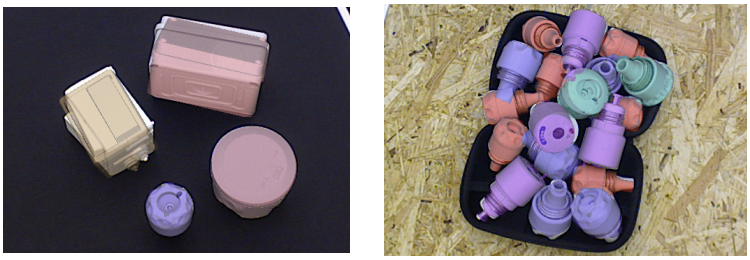
Obj.	Codebook	Ours
19	51.19	**54.15** ± 1.7
20	**40.71**	35.96 ± 1.6
21	43.25	**43.31** ± 1.4
22	**38.15**	32.03 ± 0.5
23	39.18	**56.68** ± 1.1
24	58.97	**61.93** ± 3.3
25	**69.86**	63.08 ± 1.6
26	57.94	**58.87** ± 2.3
27	68.09	**77.62** ± 1.2
28	68.06	**73.33** ± 1.3
29	76.43	**80.67** ± 0.7
30	77.81	**83.41** ± 2.1
mean	57.47	**60.09** ± 0.4

All	Codebook	Ours
mean	60.77	**62.34** ± 0.9



(a) Scene 2, image 170. **(b)** Scene 2, image 490.

Fig. E.5: Colorized renditions of pose predictions superimposed onto images from the T-LESS test dataset. In (a) the poses for objects 5 (yellow), 6 (magenta) and 7 (green) all fit well, but in (b) the pose of object 7 is severely wrong. This is probably due to occlusion.



(a) Scene 1, image 190. **(b)** Scene 20, image 10.

Fig. E.6: Colorized renditions of pose predictions superimposed onto images from the T-LESS test dataset. The pose predictions for the cylindrical objects in (a) are better than those for the rectangular objects. In (b) cylinder-shaped objects are well predicted, even though it is a complicated scene with a lot of occlusion.

test image as the cylinder-shaped objects.

The observation that our approach does better on cylinder-shaped objects is further supported by Table E.2, showing the average recall when dividing the T-LESS test dataset into objects with continuous symmetries and objects without. For objects with continuous symmetries, i.e. cylinder-shaped objects, our method outperforms [4] by a clear margin. For non-cylindrical objects there is little difference in performance between the methods.

The performance discrepancy between objects with continuous symmetries and those with discrete symmetries could be because cylinder-shaped objects are easier than other objects to estimate the pose for. This is sensible, as objects with a continuous symmetry around an axis essentially ignore any rotation around that axis. The number of degrees of freedom in the pose estimation problem are thus less for objects with continuous symmetries. This pattern of higher performance for cylindrical objects is also present in our baseline experiments using the method in [4], but to a lesser extent. However, for our proposed approach, the difference in performance

Table E.2: Average VSD recall for the T-LESS primesense test dataset divided into objects with continuous symmetries and objects with discrete symmetries. Both our and the codebook-based approach performs better on objects with continuous symmetries. However, the difference in performance between continuous and discrete symmetries is more pronounced for our method. Our results are shown with mean and standard deviation as in the previous table.

	Codebook [4]	Ours
Continuous symmetries	62.14	**67.23** ± 2.7
Discrete symmetries	**59.97**	59.51 ± 0.3

between objects with and without continuous symmetries is much more pronounced than for [4]. A possible explanation could be that our depth-based loss landscape exhibits less of the problematic local minima for objects with continuous symmetries than for those without.

4.2 Multi-Pose Ablation Study

Through an ablation study we show that using the multi-pose depth loss (in this case with 10 poses) increases the performance of the pose prediction recall considerably compared to the single-pose depth loss, as shown in Table E.3. As we discussed earlier in Sec. 3.2, the single-pose depth loss may be more prone to get stuck in local minima during training.

Table E.3: Average VSD recall for the T-LESS primesense test dataset with the single-pose loss function and with the multi-pose loss function (10 poses). Multiple poses increases performance considerably, especially for objects with discrete symmetries. Results are shown with mean and standard deviation as in previous tables.

	1 pose	10 poses	improvement
Continuous symmetries	57.37 ± 1.6	**67.23** ± 2.7	9.86
Discrete symmetries	50.62 ± 0.9	**59.51** ± 0.3	8.89
All objects	53.10 ± 0.6	**62.34** ± 0.9	9.24

4.3 Memory Consumption

A comparison of the memory consumption between our approach and the one by [4] is shown in Table E.4. The memory consumption of both the encoder and the codebook are taken directly from [4] while the memory

consumption of our pose regression network is found by calculating the theoretical size of the network and confirming it in a PyTorch implementation. Our approach consumes significantly less memory than [4]. The codebook is replaced entirely by the pose regression network, where the latter consumes $\approx$ 70 times less memory.

Loading the necessary encoder, codebooks, and pose regression network for all 30 T-LESS objects would hence require $\approx$ 1365 MB for [4] and only $\approx$ 33 MB for our method. The relative difference gets larger as the number of objects increases. It should be noted that the reported memory consumption does not include the overhead of loading the different machine learning frameworks, such as TensorFlow and PyTorch, into memory.

Table E.4: Memory consumption during inference of 30 objects. Our solution consumes $\approx$ 40 times less memory than the codebook-based solution.

	Encoder	**Codebook**	**Pose Regression Network**	**Total**
Codebook [4]	15 MB	30×45 MB	-	1365 MB
Ours		-	30×0.6 MB	33 MB

4.4 Inference Time

The inference time of our approach, implemented in PyTorch, is evaluated against the public codebase by [4], implemented in TensorFlow. All timings were measured on a laptop equipped with the following hardware: an i7-7700HQ CPU (2.80GHz) and a NVIDIA GTX 1060 6GB GPU. Note that any measurements related to the projective distance calculation originally mentioned by [4] have been excluded as it is only needed for translation estimation.

Our approach achieves real-time performance with an inference time of $\approx$ 6.2 ms, to estimate the pose of an object. This is an improvement over the current state-of-the-art codebook-based approach by [4] which takes $\approx$ 7.0ms per object. Replacing the codebook-based approach, and thereby both the cosine similarity and nearest neighbor computations, with our network, decreases inference time slightly. The computation time for the encoder should be nearly identical for both approaches as the exact same architecture is used with the only exception being the deep learning framework. Note that the slight increase in computation time when comparing to the measurements reported in [4] is due to the differences in hardware used in the two evaluations.

5 Future Work

The method in this paper predicts the rotation of each object, given a bounding box placing the object in the image. We base our output on a rotation matrix, but thanks to the differential rendering scheme this can easily be extended to output both a rotation and translation estimate instead. This could then be used to do small translation corrections within the bounding box or even determine the full translation in the input image if larger images are provided to the encoder. We expect this would improve the final 6D pose prediction without costly fine-tuning procedures.

Another natural extension for our method is to replace the individual pose regression networks for each object with a single shared pose regression network. This would decrease memory consumption further and it is possible that a pose regression network trained on all objects simultaneously would generalize better. Another benefit of using a shared pose regression network is that it does not require classification of the detected objects as our pose regression network could be trained to also perform the classification.

Our current results are based on freezing the weights of the encoder during training of our pose regression networks. It may be possible to increase the performance of our approach by fine-tuning the encoder or some part of it while training the pose regression network. Another unexplored option is to increase the size of the latent space produced by the encoder as it could increase performance of our pose regression network. Keeping the latent space small makes sense for the codebook-based approach by [4] as the memory consumption of the codebook scales linearly with it. Our approach is much less affected by the size of the latent space.

In many applications where our method would be useful, for instance real-time robotics, inference is done on video rather than independent images. In that context the system can be improved by integrating the temporal aspect, though a particle filter or similar methods [22]. For such a solution, multiple candidates from the multi-pose approach can be utilized.

Finally, we would also like to explore how each pose in our multi-pose solution behaves as a function of the pose in the input image. One question is whether each pose estimate is localized to a certain pose region, while the confidence jumps between them, or if the pose estimates vary more as the input changes. Analysis similar to the principal component analysis done by [4] could reveal this, as well as strengths and weaknesses of our approach.

6 Conclusion

In this paper, we proposed a novel multi-pose loss function to train a neural network to estimate the rotation of an object from an RGB image. This

loss is constructed such that it accounts for any symmetries, an important issue in pose estimation, without relying on predefined symmetries. Our loss is inspired by the VSD (Visible Surface Discrepancy) metric and relies on evaluating the estimated pose by depth comparison. This solution only requires RGB images as input, as the depth maps for the loss are produced by differential renderings of CAD models.

Our network is trained purely on synthetic data. We expand upon an existing state-of-the-art method which utilizes an encoder and codebook to estimate poses [4]. We show that our pose regression network can replace the codebook entirely by directly estimating poses from the output of the encoder. By making our network output multiple poses together with confidences that selects one of them, we show that the recall, as measured by the VSD metric, can be increased. When training our network on top of a pre-trained encoder, shared for all objects, we get a solution that requires a fraction of the memory and has higher pose recall than the state-of-the-art codebook-based approach. It is slightly faster and is not limited by a discretization. Our solution retains or improves many of the interesting properties for robotic applications such as real-time inference, low memory usage, training on synthetic data and only requiring RGB images.

Relying on a neural network instead of a codebook should also make our approach more easily extendable. For instance, integrating a translation estimate into the pose regression network, or training a single pose regression network for multiple objects, instead of having separate networks for each object. These extensions are left for future work.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP).

References

[1] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6d object localization," in *ECCV Workshops*, 2020.

[2] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," *WACV*, 2017.

[3] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from rgb images," in *ECCV*, September 2018.

[4] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *CVPR*, June 2020.

[5] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.

[6] S. Hinterstoisser, V. Lepetit, N. Rajkumar, and K. Konolige, "Going further with point pair features," in *ECCV*, 2016, pp. 834–848.

[7] J. Vidal, C. Lin, and R. Martí, "6d pose estimation using an improved method based on point pair features," in *ICCAR*, 2018, pp. 405–409.

[8] B. Grossmann and V. Krüger, "Fast view-based pose estimation of industrial objects in point clouds using a particle filter with an icp-based motion model," in *INDIN*, 2017, pp. 331–338.

[9] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *CVPR*, June 2019.

[10] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *CVPR*, June 2020.

[11] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *CVPR*, June 2020.

[12] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *CVPR*, June 2020.

[13] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, *CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation*. Springer International Publishing, 2020, pp. 574–591.

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

[15] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017, pp. 23–30.

[16] S. Mahendran, H. Ali, and R. Vidal, "3d pose regression using convolutional neural networks," in *ICCV Workshops*, 2017, pp. 2174–2182.

[17] Y. Zhou, C. Barnes, L. Jingwan, Y. Jimei, and L. Hao, "On the continuity of rotation representations in neural networks," in *CVPR*, June 2019.

[18] T. Hodaň, F. Michel, E. Brachmann, W. Kehl, A. G. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "Bop: Benchmark for 6d object pose estimation," in *ECCV*, Cham, 2018, pp. 19–35.

[19] T. Hodaň, J. Matas, and Š. Obdržálek, "On evaluation of 6d object pose estimation," in *ECCV Workshops*, G. Hua and H. Jégou, Eds., 2016, pp. 606–619.

[20] L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

[21] T. Hodaň, F. Michel, E. Brachmann, W. Kehl, A. Glent Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "BOP: Benchmark for 6D object pose estimation," *ECCV*, 2018.

[22] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking," in *RSS*, 2019.

Paper F

A Shared Pose Regression Network for Pose Estimation of Objects from RGB Images

Stefan Hein Bengtson, Hampus Åström, Thomas B. Moeslund, Elin A. Topp, and Volker Krueger

The paper was presented at the
The 16th International Conference on Signal Image Technology and Internet Based Systems, October 19-21, 2022.

Abstract

In this paper we propose a shared regression network to jointly estimate the pose of multiple objects, replacing multiple object-specific solutions. We demonstrate that this shared network can outperform other similar approaches that rely on multiple object-specific models by evaluating it on the T-LESS dataset using the VSD (Visible Surface Discrepancy). Our approach offers a less complex solution, with fewer parameters, lower memory consumption and less training required. Furthermore, it inherently handles symmetric objects by using a depth-based loss during training and can predict in real-time. Finally, we show how our proposed pipeline can be used for fine-tuning a feature extractor jointly on all objects while training the shared pose regression network. This fine-tuning process improves the pose estimation performance.

1 Introduction

6D pose estimation entails identifying both the orientation and position of an object. These two pieces of information are useful in multiple scenarios, for instance, the pose of an object could be used for a robotic manipulator to infer possible ways to interact with that object such as grasping.

One aspect that complicates the process of pose estimation is the presence of symmetries in some objects, making it hard or impossible to distinguish some poses from each other. This is exemplified in the T-LESS dataset [1], which features multiple highly symmetric objects, as shown in Fig. F.1. The problem of symmetries is further complicated for the type of objects found in T-LESS, as they lack textures that could help resolve such ambiguities.

Another important aspect of pose estimation is the context in which it is used, as there is often a trade-off between the accuracy of the estimates and the inference time. For instance, an approach [2] can produce highly accurate pose estimates but require several seconds to process a single image crop. This will not be ideal for some robotic applications, like bin-picking, where speed is key. Spending several seconds per object in a scene such as the ones shown in Fig. F.1, would for many applications be unacceptable. In such a scenario it may be more preferable to use an approach capable of running in real-time at the cost of a lower pose estimation accuracy [3].

This paper concerns the cases where speed is of the essence. Currently, one of the fastest [3, 4] approaches rely on a codebook-based approach [5] which can achieve a high frame rate in most scenarios, making it applicable for purposes requiring real-time execution. This approach has been improved by replacing the codebooks by small object-specific pose regression networks [6], while relying on a pre-trained feature extractor from the previous approach. By using these small regression networks, pose estimation

Fig. F.1: Four scenes from the T-LESS dataset [1] containing highly symmetric objects without any texture. The dataset contains a total of 30 objects across 20 scenes.

performance is increased and memory consumption is reduced, while maintaining a low inference time. Another benefit of these approaches is that they implicitly account for any symmetries that an object might have by considering the visual similarity of the object when comparing poses.

In this paper we show how a single shared pose regression network can replace multiple object-specific networks [6]. While doing so, we not only improve pose estimation performance, but also reduce complexity, memory consumption and training time.

The main contributions of this paper are:

- We propose a single shared pose regression network, replacing multiple separate object-specific networks. This reduces training time and memory consumption.
- We evaluate our method against two other approaches, multiple object-specific networks and multiple object-specific codebooks.
- We show that fine-tuning the pre-trained feature extractor improves the shared networks pose estimation performance noticeably.
- We show that a shared network can reduce training data needed by a third and memory usage by half, while running in real time and slightly improving VSD recall.

2 Related Work

Traditional methods relying on handcrafted descriptors have been successful in the area of pose estimation in the recent decade, with many of these methods relying on depth information for calculating hand-crafted features based on 3D points [7, 8]. These extracted features are used in various matching and voting schemes in order to produce pose estimates to align the input data with predefined models of the objects. However, computing these features and the matching process are often quite computationally heavy, and it can take several seconds to process a single object.

More recent approaches for pose estimation harness the power of deep learning to produce pose estimates directly from RGB images of the objects of interest [9, 10]. The drawback of these approaches is the huge amount of training required, which prompts the need for massive amounts of training data and expensive computing power.

Furthermore, the process of training these approaches may also prove troublesome for objects with symmetries due to the resulting pose ambiguities. For instance, training samples may be labeled as having widely different poses despite looking visually very similar due to symmetries, such as in Fig. F.2a and F.2b. Some approaches try to counteract this issue by relying on predefined symmetries for each object, such that the pose with lowest error is always chosen from the set of symmetric poses [9, 10]. However, some symmetries are not easily pre-defined as they might occur due to occlusion as illustrated in Fig. F.2c. Furthermore, identifying pre-defined symmetries often relies on a manually selected threshold specifying whether two poses are similar enough to be considered symmetric [4, 9]. This approach does hence fail to encompass the strength of the symmetries, for instance, how visually similar two poses might be. For instance, the poses in Fig. F.2a and F.2b may or may not be considered a symmetry depending on how this threshold is set.

Yet another solution is to avoid the issues of symmetries by confining the set of poses in the training data such that any symmetries are avoided altogether [11, 12]. This approach is similar to relying on pre-defined symmetries, as it also relies on specific knowledge about the symmetries of each object beforehand and therefore will fail to encompass symmetries caused by occlusion.

A lot of these problems related to symmetric objects can be avoided by focusing on the visual similarity of the object in the different poses instead of focusing on the actual poses. Examples of such includes using a codebook-based approach to identify visually similar poses [5] or training a model to predict poses which just has to be visually similar to the ground truth pose [6]. Both of these approaches can implicitly handle symmetric objects

(a) Rotation: 0^o. **(b)** Rotation: 180^o. **(c)** Occluded.

Fig. F.2: Examples of the same object in multiple poses. The poses depicted in a) and b) appear visually similar despite a difference of 180^o due to the semi-symmetric nature of the object. Finally, the object in c) is occluded causing the remaining visible part to have a high visual similarity with the poses depicted in both a) and b).

and do hence not require pre-defined symmetries for each object.

Recently, pose estimation methods have started to rely on deep learning for solely extracting local feature descriptors across an RGB image of an object. These local descriptors are then used to estimate the pose of the object by solving the PnP (Perspective-n-Point) problem in a RANSAC-like fashion [2, 3, 13]. The actual process of estimating the pose is hence not part of the training of the DNN (deep neural network) thereby avoiding the issue of pose ambiguities caused by object symmetries while achieving state-of-the-art pose estimation performance [2]. However, the process of estimating the pose from these local descriptors is often a slow process which could take several seconds per object, just like some of the traditional approaches based on hand-crafted 3D features [7, 8]. These approaches relying on local features are hence either impractical or infeasible for real-time applications.

Approaches with a decent pose estimation performance but with low inference time are hence still needed [3]. This is especially important in scenarios where real-time execution is important, which is often the case for e.g., robotic applications. Using a codebook-based approach [5] can easily achieve a frame rate of over 20 FPS in most scenarios, making it applicable for purposes requiring real-time execution. This approach was subsequently further improved by replacing the codebooks by small object-specific pose regression networks [6] which preserved the low inference time while reducing memory usage and increasing the pose estimation performance as well.

3 Method

The work in this paper is based on an approach relying on multiple object-specific pose regression networks [6] and shows how this approach can be improved and simplified by replacing the many object-specific networks with

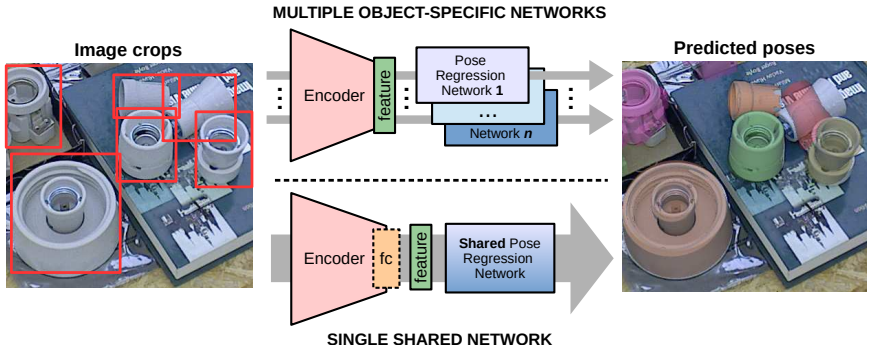


Fig. F.3: *Top*: Multiple object-specific pose regression networks are trained independently of each other [6]. These networks rely on a shared feature extractor in the form of the pre-trained encoder [5]. *Bottom:* We propose to replace all these object-specific networks with a single shared pose regression network. Furthermore, we suggest fine-tuning the last fully-connected layer of the pre-trained encoder while training the shared network jointly on all objects.

one shared pose regression network, as shown in Fig. F.3. We show that it is possible to train this single shared pose regression network for all the objects while simultaneously increasing the overall performance. Both approaches are based on the same feature extractor, in the form of the pre-trained encoder from a codebook-based approach [5]. This paper also includes an exploration of the benefits of fine-tuning parts of the pre-trained encoder for the pose regression task, while training the shared network.

The proposed method assumes input from an object detector in terms of a bounding box and an associated object ID. This is similar to the codebook-based approach [5] and the approach using multiple object-specific networks [6].

3.1 Network Architecture

The structure of the proposed shared pose regression network is illustrated in Fig. F.4. The first part of the network resembles the object-specific networks [6], where a pre-trained encoder from the codebook-based approach [5] is used as a feature extractor for the network. The input for the pose regression part of the network is thus the latent space produced by this encoder. The size of this feature vector, or latent space, is set at 128 as this is considered optimal in previous work [5].

We propose to fine-tune the last fully-connected layer of the encoder based on the hypothesis that it could provide a better feature representation for doing pose estimation later on. This fine-tuning of the encoder is done while training the shared pose regression network.

During training, synthetic RGB images of various objects in random poses

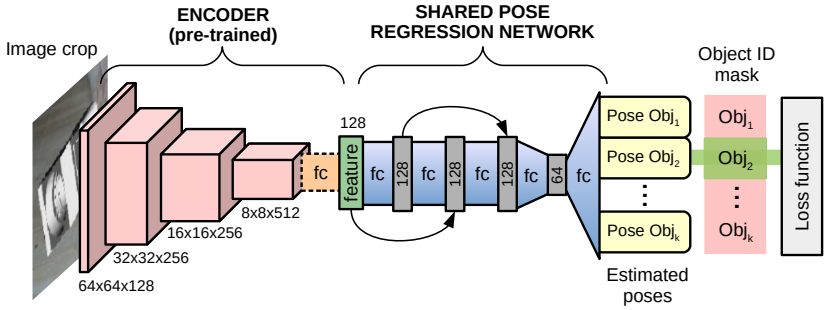


Fig. F.4: The architecture of the shared posed regression network, including the pre-trained encoder used as feature extractor [5]. The network is nearly identical to the original pipeline [6] and includes several skip connections as these were found to increase performance. The exception is the final layer of the network which has been modified to output multi-pose estimates for all the k different object categories, regardless of the class ID of the input. Additionally, a masking scheme is introduced to ensure that only the estimated poses for the correct object ID is propagated further in the pipeline. It is assumed that the object ID is available from a prior detection step.

are continuously rendered from CAD models of the objects and used as training data [5]. This avoids the issue of having to provide massive amounts of hand-labeled data while training the network, and have proven to generalize well to real data [5, 6].

Each of the estimated poses are represented using a 6D vector, as previous studies have shown that it is more stable for pose regression tasks than other representations such as quaternions and rotation matrices [14].

Finally, similar to the approach relying on multiple object-specific pose regression networks [6], only the orientation of the object is estimated by the network. Estimating the translation of the object is not included in the proposed pipeline but can be done similarly to how it was solved in connection to the codebook approach [5], or included as a task for the shared network in future work.

The proposed shared network is able to predict poses for multiple objects by expanding the size of the final output layer, as illustrated in Fig. F.4. The network therefore always outputs a multi-pose [6] estimate for each of the k different objects, where k is the number of objects the network is trained to estimate poses for. Some of the weights in the final fully-connected layer are therefor tied to a certain object and are class specific, while the rest of the model is shared between all the objects.

The size of the output layer in the proposed shared network is scaled in a linear fashion with k. One drawback of this approach is that the number of parameters in the model will increase as the number of object categories increases. However, this increase is negligible compared to the rest of the pipeline. This can be seen in Section 4.2, where the overall memory usage of

the proposed approach is compared against having multiple object-specific networks instead.

Training the proposed shared network relies on a masking scheme to ensure that only the predicted poses for the correct objects are propagated to the loss function. This is illustrated in Fig. F.4, where only one of the pose estimates propagates further to the loss function after applying the object ID mask. This approach assumes that an object ID is provided from a prior object detection step, both during training and inference, in order to apply the mask correctly.

The actual loss is calculated by comparing depth maps of the object in different poses in the same way as the individual pose regression approach [6], as illustrated in Fig. F.5.

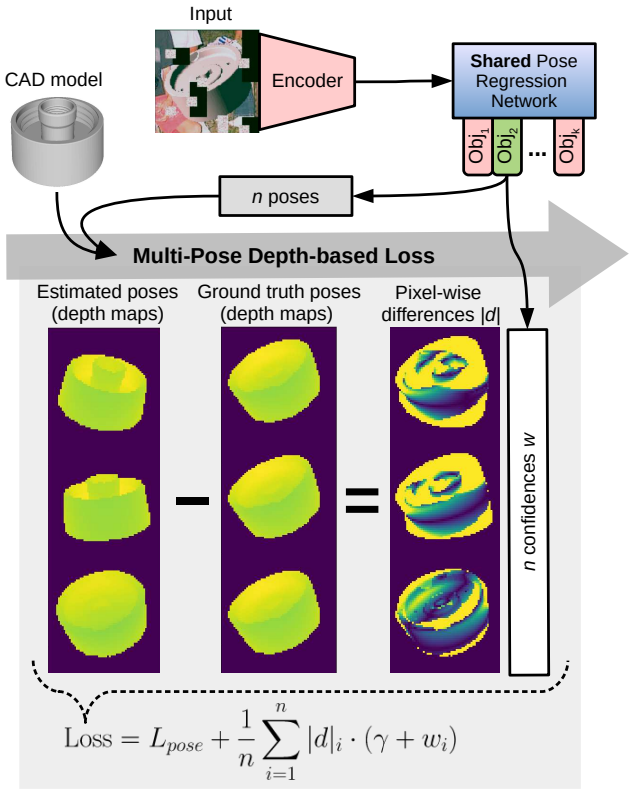


Fig. F.5: The depth-based loss [6] used when training the shared pose regression network. It avoids any symmetry-related issues by using rendered depth maps to measure the similarity of poses. Multiple possible hypotheses are predicted for each pose estimate to avoid getting stuck in local minima and the final loss is a weighted average of these hypotheses using their confidences as weights. The parameter, γ, ensure that each hypothesis contributes to the loss so that it gets updated by back-propagation. The regularization term, L_{pose}, is included to incur a high loss if the pose hypotheses becomes to similar.

This depth-based loss is heavily inspired by the VSD (Visible Surface Discrepancy) metric [15, 16], that is commonly used in pose estimation benchmarks. The idea is to render depth maps of the different objects in the estimated poses and in the groundtruth poses, respectively. These depth maps are then compared in a pixel-wise approach to identify the visual similarity of the estimated versus ground truth poses. The main motivation of this approach is that it implicitly accounts for any symmetries that the objects might have as the depth maps will look similar in that case and hence incur a small error. These depth maps are produced using differential rendering [17] in order to allow back-propagation when training the shared pose regression network.

Furthermore, each estimation consists of multiple hypotheses in order to avoid getting stuck in local minima, which can easily occur in this depth-based loss [6]. Each estimate includes n hypotheses for each of the estimated poses along with a predicted confidence for each hypothesis $w_1, ..., w_n$. This is illustrated in Fig. F.5, where $n = 3$. The final loss is essentially then a weighted average amongst the different pose hypotheses using the predicted confidences as the weights. The parameter, γ, ensures that each pose hypothesis always contributes a little to the loss, so that back-propagation updates all hypotheses during training, and the regularization term, L_{pose}, ensures that the multiple pose hypotheses do not become too similar, which would defeat the purpose of the multiple hypotheses [6]. Finally, this depth-based loss is only used during training of the shared network. No depth maps are rendered during inference.

4 Evaluation

Our approach is evaluated on the T-LESS dataset [1], containing 30 objects that are characterized by a lack of texture and by being highly symmetric. This combination, no texture and many symmetries, is what makes this dataset challenging.

The network is trained on synthetic data using the same scheme as in the original pipeline using multiple object-specific networks [6] but with a fixed learning rate of $= 0.001$ instead of an adaptive one [18]. A fixed learning rate is used as it allows training the network for an unknown amount of epochs until it converges, since the adaptive learning rate scheme originally used for training the object-specific networks requires a fixed number of epochs that has to be specified prior to training. The network is trained until convergence, with 100,000 newly generated synthetic samples each epoch. These samples are generated randomly from the 30 objects in the T-LESS dataset using the CAD models supplied in the dataset.

The vertices of the CAD models are normalized so that all objects have the

same length along their respective largest dimension. This normalization was found necessary in order to avoid some objects dominating the training phase by being bigger than other objects and therefore more likely to incur a bigger loss. Such considerations, regarding some objects being more dominant than others, is not a concern when training a separate network for each object. All the other parameters used during training are identical to the ones used for training the many object-specific pose regression networks [6].

The shared pose regression network described above is evaluated in two variations; with and without fine-tuning of the last layer in the encoder that provides the input to the shared network. We hypothesize that fine-tuning the last layer of the encoder used for feature extraction can increase the performance of the shared pose regression network. Jointly fine-tuning the encoder on all objects is easily done as only one single shared network is trained.

As a comparison, we note that when training multiple object-specific networks [6] it is complicated to jointly fine-tune the encoder in a similar way. It may be possible to train all the object-specific networks concurrently, in order to jointly fine-tune the same encoder in the process. However, such an approach is complex and requires hardware capable of training the many networks concurrently; 30 networks in case of the T-LESS dataset. Alternately, a separate encoder can be fine-tune for each specific object, but this increases the complexity of the system, both in terms of parameters to be trained and in terms of increased memory consumption. However, fine-tuning the encoder in the context of the shared pose regression network proposed in this paper does not incur any of the drawbacks mentioned above and we can therefore evaluate this network both with and without fine-tuning.

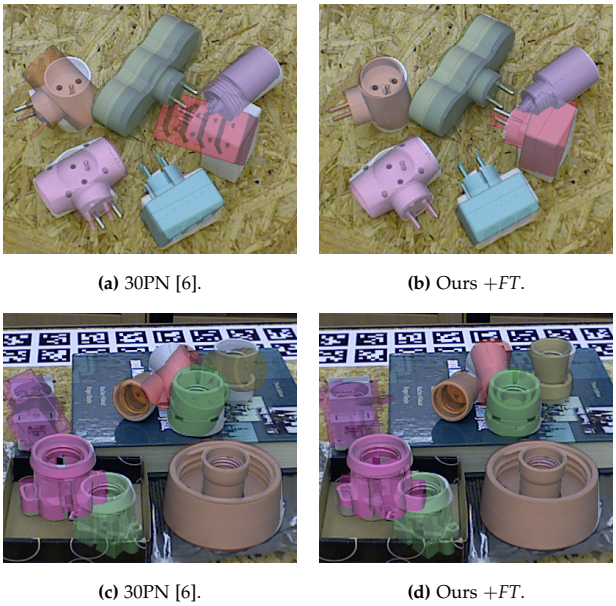
4.1 Results - Pose Estimation

We evaluate our approach on the T-LESS dataset using the VSD metric [15, 16], both with ($+FT$) and without ($\div FT$) fine-tuning of the encoder. Both are trained until convergence, which is achieved at 200 epochs. Our approach is compared against an approach using multiple object-specific pose regression networks [6] and an approach using multiple object-specific codebooks [5], as shown in Table F.1. Both ground truth translations and object IDs are used during the evaluation, as both the proposed approach and the multiple object-specific networks approach [6] do not include translation estimation nor object classification. This is done for all the evaluated approaches to ensure a fair comparison.

From these results, it is clear that fine-tuning parts of the encoder improves the performance noticeably as our approach outperforms both the codebook-based approach [5] and the object-specific model approach [6] on average, when fine-tuning is included. Furthermore, the proposed approach

Table F.1: VSD recall on the T-LESS dataset for our proposed shared pose regression network. The results with (+*FT*) and without (÷*FT*) fine-tuning of the encoder are reported along with previously published results [6] for the object-specific approaches using either 30 codebooks (30CB) [5] or 30 pose regression networks (30PN) [6].

Object	30CB [5]	30PN [6]	Ours +*FT*	Ours ÷*FT*
01	37.82	51.84	**54.00**	41.2
02	51.88	**63.74**	62.12	54.88
03	62.87	71.53	**73.03**	60.25
04	56.00	62.66	**67.14**	56.9
05	77.18	**80.82**	76.26	68.47
06	68.04	66.71	**72.27**	55.99
07	65.18	**65.68**	57.16	50.5
08	**63.11**	61.21	55.21	49.1
09	**68.96**	55.66	53.24	51.87
10	**58.55**	54.14	55.79	31.7
11	**52.15**	51.48	48.09	42.23
12	**62.19**	56.58	54.45	47.79
13	63.56	64.21	**69.19**	59.36
14	57.29	63.01	**67.89**	59.57
15	64.91	66.37	**71.98**	56.3
16	75.82	73.16	**78.25**	71.91
17	76.62	**77.72**	76.77	73.79
18	**71.26**	62.71	61.97	53.26
19	51.19	54.15	**57.50**	44.89
20	40.71	35.96	**43.71**	33.4
21	43.25	43.31	**47.63**	35.1
22	**38.15**	32.03	37.62	22.08
23	39.18	**56.68**	55.50	45.58
24	58.97	61.93	**63.64**	56.93
25	**69.86**	63.08	63.71	52.62
26	57.94	58.87	**60.24**	55.22
27	68.09	**77.62**	69.28	74.67
28	68.06	**73.33**	69.23	69.52
29	76.43	80.67	**83.48**	77.68
30	77.81	83.41	**87.42**	82.42
mean	60.77	62.34	**63.13**	54.51



(a) 30PN [6]. **(b)** Ours $+FT$.

(c) 30PN [6]. **(d)** Ours $+FT$.

Fig. F.6: Examples of pose estimates from the proposed shared network and an approach using object-specific networks [6]. Using the different pose estimates, CAD models are plotted on top of the images from the T-LESS test dataset. The colorization is solely for illustration purposes.

also results in the best performance for 15 out of the 30 objects found in the T-LESS dataset. Without fine-tuning of the encoder, the shared pose regression network performs worse than the other two approaches.

Examples of pose estimates produced using both the proposed shared network and multiple object-specific networks are shown in Fig. F.6. In the first scene (Fig. F.6a and F.6b) the predicted poses appear similar for most objects. Exceptions are object 20 (red) and object 21 (orange) where the object-specific models fail to produce feasible pose estimates. The shared network, on the other hand, produces reasonable pose estimates in both cases. Examples like these contribute to the discrepancy in performance in Table F.1, where the shared network performs the best on both these objects.

In the second scene (Fig. F.6c and F.6d) both approaches appear to perform similarly on the objects in the front region of the scene. Both approaches also struggle with object 10 (magenta, top left) but this particular object is in general difficult, as seen in the results reported in Table F.1. However, the pose prediction from the two approaches differs for the group of objects in the upper right corner, consisting of object 13 (orange), 14 (red), 15 (yellow) and 16 (green). In this case, the shared network produces pose estimates which are better aligned with the input images, particularly for the occluded objects where the multi-network approach fails.

4.2 Results - Other Metrics

Besides the pose estimation recall improvement, the proposed approach also reduces the complexity of the system by using a single shared model instead of multiple different ones. The number of parameters, the main contributor to memory usage during inference, is reduced. In the case of the 30 objects in the T-LESS dataset, the reduction in memory consumption is $\approx 51\%$ compared to multiple object-specific networks [6] and $\approx 98\%$ compared to using codebooks [5].

Additionally, using a single shared model reduces training time. The proposed approach was trained for 200 epochs with 100k samples each, amounting to 20 million samples in total. For comparison, each object-specific pose regression network was trained for 200 epochs with 10k samples each [6], resulting in 2 million samples per object. Training 30 separate pose regression networks thus requires three times as many samples as the shared pose regression network.

Finally, it takes $\approx 6.4ms$ to estimate the pose of an object during inference, for the proposed shared pose regression network. This is comparable to the inference time for the approach using multiple object-specific networks ($\approx 6.2ms$) and slightly better than the codebook-based approach ($\approx 7.0ms$). Note that all the timings in terms of the inference time exclude object detection, which is a necessary prior step for all three approaches. Finally, all timings are measured using the same hardware (i7-7700k and GTX1060). Thus, it is possible to achieve a frame rate of 20 FPS in terms of the pose estimation for scenes with 7 objects or less, even if all estimations are done in sequence.

Finally, the different characteristics for the evaluated approaches are summarized in Table F.2. This includes both pose estimation performance in terms of average VSD recall, inference time, memory usage and number of samples required during training, as discussed in detail in previous sections.

Table F.2: Summary of the main characteristics of our approach with fine-tuning (*+FT*) compared to using codebooks [5] and multiple pose regression networks [6]. **The number of training samples is not reported for the codebook-based approach as only the encoder requires training and is assumed to come pre-trained for all approaches.*

	Avg. VSD recall	Inference time	Memory usage	Training samples
30CB [5]	60.77	7.0ms	1365MB	NA*
30PN [6]	62.61	**6.2ms**	33MB	60M
Ours *+FT*	**63.13**	6.4ms	**16.6MB**	**20M**

5 Conclusion

This paper proposes a shared regression network for pose estimation of different objects, and shows that it can replace approaches with several object-specific solutions. This shared network is evaluated on the T-LESS dataset and a comparison is made to estimators with multiple object-specific models, either in the form of codebooks or pose regression networks. Our approach achieved the highest overall pose estimation recall by fine-tuning the pre-trainer encoder used for feature extraction while training the shared network. This shared network also offers a less complex solution, with fewer parameters and less memory usage, and it requires less training than the method with multiple object-specific networks. We do this while maintaining the main properties of the two other approaches, as our method handles symmetric objects similarly and has a low inference time, making it suitable for real-time applications. These results indicate that our shared model is preferable over approaches relying on multiple object-specific solutions for pose estimation.

6 Future Work

A way to improve the presented work is to consider temporal information, based on the main assumption that the pose of an object does not change in a fraction of a second. Either using various filters [19] or by integrating multiple estimates, in the form of multiple view-points [9]. The benefits of these approaches are promising given the low inference time of the proposed approach, making it possible to produce many pose estimates fast.

Another idea for future work is to include translation estimation, which the approach presented in this work currently lacks, just like the approach using multiple networks [6]. One way could be to infer translation for objects from the size of their bounding boxes in relation to the known size of the objects, from, e.g., the CAD models. However, this approach is very sensitive to noise in the bounding boxes and hence the object detector used [5, 20]. This issue could be counteracted by training a model to estimate adjustments to the bounding box of each object [10].

Yet another avenue for further research could be to expand the presented approach to also predict object IDs as it currently relies on a prior object detection step for this information. Estimating object IDs as part of the shared pose regression network could be based on the idea of visual similarity from depth renderings, just like the pose estimation. Doing so may prove beneficial as wrong predictions in terms of the object ID would be punished less harshly if the objects are visually similar and vice versa.

Finally, it would be interesting to further explore the impact of fine-tuning the pre-trainer encoder, as is essential for the performance of the shared pose regression network in this work. Exploring how similar fine-tuning would impact other approaches is thus another obvious path for future work.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP).

References

[1] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," *WACV*, 2017.

[2] R. L. Haugaard and A. G. Buch, "Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6749–6758.

[3] T. Hodan, D. Barath, and J. Matas, "EPOS: Estimating 6d pose of objects with symmetries," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020.

[4] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6d object localization," in *ECCV Workshops*, 2020.

[5] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *CVPR*, June 2020.

[6] S. H. Bengtson, H. Astrom, T. B. Moeslund, E. A. Topp, and V. Krueger, "Pose estimation from RGB images of highly symmetric objects using a novel multi-pose loss and differential rendering," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep. 2021.

[7] B. Drost, M. Ulrich, P. Bergmann, P. Härtinger, and C. Steger, "Introducing mvtec ITODD - A dataset for 3d object recognition in industry," in *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops*, 2017.

[8] J. Vidal, C.-Y. Lin, X. Lladó, and R. Martí, "A method for 6d pose estimation of free-form rigid objects using point pair features on range data," *Sensors*, vol. 18, no. 8, p. 2678, Aug. 2018.

[9] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, *CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation*. Springer International Publishing, 2020, pp. 574–591.

[10] Z. Li, G. Wang, and X. Ji, "CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[11] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6d: Making RGB-based 3d detection and 6d pose estimation great again," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.

[12] M. Oberweger, M. Rad, and V. Lepetit, "Making deep heatmaps robust to partial occlusions for 3d object pose estimation," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 125–141.

[13] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[14] Y. Zhou, C. Barnes, L. Jingwan, Y. Jimei, and L. Hao, "On the continuity of rotation representations in neural networks," in *CVPR*, June 2019.

[15] T. Hodaň, F. Michel, E. Brachmann, W. Kehl, A. Glent Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "BOP: Benchmark for 6D object pose estimation," *ECCV*, 2018.

[16] T. Hodaň, J. Matas, and Š. Obdržálek, "On evaluation of 6d object pose estimation," in *ECCV Workshops*, G. Hua and H. Jégou, Eds., 2016, pp. 606–619.

[17] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.

[18] L. N. Smith and N. Topin, "Super-convergence: very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

[19] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "PoseRBPF: A rao–blackwellized particle filter for 6-d object pose tracking," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1328–1342, Oct. 2021.

[20] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3d orientation learning for 6d object detection from rgb images," in *ECCV*, September 2018.