Tampere University

PETRI MÄKINEN

Toward Vision-based Control of Heavy-Duty and Long-Reach Robotic Manipulators

Tampere University Dissertations 899

Tampere University Dissertations 899

PETRI MÄKINEN

Toward Vision-based Control of Heavy-Duty and Long-Reach Robotic Manipulators

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Engineering and Natural Sciences
of Tampere University,
for public discussion in the auditorium Pieni sali 1
of the Festia building, Korkeakoulunkatu 8, Tampere,
on 8 December 2023 at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Engineering and Natural Sciences
Finland

Responsible supervisor and Custos	Professor Jouni Mattila Tampere University Finland	
Pre-examiners	Professor Juha Plosila University of Turku Finland	Professor Fransisco Javier Badesa Clemente Universidad Politécnica de Madrid Spain
Opponent	Professor Alessandro Saccon Eindhoven University of Technology The Netherlands	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-3145-0 (print)
ISBN 978-952-03-3146-7 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-3146-7

HIILINEUTRAALI PAINOTUOTE
ClimateCalc CC-000025/FI
PunaMusta Printing

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2023

PREFACE

The research presented in this dissertation was carried out from 2018 to 2023 at the unit of Automation Technology and Mechanical Engineering of the Faculty of Engineering and Natural Sciences at Tampere University.

I want to express my deepest gratitude to Prof. Jouni Mattila for his encouragement and for providing the opportunity to work on this dissertation. I also wish to thank my industry supervisor Sirpa Launis for her support and for providing the initial research problems.

Thanks are also due to Dr. Pauli Mustalahti for his assistance with the experimental measurements. I also wish to thank my brief co-supervisor Dr. Mohammad M. Aref for his scientific insights. For good conversations and moral support, I also thank my numerous colleagues and peers with whom I shared these years.

This work was mutually funded by Tampere University and Sandvik Mining and Construction Oy as a part of the Doctoral School of Industry Innovations. This work also received partial funding from K.F. and Maria Dunderberg's testamentary foundation, for which I am grateful. I also want to acknowledge CIVIT, the Centre for Immersive Visual Technologies at Tampere University, for providing the equipment to assist in this research.

ABSTRACT

Heavy-duty mobile machines are an important part of the industry, and they are used for various work tasks in mining, construction, forestry, and agriculture. Many of these machines have heavy-duty, long-reach (HDLR) manipulators attached to them, which are used for work tasks such as drilling, lifting, and grabbing. A robotic manipulator, by definition, is a device used for manipulating materials without direct physical contact by a human operator. HDLR manipulators differ from manipulators of conventional industrial robots in the sense that they are subject to much larger kinematic and non-kinematic errors, which hinder the overall accuracy and repeatability of the robot's tool center point (TCP). Kinematic errors result from modeling inaccuracies, while non-kinematic errors include structural flexibility and bending, thermal effects, backlash, and sensor resolution. Furthermore, conventional six degrees of freedom (DOF) industrial robots are more general-purpose systems, whereas HDLR manipulators are mostly designed for special (or single) purposes.

HDLR manipulators are typically built as lightweight as possible while being able to handle significant load masses. Consequently, they have long reaches and high payload-to-own-weight ratios, which contribute to the increased errors compared to conventional industrial robots. For example, a joint angle measurement error of 0.5° associated with a 5-m-long rigid link results in an error of approximately 4.4 cm at the end of the link, with further errors resulting from flexibility and other non-kinematic aspects. The target TCP positioning accuracy for HDLR manipulators is in the sub-centimeter range, which is very difficult to achieve in practical systems. These challenges have somewhat delayed the automation of HDLR manipulators, while conventional industrial robots have long been commercially available. This is also attributed to the fact that machines with HDLR manipulators have much lower production volumes, and the work tasks are more non-repetitive in nature compared to conventional industrial robots in factories.

Sensors are a key requirement in order to achieve automated operations and eventually full autonomy. For example, humans mostly rely on their visual perception in work tasks, while the collected information is processed in the brain. Much like humans, autonomous machines also require both sensing and intelligent processing of the collected sensor data. This dissertation investigates new visual sensing solutions for HDLR manipulators, which are striving toward increased automation levels in various work tasks. The focus is on visual perception and generic 6 DOF TCP pose estimation of HDLR manipulators in unknown (or unstructured) environments. Methods for increasing the robustness and reliability of visual perception systems are examined by exploiting sensor redundancy and data fusion. Vision-aided control using targetless, motion-based local calibration between an HDLR manipulator and a visual sensor is also proposed to improve the absolute positioning accuracy of the TCP despite the kinematic and non-kinematic errors present in the system. It is experimentally shown that a sub-centimeter TCP positioning accuracy was reliably achieved in the tested cases using a developed trajectory-matching-based method.

Overall, this compendium thesis includes four publications and one unpublished manuscript related to these topics. Two main research problems, inspired by the industry, are considered and investigated in the presented publications. The outcome of this thesis provides insight into possible applications and benefits of advanced visual perception systems for HDLR manipulators in dynamic, unstructured environments. The main contribution is related to achieving sub-centimeter TCP positioning accuracy for an HDLR manipulator using a low-cost camera. The numerous challenges and complexities related to HDLR manipulators and visual sensing are also highlighted and discussed.

CONTENTS

List of Figures

ABBREVIATIONS

AI	Artificial Intelligence
CIVIT	The Centre for Immersive Visual Technologies (at Tampere University)
CPD	Coherent Point Drift
DH	Denavit-Hartenberg
DOF	Degrees of Freedom
DSII	Doctoral School of Industry Innovations
FMM	Von Mises-Fisher Mixture Model
GMM	Gaussian Mixture Model
HDLR	Heavy-Duty, Long-Reach
IBVS	Image-based Visual Servoing
ICP	Iterative Closest Point
IMU	Inertial Measuring Unit
IP	Ingress Protection
LIDAR	Light Detection and Ranging
OOI	Object(s) of Interest
ORB	Oriented FAST and Rotated BRIEF
PBVS	Position-based Visual Servoing
RADAR	Radio Detection and Ranging
RP	Research Problem
SIFT	Scale-Invariant Feature Transform
SLAM	Simultaneous Localization and Mapping

SURF	Speeded Up Robust Features
TCP	Tool Center Point
UWB	Ultra-Wideband
VO	Visual Odometry

ORIGINAL PUBLICATIONS

P-I P. Mäkinen, M. M. Aref, J. Mattila, and S. Launis, "Application of simultaneous localization and mapping for large-scale manipulators in unknown environments," in *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2019. DOI: 10.1109/CIS-RAM47153.2019.9095770.

P-II P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Redundancy-based visual tool center point pose estimation for long-reach manipulators," in *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2020, pp. 1387–1393. DOI: 10.1109/AIM43001.2020.9159022.

P-III P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Probabilistic camera-to-kinematic model calibration for long-reach robotic manipulators in unknown environments," in *2022 IEEE 17th International Conference on Advanced Motion Control (AMC)*, 2022, pp. 48–55. DOI: 10.1109/AMC51637.2022.9729259.

P-IV P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Model-free sensor fusion for redundant measurements using sliding window variance," in *2022 22st International Conference on Control, Automation and Systems (ICCAS)*, 2022, pp. 1481–1487. DOI: 10.23919/ICCAS55662.2022.10003720.

UNPUBLISHED MANUSCRIPT

P-V P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, “Vision-aided precise positioning for long-reach robotic manipulators using local calibration,” 2023.

1 INTRODUCTION

In robotic control and tool center point (TCP) positioning, it is typically required that each joint is equipped with a high-precision sensor to measure the angle or the linear position of the joint. For high-precision control, sensors with high accuracy and repeatability are required. Especially in heavy-duty, long-reach (HDLR) manipulators, even a small error per joint results in a considerable error at the TCP of the serial chain kinematic structure. For example, a 0.5° orientation error of a 5-m-long rigid link results in a 4.4-cm positioning error at the end of the link.

In HDLR manipulators, the number of actuator degrees of freedom (DOF) is relatively high, meaning that the number of sensors per robotic manipulator is also high. Combined with the fact that the joints are not uniform, requiring multiple sensor sizes and mounting solutions, the overall cost and complexity of the sensor system can become considerable. Furthermore, HDLR manipulators attached to mobile machines work in harsh, unstructured environments, such as mines, forests, and fields. Therefore, rugged sensors with high ingress protection ratings (i.e., IP67) are required. In practice, the sensors must be dust- and waterproof.

1.1 Motivation

Mobile work machines represent a significant field of industry, and they come in many different configurations and sizes with respect to their on-board robotic HDLR manipulators. Figure 1.1 illustrates two examples of such HDLR manipulators. The first one is a HIAB articulated crane with an additional 3 DOF wrist. The second one is used in mining, where HDLR manipulators range from approximately 10-15-m-reaching tunneling machines to small-scale 1-2-m-reaching surface drilling platforms. The annual production volume for a specific machine type can be very low. Therefore, these HDLR manipulators requiring relatively high precision, with varying configurations, would benefit from sensor system solutions that reduce the manufacturing, assembly, and maintenance costs of these machines.

A typical HDLR manipulator attached to a mobile machine can have a payload-mass-to-own-weight ratio of over 1:1. For example, the HIAB HDLR manipulator in Figure 1.1 has a 4.7-m reach (without the wrist and the extensions) and weighs 445 kg, while the maximum payload is 600 kg. For comparison, conventional industrial robots working in known (structured) environments typically have a ratio of 1:10 or smaller [1]. In addition, a key differentiation between HDLR manipulators and conventional industrial robots is that the former often have prismatic joints for extended reach, and they work in dynamic, unstructured environments.

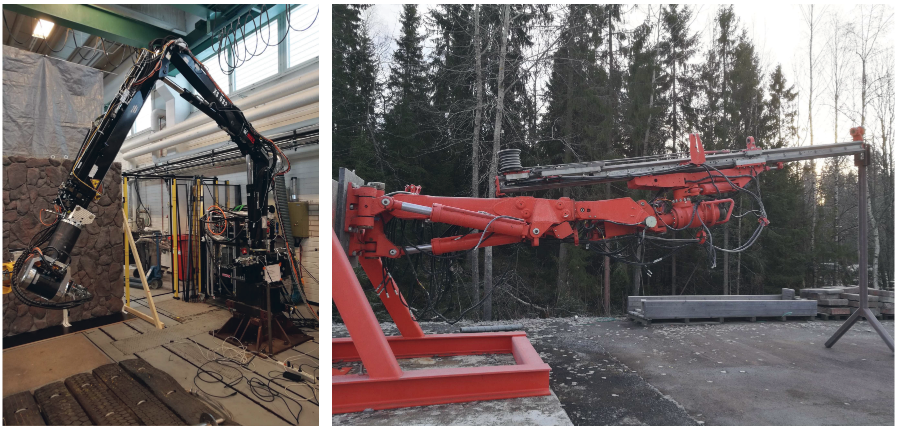


Figure 1.1 i) (left) A HIAB HDLR manipulator that can be mounted to a vehicle for heavy lifting. ii) (right) A Sandvik 8 DOF HDLR manipulator used in mining that can carry a heavy-duty rock drill.

The traditional method of operating HDLR manipulators is a human controlling joint-specific valves to drive each joint separately. The more advanced method is controlling the TCP directly instead of controlling individual joints [2]–[4], thus making the operator's work easier. Controlling the TCP requires a kinematic model of the manipulator, which is commonly formulated using Denavit-Hartenberg (DH) parameters [5]. The forward kinematic formulation computes the TCP pose (3 DOF position and 3 DOF orientation) variables using the joint states, hence requiring joint sensors. Inverse kinematic computation allows the transformation of a desired TCP pose into respective joint states, which is a prerequisite for control. Prevalent commercial systems employ rigid-body-based kinematic models, which, especially in the case of HDLR manipulators, results in significant errors at the TCP due to kinematic and non-kinematic errors. The kinematic errors include inaccuracies in the

DH parameters, for example. The non-kinematic errors include structural deformations, backlash, thermal effects, and sensor resolution. In practice, the assumption of rigidness is mainly valid for conventional industrial robots. For HDLR manipulators, however, the assumption is invalid due to long reaches and high payload-mass-to-own-weight ratios. From a control perspective, the TCP velocities of HDLR manipulators are generally designed to be relatively slow so that the non-rigid structures are quasi-static.

Some HDLR systems employ flexibility compensation computations to obtain more accurate TCP poses. However, compared to conventional industrial robots, the resulting errors are still much larger. The consensus of original equipment manufacturers (OEMs) in the heavy-duty machine industry is that for HDLR manipulators, a +/-5 mm positioning accuracy and 0.1° orientation accuracy for the TCP are desired. However, larger errors may also be allowed in the case of a very low-cost sensing solution. The errors induce challenges with respect to TCP accuracy and repeatability for HDLR manipulators. Presently, the errors are compensated for by the human operator. For autonomous operations without a human-in-the-loop, however, the TCP inaccuracies are a problem yet to be truly solved.

The heavy machinery industry is striving toward semi-autonomous and eventually fully autonomous machines capable of performing work tasks with minimal human intervention. This requires new sensor solutions and algorithms to automate work tasks currently performed by human operators. Possible advantages of increased automation levels include, for example, increased productivity, decreased operation costs, increased safety, and reduced waste [6]. Overall, autonomy for heavy-duty mobile machines is very much an emerging technical field. An overview of the safety standards for autonomous machines in [7] reveals that existing standards are directed to OEMs, whereas the worksite context (worksite operators or owners) has not been considered.

The multidisciplinary challenge of developing autonomous, heavy-duty mobile machinery is also discussed in [8]. As the authors state, sensing and perception are major elements in developing such systems. For example, operations requiring task flexibility and precision can be automated with the assistance of visual recognition and decision making, as vision-based guidance systems allow the manipulator (or machine) to vary their motion targets and achieve increased task flexibility. These technologies require advanced computer vision algorithms, which is a broad technical field aiming to process and understand visual, real-world data so that appropriate decisions can be made. Computer-vision-related topics include, for example, pose estimation, object detection and tracking, and visual servoing (or vision-based control). Frameworks related to these topics have been somewhat well established for conventional industrial robots on factory floors. To exploit visual measurements obtained using a camera, for example, sensor calibration is required. For cameras, the intrinsic calibration parameters represent a projective transformation from the 3D camera frame (or coordinate system) into the 2D pixel coordinates. Additionally, the extrinsic calibration parameters represent a rigid transformation from a 3D world frame to the 3D camera frame. Resulting from dynamic, unstructured working environments, increased uncertainties, and errors in the control system, the associated frameworks developed for conventional industrial robots are often not practical or even feasible for HDLR manipulators. For example, camera-robot cooperation requires estimating the extrinsic parameters between the sensor and the robot, which, in structured factory set-ups, is typically achieved using a specific calibration object with a predefined pattern, such as a checkerboard. The calibration procedure commonly requires taking images from different angles and distances while the calibration object is in view of the camera. While the intrinsic calibration can be performed with similar methods, it can be done while the sensor is not attached to the manipulator. Consequently, such delicate schemes for extrinsic calibration are not practical for HDLR manipulators that work in dynamic, unstructured environments. Overall, the aim of this dissertation is to investigate advanced computer vision methods for HDLR manipulators striving toward increased levels of automation.

1.2 Research Problems

This thesis was conducted under the Doctoral School of Industry Innovations (DSII) at Tampere University. Thus, the research problems (RPs) originated from the industry partner company, Sandvik Mining and Construction Oy. The first RP sought to develop a novel sensor system capable of 6 DOF TCP pose estimation. The second RP sought to improve the TCP positioning accuracy of HDLR manipulators using visual feedback.

RP-I: A novel sensor system capable of 6 DOF TCP pose measurement for

HDLR manipulators
How feasible are cameras for observing and tracking the motion of HDLR manipulators?
How can the reliability and robustness of visual sensing be improved in HDLR manipulators?

RP-II: Vision-based control for precise TCP positioning of HDLR manipulators
Can a low-cost camera be used to achieve precise (sub-centimeter) TCP positioning accuracy in HDLR manipulators?
What are the main challenges in realizing precise TCP control for HDLR manipulators using visual sensing?

The publications in this thesis are a direct result of the given RPs: RP-I is addressed in P-I through P-IV, and RP-II is addressed in P-V, while also utilizing the findings in P-I and P-III.

1.3 Scope of Research

The overall scope of the research in this thesis is under the domain of HDLR manipulators in unstructured environments. Such robotic manipulators are attached to mobile machines, are mostly driven with hydraulics, and can have up to 8 actuator DOF. Individual joints of these mechanically complex manipulators are either revolute or linear, with motion existing in both the vertical and horizontal planes. Initially, no specific restrictions were set for the research with respect to the chosen scientific approaches, sensors, sensor placements, or system costs.

In the scope of RP-I, the aim was to conduct research on novel sensing methods for generic 6 DOF TCP pose estimation of HDLR manipulators in unstructured environments. The novelty specification excluded the prevalent fine-mechanical joint sensors, such as encoders and resolvers. Moreover, an inertial sensor network is very challenging to realize because joints move in the horizontal plane. The possible accuracy of radio detection and ranging (RADAR), including ultra-wideband (UWB), was deemed insufficient. Different sensor types and technologies were examined, but it became clear that non-contact visual sensing was the most interesting approach due to the potential accuracy and task flexibility, and with visual perception being essential for autonomous machines. The overarching ambitious goal is to eventually omit the variety of embedded fine-mechanical joint sensors currently in use and to replace them with a generalized sensor system suitable for all machine types.

In the scope of RP-II, the objective was to guide the TCP of an HDLR manipulator accurately to an object of interest (OOI) using visual sensing. Related methods, such as visual servoing and extrinsic calibration between a visual sensor and a robot, have been well developed for conventional rigid-body industrial robots in factories. However, such frameworks are not practical for non-rigid HDLR manipulators in unstructured environments. Moreover, the desire was to utilize existing industry standards, meaning that basic robotic modeling and control systems were applied.

In the scope of experimental validation, testing was to be performed using an existing installation of a HIAB033 HDLR manipulator with an additional 3 DOF wrist and 5-m reach, located in the Innovative Hydraulics and Automation (IHA) heavy laboratory at Tampere University. For visual sensing solutions, commercial off-the-shelf sensors were to be utilized. Thus, part of this research benefited from a commercial OptiTrack motion capture system, borrowed from the Centre for Immersive Visual Technologies (CIVIT) at Tampere University. Additionally, a low-cost ZED/ZED2 stereo camera was utilized in this research. Moreover, real-time capable methods for vision-based control were required. The experimental setup with its main components is illustrated in Figure 1.2: A HIAB033 HDLR manipulator with an additional 3 DOF wrist, a test wall for visual feature extraction, a ZED2 stereo camera for visual odometry/simultaneous localization and mapping (VO/SLAM), and an OptiTrack marker-based tracking system. In general, all of the included publications take advantage of the same setup with some variations.

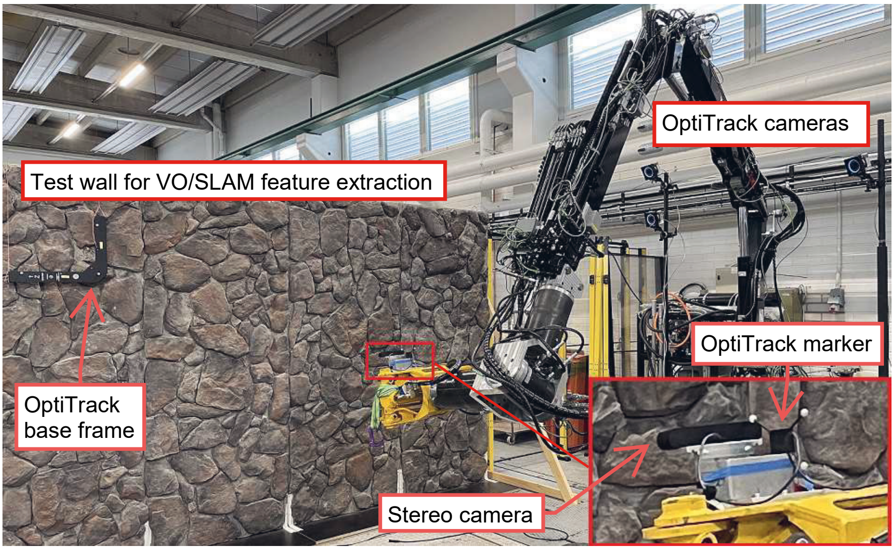


Figure 1.2 The experimental setup used throughout the thesis.

1.4 Thesis Contributions

This section details the scientific contributions of each publication presented in this compendium thesis.

P-I This publication investigates VO/SLAM for tracking the generic 6 DOF TCP pose of an HDLR manipulator in unknown or, for the first time, confined spaces. Offline data analysis was conducted using recorded real-time data, which demonstrated that the VO/SLAM poses, using the eye-in-hand configuration, correspond to the ground-truth encoder-based TCP poses in the tested cases. The initial results provide insight to the potential usefulness of VO/SLAM for tracking the 6 DOF TCP pose of non-rigid HDLR manipulators. Thus, P-I serves as a baseline for the research in P-II through P-V.

P-II This publication extends P-I by proposing a redundant visual TCP pose measurement for HDLR manipulators using marker-based tracking with an eye-to-hand configuration. The rationale was that the eye-in-hand camera (VO/SLAM) provides precision with partial sight of the environment, whereas the eye-to-hand (marker-based tracking) has a more global sight of the environment with less precision due to increased view distance. Thus, in a well-designed system, the cooperative sensing can be complementary. Offline data analysis based on recorded real-time data demonstrated that the proposed visual sensor system can effectively track the TCP pose in the measured cases. Similarly to P-I, the joint encoders were used to obtain the ground-truth TCP poses.

P-III This publication focuses on the sensor system calibration problem in P-I and P-II. The aim was to replace the basic iterative closest point (ICP) method used in P-I and P-II with a more robust solution. An overall pipeline for camera-to-kinematic calibration was proposed using coarse frame alignment followed by fine matching. A comparative study between point set matching methods suggested that a method utilizing full 6 DOF pose data during the registration process provides the most accurate results, which is optimal for robotic applications where 6 DOF pose data is readily available. A use case demonstrated the effectiveness of the calibration procedure.

P-IV This publication examines the sensor fusion problem for redundant measurements discussed in P-II. A real-time capable, model-free data fusion methodology is proposed, for which the weight parameters of the signals are computed online using sliding window (or sample) variances. Offline data analysis using recorded real-time data demonstrated that the proposed system can increase the robustness and fault tolerance of the overall visual sensor system.

P-V This publication investigates a problem, in which the objective was to drive the tool of an HDLR manipulator to a visually detected OOI as accurately as possible. Building on P-I and P-III, the pose error between the tool and an OOI is computed directly in the image frame, while using motion-based local calibration to find the extrinsic sensor-to-robot correspondence. Real-time experiments demonstrated that sub-centimeter positioning accuracy was achieved in the measured cases using the trajectory-matching-based calibration. This level of positioning accuracy is typically not achieved with state-of-the-art HDLR manipulators due to their nonlinear characteristics.

The author considers P-V as the main contribution of this dissertation, where sub-centimeter positioning accuracy was achieved for an HDLR manipulator using the proposed methods.

1.5 The Author's Contribution to the Publications

This section details the author's contribution to the publications presented in this thesis.

P-I The author wrote this paper and designed the methods and experiments for validation. Dr. Mohammad M. Aref acted as an academic co-supervisor and gave scientific insights. Sirpa Launis acted as the industrial supervisor, providing the research problem and industrial insights. Prof. Jouni Mattila, the academic supervisor, reviewed the paper and suggested improvements.

P-II The author wrote this paper, developed the conceptual visual sensor system, implemented the vision-based systems, and designed the experiments for validation. Dr. Pauli Mustalahti implemented the experimental manipulator's basic control system and assisted with the measurements. Sirpa Launis provided the research problem and industrial insights. Prof. Jouni Mattila reviewed the paper and suggested improvements.

P-III The author wrote this paper, developed the overall pipeline for camera-to-kinematic model calibration, implemented the pipeline, and designed the experiments for validation (including the use case and comparative studies). Dr. Pauli Mustalahti implemented the experimental manipulator's basic control system and assisted with the measurements. Sirpa Launis provided the research problem and industrial insights. Prof. Jouni Mattila reviewed the paper and suggested improvements.

P-IV The author wrote this paper, developed the data fusion method using sliding window variance, implemented the fusion algorithm, and designed the experiments for validation. Dr. Pauli Mustalahti implemented the experimental manipulator's basic control system and assisted with the measurements. Sirpa Launis provided the research problem and industrial insights. Prof. Jouni Mattila reviewed the paper and suggested improvements.

P-V The author wrote the paper, developed the motion-based local calibration methods, implemented the algorithms related to visual sensors, and designed the experiments for validation. Dr. Pauli Mustalahti implemented the experimental manipulator's basic control system and assisted with the measurements. Sirpa Launis provided the research problem and industrial insights. Prof. Jouni Mattila reviewed the paper and suggested improvements.

1.6 Outline of the Thesis

The introductory part of this compendium thesis is divided into five chapters. The present chapter provides the basic information regarding the objectives and contributions of the thesis.

Chapter 2 describes the state-of-the-art related to the topics of this dissertation. Chapter 3 presents a summary of each individual paper presented in this compendium thesis. Chapter 4 discusses the overall outcomes of the publications and their relation to the RPs stated in Section 1.2. They are followed by a discussion of limitations and challenges. Chapter 5 concludes the introductory part of this dissertation by summarizing the thesis contributions, remaining challenges, and future work. The structure of this thesis is outlined in Figure 1.3.

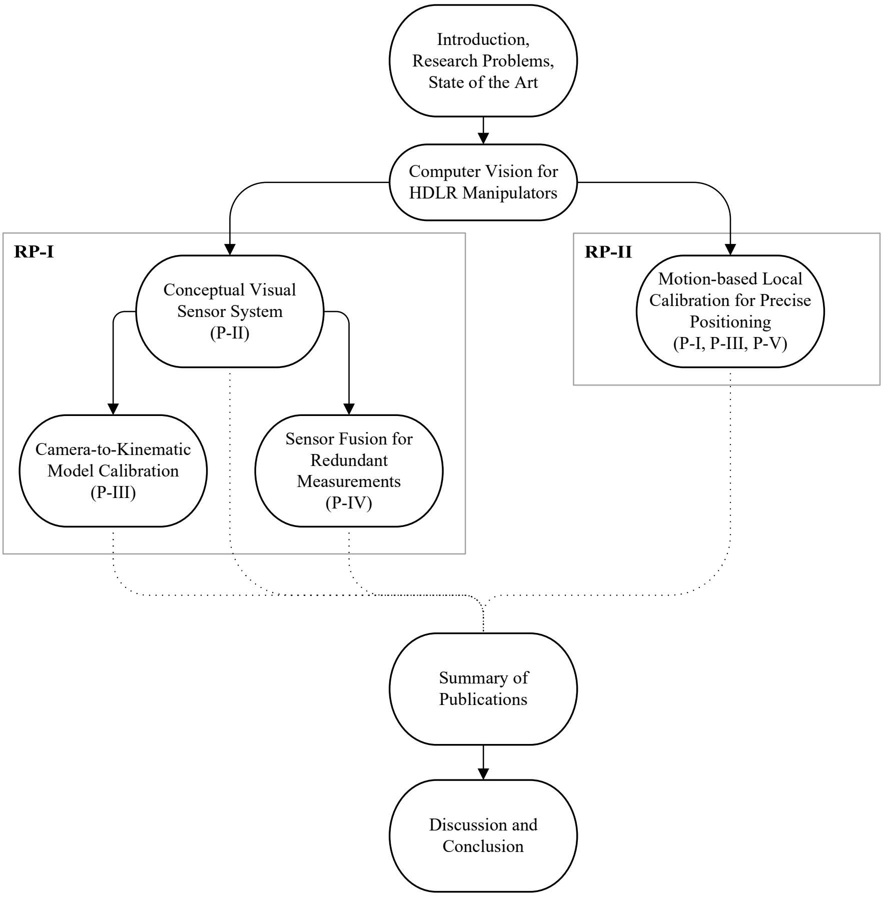


Figure 1.3 Outline of the thesis.

2 STATE OF THE ART

This chapter reviews the state-of-the-art research related to the topics of this dissertation. First, sensor technologies for HDLR manipulators are reviewed. Second, state-of-the-art methods related to visual sensing for HDLR manipulators are reviewed.

Research on visual sensing of HDLR manipulators is rather limited. Therefore, some topics are partially presented from the point of view of autonomous vehicles and conventional industrial robots. While frameworks developed for such systems may not directly be practical for HDLR manipulators, they are still relevant in the context of striving toward increased automation of HDLR manipulators via advanced perception systems.

2.1 Sensor Technologies for HDLR Manipulators

Similarly to conventional industrial robots, a possible sensing method of measuring the joint states of HDLR manipulators is to use fine mechanical sensors embedded into the mechanical structure. Typically, each joint of a robotic manipulator is embedded with a fine mechanical sensor, which also requires additional protective housing (IP67 rating for HDLR manipulators), mechanical coupling, and cabling that are suitable for a given machine type. The sensors are either angular (embedded between two revolute joints) or linear (embedded into a cylinder, for example). The joint measurement is based on mechanical coupling between the sensing element and the manipulator. This can be inconvenient with HDLR manipulators as replacing such sensors is difficult, and the couplings, while having fine mechanical tolerances for flexibility, can be subject to significant forces and moments. Thus, a large variety of spare parts is required in case of breakdowns.

In recent years, several OEMs in the heavy-duty machinery industry have adopted inertial measurement units (IMUs) for estimating the joint states. IMUs measure three-axis accelerations and three-axis angular velocities. Instead of fine mechanical sensors, IMUs are used to compute the joint angles to enable direct TCP control. However, IMUs are subject to drifting, which requires compensation algorithms. For joints moving in the vertical plane, this can be achieved using gravity as a reference. Consequently, IMUs can reliably measure joint angles in the vertical plane of motion [9]. A three-axis magnetometer can also be used to compensate for the drift, but this requires a homogeneous magnetic field, which is difficult to achieve indoors or with close proximity to metallic structures [10]. Therefore, IMUs are mostly suitable for HDLR manipulators in excavators and forestry machines, where most of the joints move in the vertical plane. For example, excavator manipulators typically have 6 actuator DOF, with base rotation and tiltrotator in the horizontal plane. Forestry machines typically have 4 actuator DOF with base rotation. The HDLR manipulators utilized in the mining industry have up to 8 joints per manipulator, several of which move in the horizontal plane. Thus, an IMU network with the current knowledge is not a feasible solution for complex mining manipulators. These state-of-the-art HDLR manipulators utilize joint sensors with mechanical couplings to compute the TCP variables (typically 5 DOF in mining applications) for control purposes, which is a requirement when designing drilling patterns, for example.

Some recent research has focused on methods related to sensing joint states or the TCP of HDLR manipulators. A 2D laser scanner was utilized to estimate the posture of a flexible forestry crane in [11]. Two scan targets were used, and the error at the tip was reportedly less than 4.3 cm. In [12], a laser-scanner-based approach is investigated to estimate the dipper pose of a mining shovel. The mean dipper positioning error is reported as 6.7 cm. In [13], an ultrasound time-of-flight ranging method is used to measure the length of a telescopic boom, but the accuracy is not sufficient for practical applications. Realizing the disadvantages of basing the control of HDLR manipulators on joint sensors, little research has observed the TCP directly. In [14], a UWB real-time location system is examined to estimate the pose of a crane. However, the method was targeted for monitoring purposes only, as the accuracy of a UWB-based system is approximately 10-30 cm. Thus, UWB is not feasible for TCP pose estimation of HDLR manipulators because the target accuracy is in the sub-centimeter range. In [15], the TCP of a large-scale manipulator is controlled using absolute position feedback from a total station network. Sub-centimeter TCP positioning accuracy is reported, but the system requires large space to operate and a considerable investment in the sensors.

Vision-guided robotic systems are essential to enable autonomous operations for intelligent machines, but analyzing the data requires advanced image processing. A current trend is combining artificial intelligence (AI) with visual sensing, as reviewed by [16]. The possibilities of AI-enhanced machine vision are also highlighted by [17]. It is only logical that visual sensing needs to be accompanied by intelligent computing in order to automate complex work tasks in challenging environments. The opportunities for sensing systems in mining applications are examined by [18]. These studies tend to focus on the machine level, while little attention is given to the robotic arms used for work tasks.

Some research has focused on automating individual work tasks. For example, an early study [19] investigates vision-based control for mining applications. Other studies include [20], which examines robotic explosive charging in mining. In [21], robotic peg-in-hole assembly strategies are compared. The peg-in-hole task is closely related to some work tasks in mining, as one of the main operations is drilling holes. This is usually followed by inserting supportive rods or explosive charges into them.

Laser scanners are a common sensing method used in mines. They can be used for various tasks: in [22], a machine learning approach is used to detect rock bolts. In [23], applications utilizing point cloud data from laser scanners are reviewed for underground mining. It is claimed that presently, the huge amount of data captured with laser scanners is mostly processed off-site, which narrows the possible applications. UWB-based sensor networks for localization in mines are also investigated in [24], [25], but the potential accuracy is seen as feasible only for navigation on the machine level.

2.2 Visual Pose Estimation in Robotic Applications

This section presents state-of-the-art visual pose estimation from two perspectives. The first one is egomotion, which aims to track the sensor's pose with respect to the environment. The second one aims to track an OOI in the image frame and estimate its pose.

2.2.1 Visual Odometry and Simultaneous Localization and Mapping

Visual odometry (VO) is used to estimate the egomotion of an object using a visual sensor attached to it [26]. Specifically, the sensor's pose trajectory is maintained based on the perceived motion of the sensor. For example, a vehicle's pose is estimated based on the motion between two image frames obtained using a camera. The motion itself is estimated using feature matching/tracking or optical flow techniques. As VO estimates the egomotion incrementally, it is subject to drifting because eventually the errors between computed frames start to accumulate. A comprehensive survey on odometry, including VO and its variants, for autonomous navigation systems is presented by [27].

In simultaneous localization and mapping (SLAM), a map is simultaneously built and maintained while performing localization. Visual SLAM algorithms can be categorized into direct and indirect methods. Direct methods base the localization on using the entire image and photogrammetry, whereas indirect methods are based on features. Roughly, SLAM comprises a VO front-end and a back-end for the mapping. The typical structure of a SLAM system is illustrated in Figure 2.1. A map can be used to refine the localization result to increase the accuracy when revisiting an area or to re-localize with respect to the surrounding environment. In addition to localization and mapping, state-of-the-art SLAM algorithms perform loop detection and closure, which aims to detect if a specific area has been visited before. This is useful for drift compensation in the VO front-end. For robotic control purposes, however, possible discontinuities in the pose variables resulting from optimization procedures are problematic.

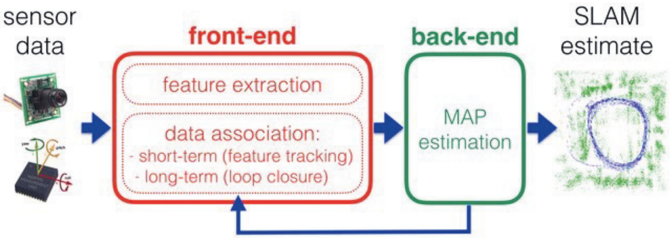


Figure 2.1 Front-end and back-end in a typical SLAM system. The back-end can provide feedback to the front-end for loop closure detection and verification. [28] ©2016 IEEE

Both VO and SLAM have attracted significant attention in recent years, as they are key technologies for autonomous machines and mobile robotics [29]. In this thesis, the general term VO/SLAM is used to refer to algorithms capable of visual localization. Currently, the most common applications for VO/SLAM are related to navigation in unstructured or semi-structured environments, such as self-driving cars on roads, heavy-duty machines at work sites, small robots (e.g., delivery) in urban areas, and drones in the air. One of the current research directions in VO/SLAM is optimizing the algorithm for challenging scenarios, which often result in drift errors, or even complete failure due to losing the tracked features, in the visual pose estimation [30]. For example, image enhancement for improving VO/SLAM performance in a challenging environment is investigated in [31].

Many algorithms have been proposed for (visual) SLAM, as evidenced by [28], [32]–[34], and they can be categorized in many ways. A simple way is to differentiate based on the sensing method used, such as monocular [35], [36], stereo [37], RGB-D [38], inertial sensor enhanced [39], and laser scanner enhanced [40]. Sensor fusion is employed to combine the data if there are many sensors, which then requires extrinsic calibration between the sensors. Another differentiation method is to consider the target application, as some algorithms are developed for drones, which have different properties from ground-level machines. For example, drones usually employ a monocular camera that weighs the least and requires the least space. Several state-of-the-art SLAM algorithms are open-sourced for research purposes. A recent survey performed an extensive comparison of the performance of existing open-source visual SLAM algorithms [41]. The authors concluded that on average, ORB-SLAM2 [42] and ORB-SLAM3 [39] provide the most reliable performance for trajectory estimation. As revealed by their names, these algorithms use oriented FAST and rotated BRIEF (ORB) [43] features for visual localization, which have become popular due to their computational efficiency as binary features. Real-time capability in VO/SLAM is essential, although features such as speeded-up robust features (SURF) and scale invariant feature transform (SIFT) could provide a more accurate result than the binary ORB features [44]. Most existing SLAM algorithms can only guarantee almost global convergence. In [45], a gradient-based hybrid observer for SLAM is proposed, ensuring global asymptotic convergence of estimation errors to zero.

2.2.2 Marker-based Tracking

In VO/SLAM, the main objective is to estimate the motion of the sensor itself. In contrast, marker-based tracking focuses on tracking the motion of an OOI in the image. Visual fiducial markers are commonly used as OOI in motion capture, and they are either passive, reflectively passive, or active (e.g., LED-based). Several high-end optical motion-capture systems exist in the market, including OptiTrack (which is used in P-II and P-IV) [46], Vicon [47], Qualisys [48], and PhaseSpace [49]. These systems mostly employ multiple cameras and markers (passive or active) for full-body pose estimation of OOI, such as human actors, in room scale and at high refresh rates. The visual pose estimation is based on triangulation by utilizing overlapping 2D data obtained from multiple cameras. Although these advanced systems are highly effective at tracking visual fiducial markers even at long distances, they require considerable investment, which is not optimal for industrial production machines.

Researchers have developed several visual fiducial marker libraries dedicated to robotics, augmented reality applications and camera calibration. These methods include ArUco markers (which are used in P-III and P-V) [50] and AprilTags [51], which are commonly used (passive) fiducial markers in robotics research. Two examples of visual fiducial markers are illustrated in Figure 2.2.

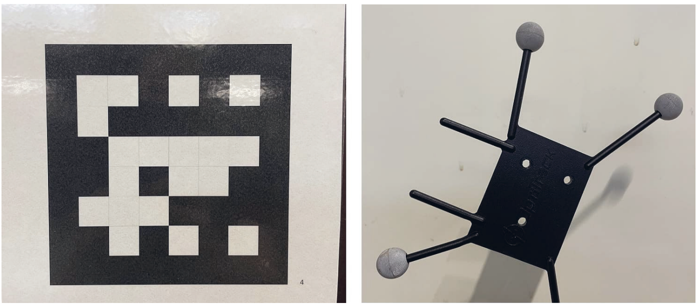


Figure 2.2 i) (left) An ArUco marker. ii) (right) An infrared-reflective marker (OptiTrack).

Vision-based fiducial marker algorithms roughly comprise object detection and pose estimation. In the case of ArUco markers, for example, the detection is based on the square shape and the binary codification matrix in the middle, whereas the pose estimation is based on the known side length and the four corners of the planar marker. The pose estimation is formulated as a perspective-n-point (PnP) [52] problem, in which the pose is computed by solving for the rotation and translation that minimize the reprojection error from the known 3D-2D point correspondences. In [53], the bucket tooth position of an articulated excavator was estimated based on visual fiducial markers. The resulting absolute positioning error was claimed to be under 2.5 cm, while many issues were identified for practical implementations. Some researchers omit visual fiducial markers. Instead, the pose estimation is based on detected parts of a manipulator [54]. Overall, an eye-to-hand pose estimation scheme alone for HDLR manipulators is unlikely to be sufficiently accurate or robust due to occlusions and long view distances. However, an eye-to-hand system can provide a good view of the manipulator and its surroundings, while visual fiducial markers are repeatable targets that are not susceptible to drifting like VO-based pose estimation.

2.3 Sensor Fusion for Autonomous Machines

Sensor fusion involves fusing data obtained from different sensors, such as RADAR, light detection and ranging (LIDAR), various cameras, and IMUs. In general, fusion algorithms can be categorized into competitive, complementary, and cooperative algorithms. The classification is dictated by the manner of fusing multiple sensor signals into a fused signal. As discussed in [55], sensor fusion has a crucial role in autonomous systems, and it is one of the fastest developing areas. The potential benefits of data fusion are related to enhancing the data authenticity and availability [56]. The authenticity includes, for example, improved confidence and reliability and reduced ambiguity. The availability infers extending the spatial and temporal coverages. Most of the research in sensor fusion has focused on autonomous vehicles, which is basically synonymous with self-driving cars. The heavy-duty machinery industry, in comparison, has such low production volume per machine type that the research lags behind. The most significant difference is that heavy-duty machinery operates in harsh, unstructured environments, which issues further challenges. On the other hand, mobile machines typically work in closed-off areas instead of public spaces (roads). This provides some potential advantages with respect to legislation.

Deep-learning-based sensor fusion for autonomous vehicle perception and localization is surveyed in [57]. It was found that deep learning can significantly improve a vehicle's perception capabilities. Challenges are also highlighted, such as the requirement of vast amounts of training data. Further areas requiring improvement are related to harsh weather conditions, reliability, and repeatability, for example. Sensor fusion technologies for autonomous vehicles are also examined in [58]. Before data fusion, sensors have to be calibrated intrinsically, extrinsically, and temporally. It is emphasized that current research focuses on offline calibration methods, for which specific calibration targets are used. This type of system is very inflexible, and the authors argue that further research on online and offline calibration methods in sensor-to-sensor fusion is required. Some research on sensor fusion of small-scale robotic arms also exists, for example, in [59], [60]. These methods utilize individual joint sensors and dynamic models.

A Kalman filter is typically employed in the data fusion process by optimally estimating the states to combine sensor data. Many studies have been undertaken related to Kalman filtering, such as [61]–[64]. As implied by state estimation, Kalman filtering requires a model of the system states. For autonomous ground vehicles, relatively simple motion models can be derived [65]. For robotic manipulators, motion in complete 3D Cartesian space has to be considered. An early study [66] argues that the key to intelligent fusion of disparate sensor data is having an effective model of the system. Few research exists on direct fusion of continuous sensor signals, but it can be based on signal statistics when a complete dynamic model of the system is difficult to formulate. For example, data fusion of continuous signals is based on confidence-weighted averaging in [67], where the weight parameters for individual signals are specified by pre-determined confidence functions utilizing signal variances.

2.4 Vision-based Control in Robotics

This section briefly examines the state-of-the-art of vision-based control in robotics. As research considering HDLR manipulators on this topic is very limited, it is partly discussed by referencing frameworks developed for conventional industrial robots. Many of such frameworks are not practical for HDLR manipulators. The two main topics of interest are extrinsic calibration and visual servoing, both of which also relate to HDLR manipulators. However, especially the calibration problem imposes challenges for HDLR manipulators, as the most common methods involving a known calibration object that acts as a world frame are practical mostly for stationary industrial robots on factory floors.

Robotic manipulators with visual sensors can be divided into two categories: the so-called eye-in-hand and eye-to-hand systems. In eye-in-hand systems, the sensor is situated near the end-effector of the robot, whereas in the latter the sensor is situated in the environment. The two configurations are illustrated in Figure 2.3. The most significant difference is that a camera in the eye-in-hand configuration is subject to the same motion as the robot's TCP. The motion is then estimated using visual feature tracking or optical flow. In contrast, an eye-to-hand system observes the robot's motion within its workspace, thus requiring trackable OOI, such as visual fiducial markers, attached to the robot. This imposes additional challenges, such as occlusion and field-of-view issues, to track the robot's pose.

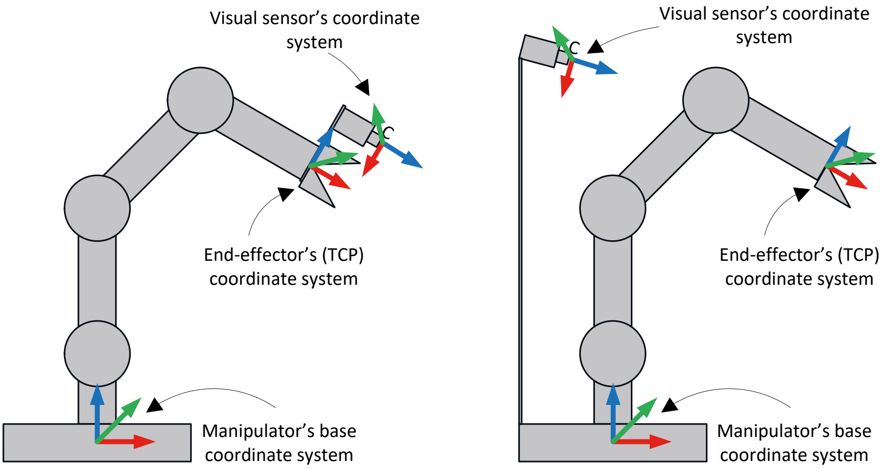

Visual sensor's coordinate system
C
End-effector's (TCP) coordinate system
Manipulator's base coordinate system
Visual sensor's coordinate system
C
End-effector's (TCP) coordinate system
Manipulator's base coordinate system

Figure 2.3 i) (left) Eye-in-hand configuration. ii) (right) Eye-to-hand configuration.

2.4.1 Extrinsic Calibration between a Visual Sensor and a Robot

In the context of robotics, extrinsic calibration describes the rigid relationship between a visual sensor's coordinate system and the robot's coordinate system. Depending on the hand-eye configuration (eye-in-hand or eye-to-hand), the calibration procedure involves computing one or more transformation matrices. After extrinsic calibration, measurements in the sensor's coordinate system can be expressed so that the robot's control system understands them. The most common solutions for hand-eye calibration utilize a calibration object (i.e., a world frame) with easily detectable geometric shapes, such as a checkerboard, a circlegrid, or visual fiducial markers [68]–[71]. Moreover, it is required that the manipulator's forward kinematic model be available.

Many algorithms for both hand-eye configurations have been proposed to solve the extrinsic calibration problem, some of them solving the rotational and positional components separately and some doing so simultaneously. In [72], a cooperative eye-in-hand/eye-to-hand visual servoing scheme is examined. The rationale is that an eye-in-hand camera provides precision with partial sight of the environment, whereas an eye-to-hand system has a wider view of the environment but less potential accuracy due to the increased view distance. Thus, in a well-designed system, the cooperative sensing can be complementary. More recently, the inconvenience of being restricted to specific calibration objects has been noted as some research exists on targetless calibration methods. For example, in [73], motion-based calibration for multi-modal sensor extrinsics and timing offset estimation is presented, without requiring calibration objects. In [74], the hand-eye transform is estimated on a surgery robot using a neural network to omit the requirement of a specific calibration object. A recent overview on hand-eye calibration methods [71] also notes that the related technologies are developing toward high precision and intelligence, though much work is required to identify the robot and camera parameters.

2.4.2 Visual Servoing

Visual servoing in robotics involves controlling the motion of a robotic system based on visual feedback [75]. Methods of visual servoing are classified into position-based (PBVS) [76], image-based (IBVS) [77], and hybrid systems [78]. In a PBVS system, the control error is defined in Cartesian coordinates, and the control algorithm utilizes the robot's kinematic model along with camera calibration parameters. In an IBVS system, the control law is defined in the image plane using image features directly. The relationship between the image plane and the robot is established using an image Jacobian matrix, which describes a nonlinear mapping between the image feature errors and the pose of the robot [79]. PBVS systems are more sensitive to calibration errors, as well as more complex to implement. IBVS systems are more robust against calibration errors, but a singularity in the image Jacobian can render the controller unstable. Hybrid systems attempt to utilize the advantages of both the PBVS and IBVS.

In general, PBVS systems are more suited for HDLR manipulators as the control structure is based on the Cartesian coordinates, enabling vision-based control in the 3D Cartesian space. Thus, the pose of a target OOI is required to determine the tool's pose with respect to the OOI. However, the task of visual pose estimation of application-specific OOI is challenging on its own. IBVS systems are more suited to structured environments, where an industrial robot has only a specific OOI in the camera view and servoing in 2D space is sufficient. Most of the research related to visual servoing is directed toward conventional industrial robots, with few papers considering HDLR manipulators. One such study is [80], in which a camera was attached near the tip of an HDLR manipulator and used for PBVS. However, no explicit procedure for extrinsic calibration between the sensor and the robot is described. Instead, a known location is assumed by approximating the sensor's displacement along two axes with reference to the forward kinematic model. In practice, the visual sensor's coordinate system is very difficult to estimate correctly without proper calibration. In [81], an eye-to-hand configuration was used for visual guidance of a heavy-duty rock-breaking manipulator. Specific markers for calibration purposes were distributed into the workspace and a considerable number of measurements were conducted to estimate the extrinsic camera-to-robot calibration parameters.

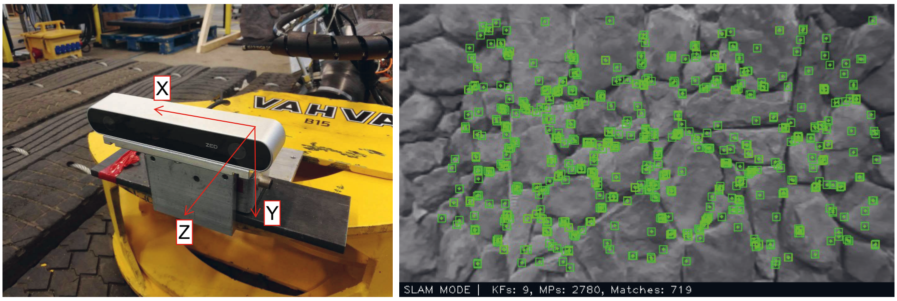
3 SUMMARY OF PUBLICATIONS

This chapter provides a summary for each publication presented in this dissertation.

3.1 Summary of P-I: Application of Simultaneous Localization and Mapping for Large-Scale Manipulators in Unknown Environments

Development drilling is one of the basic operations in underground mining. It involves mining tunnel networks, for which mobile machines with on-board HDLR manipulators are utilized. A mining manipulator holds a heavy rock drill, and the work area is a confined space within a tunnel, surrounded by walls of rock. This paper investigates the application of VO/SLAM for tracking the generic 6 DOF TCP pose of an HDLR manipulator in an unstructured environment. The focus is on a preliminary examination for the feasibility and potential accuracy of VO/SLAM-based TCP pose estimation in this type of application. Presently, OEMs use a forward kinematic model with individual joint sensors to formulate the TCP pose, which is subject to significant kinematic and non-kinematic errors accumulated along the serial chain kinematic structure. The motivations behind developing alternative sensing methods are also discussed, and a brief literature review on sensor technologies is presented.

Based on the literature review and the availability of open-source SLAM algorithms, the feature-based ORB-SLAM2 Stereo algorithm was chosen for the experiments. An experimental setup comprising an HDLR manipulator and a textured test wall mimicking rocks for feature extraction were used to collect data. A low-cost, off-the-shelf stereo camera was attached to the manipulator in the eye-in-hand configuration. Figure 3.1 illustrates the camera placement and a view of detected ORB features on the test wall.



SLAM MODE | KFs: 9, MPs: 2780, Matches: 719

Figure 3.1 i) (left) ZED stereo camera in the eye-in-hand configuration. ii) (right) A view of detected ORB features during SLAM. [Source: P-I]

The manipulator was moved around its workspace while the camera was pointing toward the test wall. Simultaneously, the pose trajectory data of the TCP were obtained using the visual stereo SLAM algorithm and a forward kinematic model with joint encoders. The encoder-based TCP pose variables were assigned as the ground-truths, and an iterative closest point (ICP) algorithm was employed for extrinsic calibration between the SLAM-based and the encoder-based pose variables. Offline data analysis was conducted, and the resulting comparative study demonstrates promising results with reasonable accuracy for the considered application. Specifically, the extrinsically calibrated VO/SLAM-based pose variables represent the dynamic behavior of the TCP well when compared with the ground-truth poses. The method's limitations are also highlighted.

3.2 Summary of P-II: Redundancy-Based Visual Tool Center Point Pose Estimation for Long-Reach Manipulators

This paper proposes a new visual sensor system concept for HDLR manipulators in unstructured environments, in which an eye-in-hand/eye-to-hand cooperative scheme is utilized. Specifically, the VO/SLAM-based TCP pose estimation approach presented in P-I was extended to a more complete, conceptual visual sensor system by incorporating a marker-based tracking module. For a robotic manipulator, a visual fiducial marker provides a sufficient and repeatable target. The overall rationale is that a camera near the TCP (eye-in-hand) has a narrow view but is able to provide more accurate pose estimates. On the other hand, a camera placed in the environment (eye-to-hand), such as on the roof of a machine, has a wide view of the surroundings but less accurate pose estimation capability due to the increased distance. The aim of the proposed conceptual visual sensor system was to increase the robustness and fault tolerance of the vision-based TCP pose estimation scheme via sensor redundancy and data fusion.

The experimental setup comprised the same system that was used in P-I, including the state-of-the-art ORB-SLAM2 Stereo algorithm. For marker tracking, a high-end commercial OptiTrack motion capture system was utilized. The experimental setup is illustrated in Figure 3.2. The ground-truth TCP pose variables were obtained using the forward kinematic model with joint encoders. The vision-based TCP pose estimates were fused in a competitive manner using confidence weighted averaging, which is a model-free method utilizing specified confidence functions to compute the weight parameters. The proposed visual sensor system was studied in offline data analysis by using offline-capable methodologies for the extrinsic calibration and the data fusion, with recorded real-world data. Equal availability (same frequency) was assumed for the pose estimates provided by each measurement method.

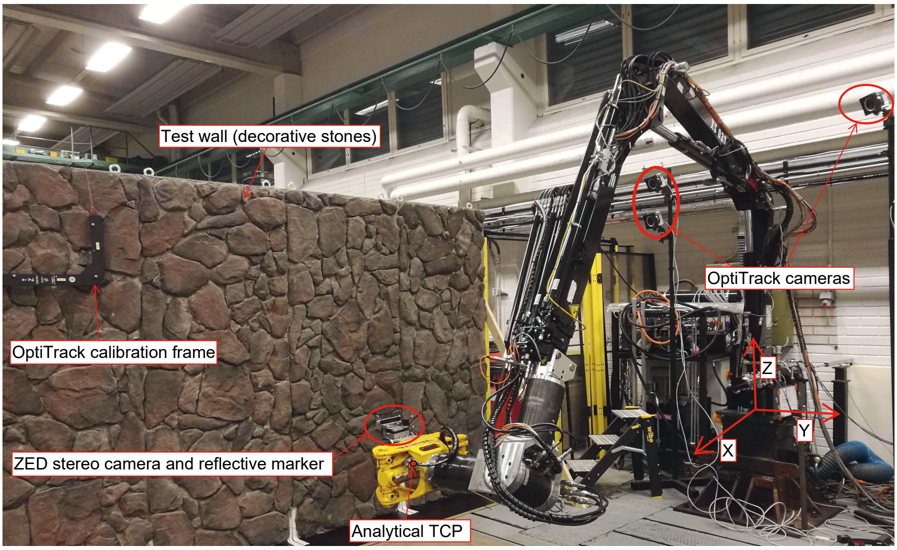

Test wall (decorative stones)
OptiTrack cameras
OptiTrack calibration frame
Z
Y
X
ZED stereo camera and reflective marker
Analytical TCP

Figure 3.2 The experimental setup for eye-in-hand/eye-to-hand cooperative visual TCP pose estimation. [Source: P-II]

The experiments demonstrated that in the tested cases, the VO/SLAM module with a narrow view was able to provide orientation signals with better quality. The marker tracking module with a wider view suffered from minor deterioration in the orientation measurements due to the increased distance from the observed target. Overall, the results suggest that the state-of-the-art visual pose estimation methods, directly observing the motion of the TCP, provide more accurate TCP poses in the absolute coordinates compared to the rigid-body-based forward kinematic model and joint encoders.

3.3 Summary of P-III: Probabilistic Camera-to-Kinematic Model Calibration for Long-Reach Robotic Manipulators in Unknown Environments

This paper follows P-I and P-II by focusing on developing a robust methodology for on-site extrinsic camera-to-kinematic model calibration. The aim was to examine point set matching methods and to replace the basic ICP algorithm utilized in P-I and P-II with a state-of-the-art alternative. For this purpose, a probabilistic point set matching method capable of utilizing the full 6 DOF pose data was applied. Commonly, point set matching methods utilize only 3 DOF position data. Thus, the rationale was that for robotic applications with full 6 DOF pose data available, the matching result should benefit from the increased amount of data used in the point set registration process.

The probabilistic 6 DOF point set matching method utilized a Gaussian mixture model (GMM) to model positional uncertainties and a von Mises-Fisher mixture model (FMM) to model orientational uncertainties. This hybrid mixture model-based method was compared with two other point set matching methods utilizing only the 3 DOF position data. The first method was a classic least-squares-based method, while the second was the coherent point drift (CPD) algorithm, a probabilistic method closely related to the applied 6 DOF hybrid mixture model method. Specifically, they both share the GMM part, while the 6 DOF method is extended with the FMM. The overall pipeline for camera-to-kinematic model calibration comprised coarse frame alignment and fine matching. Point set matching methods are mostly guaranteed to converge to a locally optimal solution. Therefore, it was required that the point sets were roughly aligned based on the known initial TCP pose, before running the point set registration for fine matching.

The methods in P-I and P-II were extended to a real-time setting instead of offline data analysis, and arbitrary manipulator motions were used during the calibration sequence instead of a predefined path. The camera was still pointed toward the test wall for visual feature extraction. Pose trajectories were obtained using the ORB-SLAM2 Stereo algorithm and the joint encoders. A specific orientation magnitude correction was used to obtain comparable VO/SLAM-based orientation signals with the encoder-based TCP pose data. The results demonstrate that the proposed pipeline with 6 DOF data provide the smallest calibration errors, making it effective for robotic applications. Finally, a simple use case of utilizing the camera-to-kinematic calibration was executed by driving the image center to detected ArUco markers. An example result is illustrated in Figure 3.3. The mean Euclidean distance errors were in the sub-centimeter range, which can be considered excellent in this application. However, the depth parameter was not included in the use case.

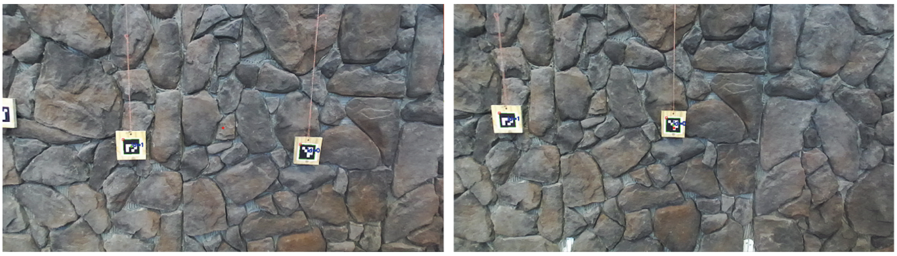


Figure 3.3 i) (left) The initial pose. ii) (right) The pose after driving the image center to a detected marker. [Source: P-III]

3.4 Summary of P-IV: Model-Free Sensor Fusion for Redundant Measurements Using Sliding Window Variance

The aim of this paper was to develop a real-time capable data fusion method for the redundancy-based visual sensor system described in P-II. The conceptual design comprising VO/SLAM and marker-based tracking modules is illustrated in Figure 3.4. The confidence-weighted, averaging-based method used in P-II is suitable for offline data analysis. For real-time data fusion of continuous sensor readings, the use of pre-specified confidence functions is not practical. Before fusion, the pose variables also have to be extrinsically calibrated to a common frame, for which the pipeline presented in P-III was used.

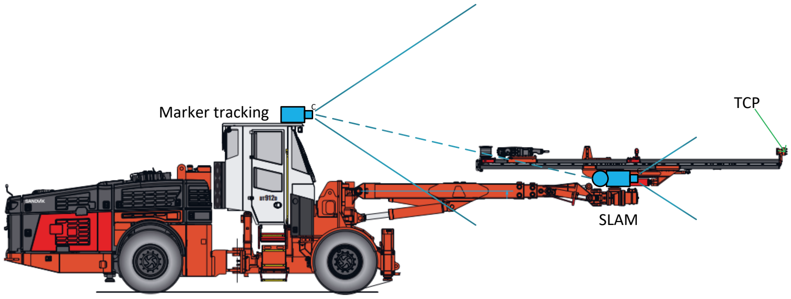


Figure 3.4 The overall conceptual design for eye-in-hand/eye-to-hand cooperation: The Sandvik DT912D single boom tunneling jumbo. [Source: P-IV]

The proposed model-free data fusion method was based on weighted averaging, but the weights for each signal were computed using sliding window variance (or sample variance) with N latest observations. The rationale is that the signal with less variance is assumed to be of better quality, thus it is given a larger weight. The window length N was updated after each individual sliding window so that the fusion algorithm can react to dynamic changes in the signals. Updating the window length N was conducted by using the largest absolute mean difference between redundant signals. Having a specified window length implies that the system reacts to signal changes with a slight delay. In the experiments, the longest window used was 4,000 samples (or 4 s). A simple method for transition smoothing was also presented, which aimed to mitigate the effects of outliers.

Real-time experiments were conducted using the HIAB033 setup. The TCP pose was estimated using the ORB-SLAM2 Stereo algorithm and the OptiTrack motion capture setup. The objective was to maintain a reliable TCP pose measurement of the HDLR manipulator using the conceptual visual sensor system presented in P-II. The 6 DOF pose trajectory data from the two independent, redundant visual sensors were fused in an optimal manner in the sense that the variances of the fused signals were minimized with respect to the input variances, computed over the current sliding windows. The results demonstrated that the proposed data fusion methodology can increase the overall robustness and fault tolerance of the visual sensor system. The challenges of the proposed methods include having to detect and discard any grossly faulty measurements before the fusion occurs, which implies relying on advanced sensor self-diagnostics.

3.5 Summary of P-V: Vision-Aided Precise Positioning for Long-Reach Robotic Manipulators Using Local Calibration

This paper investigates motion-based sensor-to-robot calibration methods for vision-based guidance of HDLR manipulators. The objective was to achieve precise (sub-centimeter) absolute TCP positioning accuracy using a low-cost visual sensor. Compared to conventional industrial robots, the error tolerances in the TCP position are much larger. As previously highlighted, the existing extrinsic calibration frameworks using specific calibration objects (e.g., checkerboards) are not practical for HDLR manipulators in dynamic, unstructured environments. Thus, this paper proposes motion-based calibration in a local plane, while the pose error between the tool (TCP) and an OOI is computed directly from an image. Consequently, the view of the tool and the OOI, along with the capability to estimate their poses, were assumed. The presented methodology comprises orientation adjustment for aligning the tool and the OOI, as well as range adjustment, for which two methods are proposed and compared. The first method is a line-equation-based method using a circular calibration path. The second is a trajectory-matching-based method using an asymmetric calibration path.

Real-time experiments were conducted using the HIAB033 setup, and a low-cost ZED2 stereo camera was used for visual sensing in the eye-in-hand configuration. ArUco markers were used to represent a tool and an OOI in the experiments. The experimental setup is illustrated in Figure 3.5. Using motion-based local calibration, the objective was to position the tool (the marker at the tip) in the middle of three markers attached to the board. Open-loop visual control was employed by first looking and then moving. The motivation was to avoid closed-loop visual control due to robustness and occlusion issues. The line-equation-based range adjustment suffered from accumulated errors and had considerable variation in the final tool positioning result. The trajectory-matching-based method was able to reliably achieve sub-centimeter positioning accuracy, although two visual control inputs were required in each measured case. The precise positioning was enabled by computing the pose error between a detected tool and an OOI directly in the image frame, while using highly accurate local calibration resulting from the trajectory matching. Such accuracy is typically not achieved with HDLR manipulators due to the kinematic and non-kinematic errors present in the system. Moreover, the motion-based local calibration does not require external objects placed in the environment for calibration purposes.

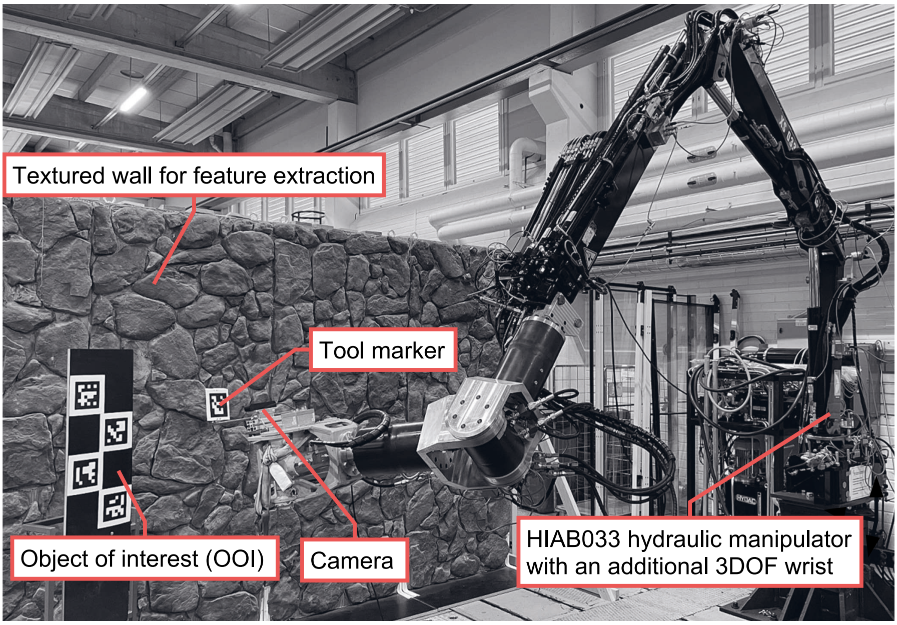

Textured wall for feature extraction
Tool marker
Object of interest (OOI)
Camera
HIAB033 hydraulic manipulator with an additional 3DOF wrist

Figure 3.5 The experimental setup for motion-based local calibration and visual guidance of a HDLR manipulator. [Source: P-V]

The challenges and limitations are also discussed. For a practical application, the object detection and pose estimation have to be realized with application-specific parameters, which is a challenge. Further issues include lighting and sufficient textures for VO/SLAM.

4 DISCUSSION

This chapter discusses the publications and their outcomes in relation to the RPs. The significance and validity of the overall results are discussed, along with limitations and persisting challenges.

4.1 A Novel Sensor System Capable of 6 DOF TCP Pose Measurement for HDLR Manipulators (RP-I)

How feasible are cameras for observing and tracking the motion of HDLR manipulators? How can the reliability and robustness of visual sensing be improved in HDLR manipulators?

RP-I focuses on examining advanced visual sensing methods for estimating the generic 6 DOF TCP pose of HDLR manipulators in unstructured environments. Notably, HDLR manipulators are subject to structural deformations, which suggests that observing the TCP directly can provide a more accurate result in comparison to using individual joint sensors with conventional rigid-body-based kinematic modeling. P-I investigates the performance and feasibility of a state-of-the-art, feature-based SLAM system, although the VO front-end is mainly utilized. Further experiments are required to determine if VO alone is sufficient or if the back-end of SLAM is required for extended operation. While drift correction computations would likely be a requirement for long-term operation, correcting the pose in a discontinuous manner is not allowed for real-time TCP pose-tracking applications. The initial results of P-I indicate that under sufficient conditions regarding lighting and available textured surfaces, the method is capable of adequately tracking the 6 DOF TCP pose. P-II extends the visual TCP pose estimation scheme of P-I into a more complete sensor system by incorporating an eye-to-hand configuration. A marker-based tracking system is utilized, as visual fiducial markers provide repeatable targets suitable for robotic manipulators. The rationale is that the eye-in-hand camera has a narrow view of the surroundings but is able to provide more accurate TCP pose estimates. The eye-to-hand system has a wider view of the surroundings but provides less accurate TCP pose estimates due to the increased view distance.

Moreover, as discussed later, the eye-in-hand camera can be used for vision-based control in auxiliary work tasks. The eye-to-hand camera system can also be used for auxiliary tasks. Possible applications include detecting humans or obstacles in the workspace, for example. Overall, P-I and P-II seek to address the first research question of RP-I regarding the feasibility of visual TCP pose estimation for HDLR manipulators in dynamic, unstructured environments. As discovered in P-II, both of the visual TCP pose estimation methods appear to represent the dynamic motion of the TCP more accurately than the TCP formulated using the rigid-body-based forward kinematic model and joint encoders. This was expected due to the non-rigid nature of HDLR manipulators, but it is problematic as the control system itself is based on the less accurate forward kinematic TCP. Thus, the outcome is that based on the results of P-I and P-II, visual TCP pose estimation based on both VO/SLAM systems and marker-based tracking can provide highly accurate measurements in the right conditions. Further challenges related to robotic control methods and long-term performance of VO/SLAM are discussed later.

The second research question of RP-I is partially addressed in P-II via a proposal of a visual TCP pose estimation scheme using redundant measurements. As a non-contact sensing method, visual pose estimation includes many challenges with respect to robustness and reliability. These challenges arise from occlusions, outliers, and calibration parameters, for example. The second research question of RP-I seeks to find measures to improve these aspects in the visual sensor system proposed for HDLR manipulators in P-II. As a result, P-III examines a robust pipeline for the camera-to-kinematic model calibration. A probabilistic point set matching method capable of utilizing the full 6 DOF pose data, which is readily available in robotic systems, is utilized for this purpose. The comparative results with added Gaussian noise, along with the studied use case, suggest that this method provides the least matching errors; however it still suffers from the same issues as all point set registration methods. Specifically, coarse frame alignment is required to minimize the possibility of the point set matching algorithm converging to a local minimum. A future development should be the formulation of a global solution for the coarse frame alignment.

To address the data fusion problem in P-II, P-IV proposes a real-time capable method for fusing redundant sensor data using sliding window variances. The aim of sensor fusion, along with redundant visual TCP pose estimation, is to increase the system's overall robustness, reliability, and fault tolerance. Thus, the fusion method in P-IV is based on signal statistics, with the assumption that out of two redundant signals, the one with less (sample) variance is of higher quality. Therefore, it is given a larger weight parameter when forming the fused signal. As the method is solely based on signal statistics, advanced sensor self-diagnostics would likely be required to detect and discard grossly faulty measurements before the fusion occurs. P-IV also presents a simple method for transition smoothing by predicting the next fused value in a naïve manner by using linear interpolation, but a direction for future research would be to increase the fusion algorithm's capability to detect and handle outliers. A downside of the data fusion method is that the signal variances are computed over a sliding window of N samples, which infers that if a signal's quality suddenly changes, the fusion algorithm reacts with a delay. However, the experiments in P-IV suggest that the window length N can be maintained as relatively short. Overall, P-II through P-IV address the second research question of RP-I in the scope of the visual sensor system examined in P-II.

The main result of P-I and P-II is a visual sensor system targeted for generic 6 DOF TCP tracking of HDLR manipulators, whereas P-III and P-IV propose real-time capable methods for increasing the robustness and reliability of the visual sensor system. Although based on experimental real-time measurements, the results presented in this dissertation were validated in a laboratory setting. Thus, implementing the proposed methods into practical systems is a bridge yet to be crossed. Practical implementations of vision-based sensing for HDLR manipulators face many challenges, of which the placement of the sensors is the foremost. HDLR manipulators work in harsh conditions, which requires rugged sensors that can withstand those conditions. Unfortunately, non-contact visual sensors are sensitive and easily disturbed. It is also noteworthy that the proposed visual sensor system could contribute to increasing the automation level and task flexibility of HDLR manipulators in areas other than just visual TCP pose estimation (RP-I). Other topics include vision-aided control in auxiliary tasks, as discussed in P-V, along with object detection and recognition in the workspace. Presently, some OEMs employ laser scanners in their

HDLR manipulators, while the potential of computer vision paired with cameras is not yet unlocked. Therefore, it is likely that camera-based systems are first adopted to simple, auxiliary functions to which computer vision can add value.

4.2 Vision-based Control for Precise TCP Positioning of HDLR Manipulators (RP-II)

Can a low-cost camera be used to achieve precise (sub-centimeter) TCP positioning accuracy in HDLR manipulators?
What are the main challenges in realizing precise TCP control for HDLR manipulators using visual sensing?

For RP-II, the main objective is to accurately position the TCP of an HDLR manipulator to a visually detected OOI using a low-cost camera. Practical applications related to this problem include, for example, tool swapping and positioning the tool to pre-drilled holes. As previously discussed, hand-eye calibration methods developed for conventional industrial robots are not practical for HDLR manipulators in dynamic, unstructured environments. To solve the extrinsic sensor-to-robot calibration problem, P-V proposes motion-based local calibration. Using the trajectory-matching-based method, the rotation difference between the visual sensor's frame and the TCP frame can be computed with high accuracy. Then, the position error between the tool and an OOI is computed directly from an image. The visual control input for the robotic system is obtained by applying the rotation transformation to the position error. This enables the circumvention of numerous errors in the serial chain kinematic model used to formulate the TCP for control purposes, along with errors arising from nonlinearities. Combined with highly accurate local calibration, minimizing the image-based position and orientation errors will, in theory, position the tool at the target OOI. Some aspects related to robotic control, such as joint constraints and singularities, are not in the scope of P-V. It should also be highlighted that the aim is to maintain conventional rigid-body-based modeling and control systems that are prevalent in the industry. In [82], it was found that nonlinear model-based control methods produce the most advanced control performance for (hydraulic) HDLR manipulators. However, as shown in P-V, such complex controller structures may be averted by employing vision-based control for tasks requiring high precision.

While the formalism for motion-based local calibration is fully described in P-V, the research also benefits from P-I, where the eye-in-hand configuration for visual SLAM-based TCP pose tracking of an HDLR manipulator is initially deemed as sufficiently accurate. The research also utilizes the findings of P-III, which attempts to find the optimal method for robust pose trajectory matching in robotic applications. The experimental results presented in P-V demonstrate that sub-centimeter positioning accuracy is reliably achieved in the tested cases using the trajectory-matching-based method. As discussed, however, the performance relies of multiple aspects, such as the VO/SLAM system, pose trajectory qualities, and small trajectory matching errors. The line-equation-based method, while theoretically viable, suffers from accumulated errors and does not perform reliably. The local calibration also only holds in the local plane, in which the calibration is performed. Changing the orientation of the tool requires another calibration, which can be cumbersome to repeat depending on the application. Therefore, the outcome related to the first research question of RP-II is that sub-centimeter absolute positioning accuracy can be achieved for HDLR manipulators using a low-cost camera, which is regarded as the main contribution of this dissertation. However, the success highly depends on the setup and circumstances. Thus, further tests are required for practical applications with more complex HDLR manipulators. It is shown in the laboratory experiments that the camera frame and the TCP frame can be matched accurately using the rotation difference. However, this is reliant on an accurate estimation of the visual sensor's egomotion, and the assumption that the kinematic TCP is accurate enough, so that pose trajectory matching can be executed with small errors.

The line-equation-based method for range adjustment, especially, highlights the challenges of realizing visual control. While the idea is simplistic, the errors accumulate from various sources. These include camera placement and alignment, orientation adjustment, and tracking the circular path. The limitations and challenges related to the topics of this dissertation are discussed further below.

While P-V examines precise vision-aided control for HDLR manipulators, enabled by highly accurate alignment between the camera frame and the TCP frame, using the eye-in-hand configuration, it should be mentioned that the results of P-II and P-IV suggest that similar accuracy can be achieved with the eye-to-hand configuration. However, degradation is to be expected, especially in the orientation signals, with increased view distances. As highlighted in P-II and P-IV, the quality of the orientation signals can be improved to some extent with signal filtering. For TCP control purposes, the eye-in-hand configuration is still perceived as much more useful.

4.3 Limitations and Challenges

This section discusses the limitations and challenges faced in this dissertation. It also aims to address the second research question of RP-II, which relates to highlighting the main challenges in realizing vision-based control for HDLR manipulators.

4.3.1 Robustness and Reliability with Visual Sensors

Visual sensors are inherently non-contact sensors that provide a large amount of data. Non-contact infers that the sensor element is not in direct physical contact with the sensed target. An image contains a vast amount of information, but only a small part of it is typically useful. Extracting the useful information, such as detecting an OOI, can be challenging. In general, visual measurements are prone to challenges in robustness and reliability. These issues arise from the fact that visual measurements are easily degraded due to various reasons, such as mechanical vibrations, water and vapor, occlusions, and insufficient lighting. These challenges are highlighted for HDLR manipulators working in dynamic, unstructured environments, which also require the sensors to be rated for IP67 protection. While such rugged cameras can be realized with increased manufacturing costs, the other discussed environmental factors still pose significant challenges to visual sensing.

Regarding VO/SLAM systems, this dissertation does not consider long-term usage. VO/SLAM systems are prone to drifting and outliers that degrade the performance after a longer period of continuous execution. As discussed, the SLAM front-end provides real-time localization using VO, while the back-end of SLAM handles mapping. In a confined workspace, it is theoretically possible to first map the environment and then localize based on the map. Some features related to SLAM systems, such as loop closing and global optimization, are not preferential if they impose discontinuities on the real-time pose variables. However, as highlighted, VO alone is subject to drifting, especially in long-term usage, which is not acceptable for robotic control purposes. The research of this dissertation also makes the assumption that there are enough textured surfaces in the HDLR manipulator's workspace to utilize feature-based VO/SLAM systems. While this may be a fair assumption for underground mines, other types of environments, such as construction sites or forests, are not considered in the scope of this research. Even for underground mines, VO/SLAM systems optimized to such environments are likely required. As highlighted in Section 2, one of the current research directions in VO/SLAM systems is optimization for challenging scenarios.

4.3.2 Robotic Control

The scope of this dissertation involves generic 6 DOF TCP pose estimation, which alone is not sufficient to realize complete manipulator control, as modern control system frameworks also require information of the individual joint states. This work examines direct TCP pose estimation using a visual sensor system, whereas the established method is to formulate the TCP based on joint sensors and a rigid-body-based forward kinematic model. As discussed, the assumption of rigidness in the modeling deteriorates the TCP positioning accuracy and repeatability in HDLR manipulators due to the numerous kinematic and non-kinematic errors present in the system. While the control of structurally flexible manipulators is a research field of its own, robotic control methods in the industry are still closely dependent on the rigid-body-based kinematic formulation and simple controller structures (e.g., Proportional-Integral-Derivative). As mentioned, it has been shown that nonlinear model-based control methods can provide the most advanced control performance for hydraulic HDLR manipulators. However, these methods are also based on rigid bodies and have not (yet) been adopted to the industry. Control of flexible manipulators is even more complex, and consequently, accounting for structural bending in HDLR manipulators in the industry is mostly limited to static compensation computations to enhance the rigid-body-based forward kinematic TCP accuracy.

To utilize predominant robotic control methods, the inverse kinematic relation from the TCP to the manipulator's base is required. For a system with direct TCP pose estimation, this would involve estimating the individual joint states based on the current TCP pose, which is not investigated within this research. The matter is further complicated by the extrinsic sensor-to-robot calibration, which is discussed below. An ultimate goal is to realize TCP control for HDLR manipulators, while the individual joint sensors could be completely omitted to streamline the mechanical structure, at least for non-redundant manipulators.

4.3.3 Sensor-to-Robot and Sensor-to-Sensor Calibration

The methods used in this dissertation take advantage of the joint encoders to compute the rigid-body-based forward kinematic TCP and then use it for calibration and ground-truth purposes, which remains another challenge. Point set matching is not sufficient to solve the extrinsic sensor-to-robot calibration, when the individual joint states (and thus, the kinematic TCP) is not available. The original idea was to use the point set matching for initial calibration and then to use the visual sensor system to track the dynamic motion of the TCP. The sensor-to-robot calibration issue is one of the most fundamental challenges, which is a limiting factor in benefiting from visual measurements in applications with HDLR manipulators. Consequently, the proposed conceptual visual sensor system estimating the TCP pose may also be useful as a secondary sensing system (joint sensors being the primary) in auxiliary tasks or bending estimation, for example. As shown in P-V, there are potential benefits in using the eye-in-hand configuration for visual control purposes. This can ultimately be used to increase the automation level of HDLR manipulators, including task flexibility.

This research takes advantage of a state-of-the-art motion capture system, which provides the necessary multi-camera system calibrations, marker detection and pose estimation. Therefore, the global poses of the markers, their IDs, and accurate tracking capabilities are readily accessible. The high performance is reflected in the cost of the system, which is currently not realistic for commercial HDLR manipulators. In-house development of similar high-end systems would be a challenging task, but the results presented in this thesis demonstrate the potential of such a system. The multi-camera system utilizes a reference world frame, which is required in view of the cameras. Then, the marker poses are expressed with respect to the world frame. Thus, the issue of extrinsic sensor-to-robot calibration persists also in the eye-to-hand system as there is no straightforward manner to find the precise relation between the set world frame and the base frame of the manipulator. As discussed, relevant calibration frameworks developed for conventional industrial robots are generally not practical for HDLR manipulators due to their nonlinear characteristics and harsh working environments.

4.3.4 Visual Object Detection and Pose Estimation

This work utilizes two types of visual fiducial markers for OOI detection and pose estimation. In RP-I, the state-of-the-art marker-based tracking system utilizes reflective markers. In addition, ArUco markers are utilized to represent the tool and an OOI for the vision-aided control system in RP-II. For a practical system, the ArUco markers need to be replaced with application-specific OOI. Presently, state-of-the-art visual object detection methods are mostly learning-based. Thus, solving the problem for application-specific OOI will likely require advanced learning-based methods to detect different tools and target OOI in images.

Visual detection of a specific OOI in the camera's view is only the first step. The second step is visual pose estimation of the target OOI, which is a requirement for accurate vision-based control purposes. The accuracy of visual pose estimation highly depends on the quality of the OOI detection outcome. Compared to a planar ArUco marker with 4 well-defined corners, for example, pose estimation of geometrically complex OOI can be challenging to realize with sufficient accuracy. The results of P-V show that precise tool positioning for HDLR manipulators can be achieved if the poses of the tool and the target OOI can be accurately estimated.

5 CONCLUSION

To conclude, the heavy machinery industry, including the TCP control of HDLR manipulators, is taking steps toward increased levels of automation and autonomous systems, which require new intelligent algorithms and sophisticated sensing solutions. In order to automate work tasks of HDLR manipulators and increase their task flexibility, a key challenge is replacing human vision and decision making with sensors and computerized algorithms. This dissertation investigates the possibilities of exploiting advanced visual sensing solutions for HDLR manipulators working in unstructured environments.

In RP-I, the overarching objective was to estimate the generic 6 DOF TCP pose of HDLR manipulators using visual sensing. The aim was to study the feasibility of camera-based sensing for this purpose, and to examine methods for increasing its robustness and reliability. P-I examines a state-of-the-art feature-based VO/SLAM system for tracking the generic 6 DOF TCP pose of a HDLR manipulator in a confined space. It also serves as a basis for the follow-up research. P-II proposes a conceptual visual sensor system consisting of eye-in-hand/eye-to-hand cooperation, with emphasis placed on increasing the robustness and reliability of the system via sensor redundancy. P-III investigates the sensor system's calibration problem and utilizes a robust point set matching pipeline, taking advantage of the full 6 DOF TCP pose data in the process. P-IV addresses the data fusion problem by proposing a real-time capable data fusion method based on sliding window variances. Together, P-I through P-IV seek to address RP-I by presenting a conceptual visual sensor system utilizing eye-in-hand/eye-to-hand cooperation. However, as discussed, many challenges related to control and calibration remain.

In RP-II, the main objective was to investigate vision-based sensing methods for precise TCP positioning of HDLR manipulators. The aim was to utilize a low-cost camera, while also examining the main challenges related to vision-based control of HDLR manipulators. P-V proposes a methodology to enable precise TCP positioning by computing the pose error between a tool and an OOI directly in the image frame while using highly accurate motion-based local calibration to align the rotation of the camera frame with the rigid-body-based forward kinematic TCP frame, which is used for robotic control. For the tested cases in P-V, it is shown that sub-centimeter accuracy is reliably obtained using the trajectory-matching-based method. This level of accuracy is challenging to achieve with HDLR manipulators due to their nonlinear characteristics. Thus, RP-II is addressed by P-V, which is a result of the developments in P-I and P-III.

Overall, the initially set RPs are successfully addressed within this dissertation, and the results contribute toward realizing vision-based control for HDLR manipulators. Five original papers are presented in this compendium thesis, with the main contribution arising from P-V, in which sub-centimeter TCP positioning accuracy is achieved for an HDLR manipulator using the developed methods.

The proposed methods in this dissertation were validated in a laboratory setting and thus are mainly on a conceptual level. Therefore, further development is required for practical implementations. While mobile machines with on-board HDLR manipulators could benefit greatly from advanced computer vision systems, harsh working environments issue many practical challenges related to robustness, reliability, and fault tolerance. These challenges include lighting conditions, varying surface textures, dust and vapor, water, sensor mounting positions, and mechanical vibrations. Presently, advanced computer vision algorithms and camera-based vision are emerging fields for HDLR manipulators. Therefore, it is expected that the first practical applications in the industry are related to automating relatively simple, auxiliary tasks that can be performed under human supervision.

Related to RP-I, the ultimate future goal is to eventually omit the individual joint sensors, which are currently used in mechanically complex HDLR manipulators, completely. While visual TCP pose estimation (RP-I) may provide a solution for this, controlling the manipulator remains a challenge as modern control systems require knowledge of the joint states. Furthermore, the extrinsic sensor-to-robot calibration in the long term remains a significant challenge. Related to RP-II, a future desire is to utilize the proposed vision-aided control method using motion-based local calibration for practical applications in the industry that require precise TCP positioning. This requires further development of advanced vision-based OOI detection and pose estimation methods to realize the application-specific parameters.

REFERENCES

[1] S. Popić and B. Miloradović, “Light weight robot arms-an overview,” *INFOTEH-JAHORINA*, vol. 14, 2015.

[2] *Ponsse, Active Crane*, https://www.ponsse.com/en/products/tailored-solutions/product/-/p/activecranebisonbuffaloelephant, Accessed: 05.06.2023.

[3] *HIAB, Crane Tip Control*, https://www.hiab.com/sv/media/newsroom/hiab-crane-tip-control, Accessed: 05.06.2023.

[4] *John Deere, Intelligent Boom Control (IBC)*, https://www.deere.com/en/technology-products/forestry-and-logging-technology/operator-assistance-technology/, Accessed: 05.06.2023.

[5] J. Denavit and R. S. Hartenberg, “A kinematic notation for lower-pair mechanisms based on matrices,” *Journal of Applied Mechanics*, vol. 22, no. 2, pp. 215–221, 1955.

[6] L. Lopes, T. Miklovicz, E. Bakker, and Z. Milosevic, “The benefits and challenges of robotics in the mineral raw materials sector-an overview,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1507–1512.

[7] R. Tiusanen, T. Malm, and A. Ronkainen, “An overview of current safety requirements for autonomous machines–review of standards,” *Open Engineering*, vol. 10, no. 1, pp. 665–673, 2020.

[8] T. Machado, D. Fassbender, A. Taheri, *et al.*, “Autonomous heavy-duty mobile machinery: A multidisciplinary collaborative challenge,” in *2021 IEEE International Conference on Technology and Entrepreneurship (ICTE)*, IEEE, 2021, pp. 1–8.

[9] J. Vihonen, J. Mattila, and A. Visa, "Joint-space kinematic model for gravity-referenced joint angle estimation of heavy-duty manipulators," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 12, pp. 3280–3288, 2017.

[10] F. Wittmann, O. Lambercy, and R. Gassert, "Magnetometer-based drift correction during rest in IMU arm motion tracking," *Sensors*, vol. 19, no. 6, p. 1312, 2019.

[11] H. Hyyti, V. V. Lehtola, and A. Visala, "Forestry crane posture estimation with a two-dimensional laser scanner," *Journal of Field Robotics*, vol. 35, no. 7, pp. 1025–1049, 2018.

[12] A. H. Kashani, W. S. Owen, N. Himmelman, *et al.*, "Laser scanner-based end-effector tracking and joint variable extraction for heavy machinery," *The International Journal of Robotics Research*, vol. 29, no. 10, pp. 1338–1352, 2010.

[13] P. Cheng, B. Oelmann, and F. Linnarsson, "A local positioning system for loader cranes based on wireless sensors—a feasibility study," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 8, pp. 2881–2893, 2011.

[14] C. Zhang, A. Hammad, and S. Rodriguez, "Crane pose estimation using UWB real-time location system," *Journal of Computing in Civil Engineering*, vol. 26, no. 5, pp. 625–637, 2012.

[15] A. P. R. Lauer, O. Lerke, B. Blagojevic, *et al.*, "Tool center point control of a large-scale manipulator using absolute position feedback," *Control Engineering Practice*, vol. 131, p. 105 388, 2023.

[16] S. Cebollada, L. Payá, M. Flores, *et al.*, "A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data," *Expert Systems with Applications*, vol. 167, p. 114 195, 2021.

[17] A. Singh, V. Kalaichelvi, and R. Karthikeyan, "A survey on vision guided robotic systems with intelligent control strategies for autonomous tasks," *Cogent Engineering*, vol. 9, no. 1, p. 2 050 020, 2022.

[18] M. E. Kiziroglou, D. E. Boyle, E. M. Yeatman, and J. J. Cilliers, "Opportunities for sensing systems in mining," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 278–286, 2016.

[19] P. Corke, J. Roberts, and G. Winstanley, "Vision-based control for mining automation," *IEEE Robotics & Automation Magazine*, vol. 5, no. 4, pp. 44–49, 1998.

[20] A. Bonchis, E. Duff, J. Roberts, and M. Bosse, "Robotic explosive charging in mining and construction applications," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 1, pp. 245–250, 2013.

[21] J. Xu, Z. Hou, Z. Liu, and H. Qiao, "Compare contact model-based control and contact model-free learning: A survey of robotic peg-in-hole assembly strategies," *arXiv preprint arXiv:1904.05240*, 2019.

[22] J. Gallwey, M. Eyre, and J. Coggan, "A machine learning approach for the detection of supporting rock bolts from laser scan data in an underground mine," *Tunnelling and Underground Space Technology*, vol. 107, p. 103 656, 2021.

[23] S. K. Singh, B. P. Banerjee, and S. Raval, "A review of laser scanning for geological and geotechnical applications in underground mining," *International Journal of Mining Science and Technology*, 2022.

[24] A. Chehri, P. Fortier, and P. M. Tardif, "UWB-based sensor networks for localization in mining environments," *Ad Hoc Networks*, vol. 7, no. 5, pp. 987–1000, 2009.

[25] M.-G. Li, H. Zhu, S.-Z. You, and C.-Q. Tang, "UWB-based localization system aided with inertial sensor for underground coal mine applications," *IEEE Sensors Journal*, vol. 20, no. 12, pp. 6652–6669, 2020.

[26] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part I: The first 30 years and fundamentals," *IEEE Robotics and Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[27] S. A. Mohamed, M.-H. Haghbayan, T. Westerlund, *et al.*, "A survey on odometry for autonomous navigation systems," *IEEE Access*, vol. 7, pp. 97 466–97 486, 2019.

[28] C. Cadena, L. Carlone, H. Carrillo, *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[29] K. Yousif, A. Bab-Hadiashar, and R. Hoseinnezhad, "An overview to visual odometry and visual SLAM: Applications to mobile robotics," *Intelligent Industrial Systems*, vol. 1, no. 4, pp. 289–311, 2015.

[30] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, and H. Voos, "Visual SLAM: What are the current trends and what to expect?" *Sensors*, vol. 22, no. 23, p. 9297, 2022.

[31] M. Etxeberria-Garcia, M. Zamalloa, N. Arana-Arexolaleiba, and M. Labayen, "Visual odometry in challenging environments: An urban underground railway scenario case," *IEEE Access*, vol. 10, pp. 69 200–69 215, 2022.

[32] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: A survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[33] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: A survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.

[34] A. Macario Barros, M. Michel, Y. Moline, *et al.*, "A comprehensive survey of visual SLAM algorithms," *Robotics*, vol. 11, no. 1, p. 24, 2022.

[35] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2014, pp. 15–22.

[36] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European conference on computer vision*, Springer, 2014, pp. 834–849.

[37] R. Gomez-Ojeda, F.-A. Moreno, D. Zuñiga-Noël, *et al.*, "PL-SLAM: A stereo SLAM system through the combination of points and line segments," *IEEE Transactions on Robotics*, 2019.

[38] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, pp. 2100–2106.

[39] C. Campos, R. Elvira, J. J. G. Rodríguez, *et al.*, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[40] M. Labbé and F. Michaud, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[41] D. Sharafutdinov, M. Griguletskii, P. Kopanev, *et al.*, "Comparison of modern open-source visual SLAM approaches," *Journal of Intelligent and Robotic Systems*, vol. 107, no. 3, p. 43, 2023.

[42] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017. DOI: 10.1109/TRO.2017.2705103.

[43] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 Internatioal Conference on Computer Vision*, IEEE, 2011, pp. 2564–2571.

[44] H.-J. Chien, C.-C. Chuang, C.-Y. Chen, and R. Klette, "When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, IEEE, 2016, pp. 1–6.

[45] S. H. Hashemi and J. Mattila, "Global asymptotic convergent observer for SLAM," *IEEE Access*, vol. 10, pp. 122 414–122 422, 2022.

[46] *OptiTrack*, https://optitrack.com/applications/robotics/, Accessed: 05.06.2023.

[47] *Vicon*, https://www.vicon.com/applications/engineering/, Accessed: 05.06.2023.

[48] *Qualisys*, https://www.qualisys.com/engineering/robotics-and-uav/, Accessed: 05.06.2023.

[49] *PhaseSpace*, https://www.phasespace.com/applications/robotics/, Accessed: 05.06.2023.

[50] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[51] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *2011 IEEE International Conference on Robotics and Automation*, IEEE, 2011, pp. 3400–3407.

[52] X. X. Lu, "A review of solutions for perspective-n-point problem in camera pose estimation," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1087, 2018, p. 052 009.

[53] K. M. Lundeen, S. Dong, N. Fredricks, *et al.*, "Optical marker-based end effector pose estimation for articulated excavators," *Automation in Construction*, vol. 65, pp. 51–64, 2016.

[54] M. M. Soltani, Z. Zhu, and A. Hammad, "Skeleton estimation of excavator by detecting its parts," *Automation in Construction*, vol. 82, pp. 1–15, 2017.

[55] J. Kocić, N. Jovičić, and V. Drndarević, "Sensors and sensor fusion in autonomous vehicles," in *2018 26th Telecommunications Forum (TELFOR)*, IEEE, 2018, pp. 420–425.

[56] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[57] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, p. 4220, 2020.

[58] D. J. Yeong, G. Velasco-Hernandez, J. Barry, J. Walsh, *et al.*, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[59] J. G. García, A. Robertsson, J. G. Ortega, and R. Johansson, "Sensor fusion for compliant robot motion control," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 430–441, 2008.

[60] B. Olofsson, J. Antonsson, H. G. Kortier, *et al.*, "Sensor fusion for robotic workspace state estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 5, pp. 2236–2248, 2015.

[61] S.-L. Sun and Z.-L. Deng, "Multi-sensor optimal information fusion Kalman filter," *Automatica*, vol. 40, no. 6, pp. 1017–1023, 2004.

[62] S. Yazdkhasti and J. Z. Sasiadek, "Multi sensor fusion based on adaptive Kalman filtering," in *Advances in aerospace guidance, navigation and control*, Springer, 2018, pp. 317–333.

[63] J. Gross, Y. Gu, S. Gururajan, *et al.*, "A comparison of extended Kalman filter, sigma-point Kalman filter, and particle filter in GPS/INS sensor fusion," in *AIAA guidance, navigation, and control conference*, 2010, p. 8332.

[64] E. Bostanci, B. Bostanci, N. Kanwal, and A. F. Clark, "Sensor fusion of camera, GPS and IMU using fuzzy adaptive multiple motion models," *Soft Computing*, vol. 22, no. 8, pp. 2619–2632, 2018.

[65] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *2008 11th international conference on information fusion*, IEEE, 2008, pp. 1–6.

[66] H. F. Durrant-Whyte, "Sensor models and multisensor integration," in *Autonomous Robot Vehicles*, Springer, 1990, pp. 73–89.

[67] W. Elmenreich, "Fusion of continuous-valued sensor measurements using confidence-weighted averaging," *Journal of Vibration and Control*, vol. 13, no. 9-10, pp. 1303–1312, 2007.

[68] F. Dornaika and R. Horaud, "Simultaneous robot-world and hand-eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 4, pp. 617–622, 1998.

[69] I. Enebuse, M. Foo, B. S. K. K. Ibrahim, *et al.*, "A comparative review of hand-eye calibration techniques for vision guided robots," *IEEE Access*, vol. 9, pp. 113 143–113 155, 2021.

[70] E. Pedrosa, M. Oliveira, N. Lau, and V. Santos, "A general approach to hand–eye calibration through the optimization of atomic transformations," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1619–1633, 2021.

[71] J. Jiang, X. Luo, Q. Luo, *et al.*, "An overview of hand-eye calibration," *The International Journal of Advanced Manufacturing Technology*, vol. 119, no. 1-2, pp. 77–97, 2022.

[72] G. Flandin, F. Chaumette, and E. Marchand, "Eye-in-hand/eye-to-hand cooperation for visual servoing," in *Proc. 2000 ICRA. Millennium Conf. IEEE Int. Conf. Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, IEEE, vol. 3, 2000, pp. 2741–2746.

[73] Z. Taylor and J. Nieto, "Motion-based calibration of multimodal sensor extrinsics and timing offset estimation," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1215–1229, 2016.

[74] K. Pachtrachai, F. Vasconcelos, P. Edwards, and D. Stoyanov, "Learning to calibrate-estimating the hand-eye transformation without calibration objects," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7309–7316, 2021.

[75] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[76] W. J. Wilson, C. W. Hulls, and G. S. Bell, "Relative end-effector control using Cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 684–696, 1996.

[77] F. Chaumette, "Potential problems of stability and convergence in image-based and position-based visual servoing," in *The Confluence of Vision and Control*, Springer, 2007, pp. 66–78.

[78] A. A. Hafez, E. Cervera, and C. Jawahar, "Hybrid visual servoing by boosting IBVS and PBVS," in *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, IEEE, 2008, pp. 1–6.

[79] W. Pan, M. Lyu, K.-S. Hwang, *et al.*, "A neuro-fuzzy visual servoing controller for an articulated manipulator," *IEEE Access*, vol. 6, pp. 3346–3357, 2018.

[80] S. Zhou, C. Shen, F. Pang, *et al.*, "Position-based visual servoing control for multi-joint hydraulic manipulator," *Journal of Intelligent and Robotic Systems*, vol. 105, no. 2, p. 33, 2022.

[81] S. Lampinen, L. Niu, L. Hulttinen, *et al.*, "Autonomous robotic rock breaking using a real-time 3D visual perception system," *Journal of Field Robotics*, vol. 38, no. 7, pp. 980–1006, 2021.

[82] J. Mattila, J. Koivumäki, D. G. Caldwell, and C. Semini, "A survey on control of hydraulic robotic manipulators with projection to future trends," *IEEE/ASME Transactions on Mechatronics*, vol. 22, no. 2, pp. 669–680, 2017.

PUBLICATIONS

PUBLICATION
I

Application of simultaneous localization and mapping for large-scale manipulators in unknown environments

P. Mäkinen, M. M. Aref, J. Mattila, and S. Launis

In *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, 2019
DOI: 10.1109/CIS-RAM47153.2019.9095770

Application of Simultaneous Localization and Mapping for Large-scale Manipulators in Unknown Environments

Petri Mäkinen, Mohammad M. Aref and Jouni Mattila
Unit of Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
Emails: petri.makinen@tuni.fi,
m.aref@ieee.org,
jouni.mattila@tuni.fi

Sirpa Launis
Sandvik Mining and Construction
Tampere, Finland
Email: sirpa.launis@sandvik.com

Abstract—**In this paper, we study the application of simultaneous localization and mapping (SLAM) for estimating the tool center point (TCP) 6 degrees-of-freedom (DOF) pose of a large-scale hydraulic manipulator without a priori knowledge of the environment. We attach a stereo camera near the TCP of the manipulator and perform SLAM by utilizing the open source version of ORB-SLAM2. In offline experiments, the camera frame and the TCP frame are extrinsically calibrated using an iterative closest point search to match a point cloud of poses from the SLAM module with a point cloud of ground-truth TCP poses, which are obtained from joint encoder measurements along with a kinematic model of the manipulator. The estimated TCP trajectory provided by the SLAM is then compared to the ground-truth TCP trajectory. These preliminary experiments show that a pure visual SLAM algorithm can perform reasonably well in this application scenario. Limitations and future work are also discussed.**

Index Terms—**manipulators, simultaneous localization and mapping, position measurement, rotation measurement**

I. Introduction

Autonomous vehicles have been avidly studied in the past decade, as the trend tends toward fully independently operating systems. This includes a multitude of mobile, heavy-duty machines that utilize large-scale manipulators (also called booms) to complete work tasks. In underground mining applications, such manipulators are utilized in tunneling jumbos and drill rigs, for example. Although these machines are currently human operated, on-site or remotely via teleoperation, each joint has a sensor so that the tool center point (TCP) of the manipulator can be measured based on the joint states. This is because knowledge of the TCP pose is essential for accurate drilling, as it has a direct effect on the mining progress made with each blast. The number of actuator degrees-of-freedom (DOF) in these manipulators is typically up to 8, which means that 8 sensors, including their waterproof enclosures, associated cabling, and high-precision mechanical couplings, add up to a significant bill of materials, and thus, the cost per machine. Moreover, a single machine can have multiple booms. For instance, tunneling jumbos typically have one to three drilling booms, which further add to the bill of materials and motivate research for new, alternative TCP measurement methods designed for this type of application.

Such methods have been explored for many similar applications in the literature. For example, in [1], a laser scanner was used for end-effector tracking and joint variable extraction of a heavy mining shovel's dipper. In [2], a low-cost 2D laser scanner was used to estimate the posture of a forestry crane, with a reported average tip position accuracy of 4.3 cm. With scanner-based measurements, the sparsity of the acquired point cloud becomes an issue at larger distances due to the low spatial resolution of the sensor. A part detection-based scheme for estimating the 2D pose of an excavator was studied in [3]. The method used a database of synthetic images comprising different parts of an excavator, which were used to train part detectors. A single camera was then used for extracting the skeleton of the excavator based on the detected parts. The feasibility of a local positioning system for loader cranes using wireless sensors was studied in [4]. The described method used inertial sensors for joint angle measurements and an ultrasonic transducer for measuring the length of a telescopic joint. In [5], a gravity-referenced joint angle estimation scheme using three-axis linear accelerometers and three-axis rate gyros was proposed, with reported joint angle sensing errors of ±1 degree. An ultra-wide band (UWB) based real-time location system for estimating crane poses was studied in [6]. However, based on the UWB system error alone, which was approximately 30 cm, the accuracy is not sufficient for applications requiring precise positioning. In [7], an optical marker-based end-effector pose estimation scheme was presented for articulated excavators, with encouraging initial results. Nonetheless, these sensing methods do not particularly fit the present application of interest, as drilling booms are typically complex structures (8 actuator DOF) with considerable maximum lengths. The confined workspace also restricts the placement of sensors, such as cameras, in the environment around the machine.

This work was funded by the Doctoral School of Industry Innovations (DSII) of Tampere University.

A potential solution would be to place a camera directly at the TCP and perform pose estimation based on visual odometry or more advanced simultaneous localization and mapping (SLAM) algorithms. SLAM has attracted considerable attention in the past few decades, as this technology is a vital component of any autonomous vehicle: A machine cannot operate independently unless it is aware of its location in relation to the environment. Thus, the main objective of visual SLAM is to constantly perform localization based on visual feeds, while simultaneously building a map of the surroundings. Only recently have SLAM technologies showed signs of advancing toward the levels of maturity and reliability that are required for autonomous systems. Some areas, however, such as fail-safe systems, are still relatively unexplored, as discussed in a recent survey paper [8].

Numerous SLAM methods have been engineered over the years, with previous ones surveyed in [9]. Some of the more recent and most popular monocular schemes include DVO [10] and SVO [11]. Stereo and RGB-D methods have also been presented, for example, RTAB-MAP [12], in which the authors also provided comparative results using many SLAM algorithms available on ROS. A stereo SLAM method using ORB [13] features and line segments was proposed in [14]. The rationale was that the inclusion of line segments improves the performance in low-textured environments with planar structures, where a low number of point features can be extracted. The reported performance was similar to that of ORB-SLAM2 [15], which is another SLAM method. ORB-SLAM2 is fully based on ORB features, and can be used with mono, stereo, or RGB-D input. The authors extended their work to include inertial sensors in [16].

In this paper, inspired by the recent advances in SLAM technologies in vehicle positioning, we study the application and feasibility of SLAM in a much different setting. Specifically, we apply SLAM to estimate the TCP pose of a large-scale articulated crane in an unknown, confined space. The motivation is that for mining manipulators underground, the workspace area is typically small and confined, with walls closing in on each direction. The laboratory-grade simulation of such an environment is a test wall built from decorative stones. We attach a low-cost stereo camera near the TCP of an articulated heavy-duty crane so that the camera faces the test wall. Based on the literature review and the availability of state-of-the-art open source SLAM algorithms, we utilize the feature-based ORB-SLAM2[1], which is a tried and tested algorithm with excellent localization capabilities in varying environments. To avoid the issue of scale ambiguity, stereo vision is employed in the experiments. Data analysis is performed by comparing ground-truth TCP poses, obtained using joint encoders, with calibrated SLAM output poses. For offline data analysis, the calibration between the SLAM poses and the ground-truth poses is conducted using an iterative closest point (ICP) algorithm.

The remainder of this paper is outlined as follows: In Section II, a description of the application is provided; it is followed by Section III, in which the experimental setup is presented. In Section IV, the data analysis is presented with comparative results. Finally, in Section V, the conclusion is provided.

II. Simultaneous Localization and Mapping in the Proposed Application

Development drilling, in which tunnel networks are formed, is one of the basic operations in underground mining. Using this as an example, the TCP of a drilling boom is typically moved within an area resembling a rectangle that corresponds to the profile of the tunnel being mined. Dozens of holes, many meters in depth, are drilled inside the profile. The TCP is driven from drilling point to drilling point in a pre-planned manner, while during a drilling operation the current TCP pose is maintained. As the TCP is moved within a small area, with back-and-forth motions, it is possible to establish a comprehensive local map with SLAM before even beginning the drilling. This also suggests that loop detection and closing features of SLAM can be highly useful in correcting any drift in the pose estimates. Especially for the proposed application of TCP pose estimation, it is desirable that the tracking can be maintained at all times. Notably, heavy drilling induces severe vibrations in the manipulator and the machine, which could affect the tracking performance. In this case, having a local map would be useful, as relocalization based on the map can be instantly performed after a drilling operation, during which the joint positions can be locked in place by the control system. After the drilling plan is completed, explosive charges are placed inside the drill holes, after which blasting occurs. As the depth of a drill hole is measured in meters, any error in the TCP's pose will directly degrade the blasting result. Regarding the desired accuracy in this type of application, the rough target values for positioning and orienting are 1 cm and 1°, respectively.

The SLAM we utilize in this work, ORB-SLAM2, is a feature-based method that utilizes only ORB features, which have quickly become a popular choice due to their computational efficiency and good invariance to viewpoint changes. ORB-SLAM2 consists of three parallel threads: tracking, local mapping, and loop closing. If the tracking is lost, the system is capable of relocalization using a bag-of-words place recognition module, which is based on DBoW2 [17]. A pure localization mode is also available, in which the mapping and loop closing features are disabled. Importantly, the system applies bundle adjustment (BA) for optimization purposes at various stages of the algorithm: for optimizing the camera's orientation and position by minimizing the reprojection error between matched 3D landmarks and 2D key points in the tracking thread (motion-only BA); for optimizing a window of keyframes and points in the local mapping thread (local BA); and after a loop closure for optimizing all keyframes and points (full BA) [15]. The Levenberg-Marquardt method used for optimization is implemented in [18].

III. Experimental Setup

A. System Description

The manipulator we used for testing was a HIAB033, which is a hydraulically actuated crane. The manipulator, illustrated in Fig. 1, additionally had a spherical wrist with a gripper attached to it, yielding a total of 6 active actuator DOF: rotate,

[1] https://github.com/raulmur/ORB_SLAM2

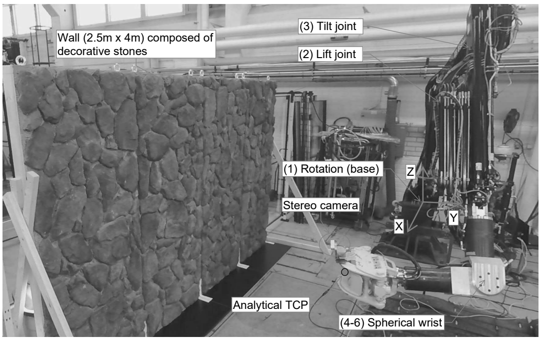


Fig. 1. The figure illustrates the test setup, in which the manipulator was positioned so that the stereo camera attached to the gripper faced the test wall. The goal was to estimate the TCP pose of the manipulator based on pure visual SLAM, which extracted the required features from the test wall. The joints of the manipulator are labeled from 1 to 6, with the base coordinate system also shown.

lift, tilt, and 3 DOF in the wrist. The manipulator also had two extension cylinders, which were disabled during the tests. Each active joint was instrumented with an incremental encoder.

The control system of the HIAB was a dSPACE DS1005 PPC controller board and a development PC, which ran in 3 ms sampling time. A Stereolabs ZED stereo camera was used for visual measurements. The ZED was connected to a laptop via its USB interface, and images were captured using ZED SDK's Matlab plug-in by using a UDP trigger signal, which was established from the dSPACE development PC to the ZED laptop. The UDP trigger signal was transmitted at 8×3 ms time intervals, which was dictated by the time it took for the ZED SDK to capture and save a pair of grayscale, 672×376 resolution images. The trigger signal ensured that the image data recorded with the laptop and the encoder data recorded with dSPACE could be synchronized with each other.

For the SLAM experiments, a wall made out of decorative stones was built. This was our laboratory-grade simulation of a mine wall. The wall was 2.5×4 m in dimensions, and the stones were cemented to the wall randomly,, albeit they comprised some recurring shapes. Varying motions were applied to the manipulator, and data was recorded using the camera and the joint encoders. Then, the corresponding image sequences were extracted from the data. Each sequence library was then processed by the ORB-SLAM2 stereo algorithm, which provided a pose sequence corresponding to each input sequence. In the SLAM settings, the number of ORB features was set to 2000, and the camera FPS was set to 41.6667, according to the UDP trigger signal. Fig. 2 shows the ZED stereo camera attached to the gripper of the manipulator, as well as an example view of the detected ORB features from the test wall.

B. Ground-truth TCP Pose

As a ground truth, or reference, measurement of the TCP pose is required to evaluate the estimated SLAM poses, a kinematic model of the manipulator was formulated. The states of the six active joints, measured with encoders, were then used with the forward kinematic model of the manipulator to formulate the TCP pose, which was used as the ground-truth measurement.

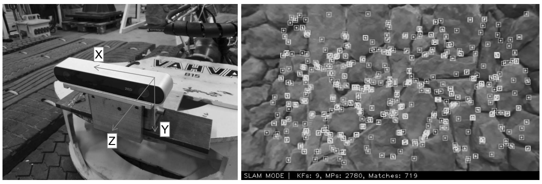


Fig. 2. The left image shows the stereo camera attached to the gripper. The camera's coordinate system is also shown. The right image displays an example view of the detected ORB features during SLAM.

TABLE I
DH Parameters Used for TCP Pose Formulation.

No.	Joint	α_i	a_i	θ_i	d_i
1.	Rotation	$\pi/2$	a_1	θ_1	d_1
2.	Lift	0	a_2	θ_2	0
3.	Tilt	$\pi/2$	a_1	$\theta_3 + \pi/2$	d_3
4.	Wrist 1	$\pi/2$	0	θ_4	d_4
5.	Wrist 2	$-\pi/2$	0	θ_5	0
6.	Wrist 3	0	0	θ_6	d_6

Table I presents the Denavit-Hartenberg (DH) parameters of the manipulator, which comprised an anthropomorphic arm with a spherical wrist [19]. The exact parameters were not used, as they were not available from the manufacturer. Instead, the parameters were self-measured, and are presented only symbolically here. The forward kinematic relationship between the base and the analytical TCP of the manipulator (see Fig. 1) was then formulated as follows:

$$^{1}\mathbf{T}_6 = \mathbf{T}_1\mathbf{T}_2\mathbf{T}_3\mathbf{T}_4\mathbf{T}_5\mathbf{T}_6, \tag{1}$$

where $^{1}\mathbf{T}_6$ denotes the transformation matrix between the base frame and the TCP frame, and $\mathbf{T}_i$, $i \in \{1, ..., 6\}$, was formulated using the following general equation by substituting the DH parameters of the ith joint:

$$\mathbf{T}_i = \begin{bmatrix} c\theta_i & -s\theta_i c\alpha_i & s\theta_i s\alpha_i & a_i c\theta_i \\ s\theta_i & c\theta_i c\alpha_i & -c\theta_i s\alpha_i & a_i s\theta_i \\ 0 & s\alpha_i & c\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where $\cos$ and $\sin$ are abbreviated as c and s, respectively. The ground-truth TCP pose was then extracted from the $^{1}\mathbf{T}_6$ transformation matrix.

IV. Data Analysis

A. Extrinsic Calibration

As visual SLAM estimates the pose with respect to the camera frame, calibration between the camera frame and the ground-truth TCP frame is required to obtain comparable results. In more detail, the transformation matrix between the camera's coordinate system (or frame) and the coordinate system of the ground-truth TCP must be known, as the two are not inherently aligned. This results from the ambiguous camera attachment and the camera model. For the initial experiments presented in this paper, we used only recorded data, which

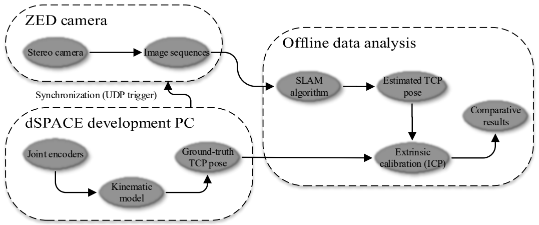


Fig. 3. A diagram illustrating how the comparative results were obtained from the measurements.

permitted the use of the ICP algorithm [20]. In essence, the SLAM pose trajectory is modified into a point cloud, which is then matched to the respective ground-truth pose trajectory point cloud by using an ICP search.

However, a weakness of the ICP is that it can provide an erroneous fitting, while the mean error between the two matched point clouds appears small. In attempt to avoid such a scenario, the camera frame was first modified so that the positive direction of each axis corresponded to that of the ground-truth TCP frame. The second step was to use the ICP to find the transformation matrix between the two frames.

B. Comparative Results

The performance of SLAM in estimating the TCP pose was experimented with motions into different directions. The TCP poses are expressed with respect to the base coordinate system (see Fig. 1), and the manipulator was automatically driven to the same initial TCP position before each experiment. The initial poses between experiments had minor variances, because P-control was used when the joints were driven. The procedure for how the results were obtained is further visualized in Fig. 3.

The first three experiments observed the orientation of the TCP with respect to each axis, whereas in the fourth and final experiment a longer trajectory with loops and multiple laps was studied. In the results, black lines always represent ground-truth variables obtained using encoder measurements and forward kinematics, whereas red lines represent calibrated SLAM estimates in each case. Furthermore, orientation is expressed with XYZ Euler angles.

The camera frame was first aligned with the ground-truth TCP frame using the ICP procedure described in the previous subsection. Static biases with respect to the ground-truth initial poses were also removed from the estimates. Table II shows the root mean square errors calculated during the ICP procedures. The errors are very small, implying that the calibrated camera frame should closely match the ground-truth TCP frame so that the poses are comparable.

In the first experiment, motion was applied only to the lift joint (see Fig. 1) so that the TCP rotated about the Y-axis. The resulting 6 DOF TCP pose is shown in Fig. 4, in which the translational motions and the respective orientations in relation to each axis are illustrated. The positional variables and the Y-axis orientation demonstrate good matching with their respective ground-truth measurements. The remaining two orientation estimates from SLAM show larger amplitudes of motion than their corresponding ground-truth measurements.

In the second experiment, motion was applied only to the base rotation. In this case, the camera (and the TCP) moved mainly in the depth direction along the Y-axis and around the Z-axis of the base frame. The outcome is illustrated in Fig. 5. The results are similar to those of the first experiment: The positional variables and the orientation of the main motion axis show good matching with the ground-truth measurements, while the other two orientations display larger motions.

In the third experiment, the goal was to rotate the TCP around the X-axis by moving the second wrist joint. With the present test setup, however, achieving rotational motion purely around the X-axis was not possible due to the wrist's structure. The TCP was also lifted upward before the second wrist joint was moved, which was to allow larger motion while the wall is maintained in the camera's view. The results are illustrated in Fig. 6. As shown, the positional variables and the main orientation match well in this case also. However, there are slight differences in the remaining two orientation variables, with the estimated angles displaying larger amplitudes.

For the fourth experiment, a TCP trajectory with multiple loops and laps was designed so that the loop detection and closing features of the SLAM algorithm could be tested. The complete path is illustrated in Fig. 7, where the black point cloud represents the ground-truth TCP trajectory. It is compared with the red point cloud, which visualizes the calibrated TCP trajectory obtained from SLAM. The first rectangular part of the path in the XZ plane was completed 3 times, after which the TCP was moved closer to the wall along the Y-axis. Then, the second rectangular part of the path in the XZ plane was also completed three times. Finally, the TCP was driven back to the initial position along the Y-axis. As the results in Fig. 7 show, the multiple laps during each loop in the XZ plane are barely visible in the point clouds and the loops are also closed. The respective 6 DOF TCP pose during the measurement is illustrated in Fig. 8. The results are in line with the previous experiments; the positional variables match well with the corresponding ground-truth measurements. The orientations also match relatively well, albeit the estimated angles show larger amplitudes of motion by a few degrees in relation to their ground-truth measurement counterparts.

To sum up the results of the four experiments, ORB-SLAM2 performed surprisingly well in the tested cases. The mean absolute errors of the TCP pose variables in each experiment (1–4) are documented in Table III, where $\gamma_{x,y,z}$ denote the XYZ Euler angles. Respectively, the maximum absolute errors are shown for each case in Table IV. The estimated orientation angles generally demonstrated larger amplitudes than the ground-truth measurements, which is expected to be at least partly attributed to the flexibility of the manipulator. The differences could also have followed from inaccuracies in the DH parameters (namely, the angles) or in the calibration step.

Finally, this work considered only a laboratory setting. In addition, a specifically designed test wall with a relatively textured surface was used. As visual SLAM is completely dependent on what the camera has in its view, the performance is strictly tied to the environment. Thus, real-world measurements from actual mines are required for further experimenting. In this work, we also used a stereo camera to avoid scale ambiguity, which is a well-known issue with

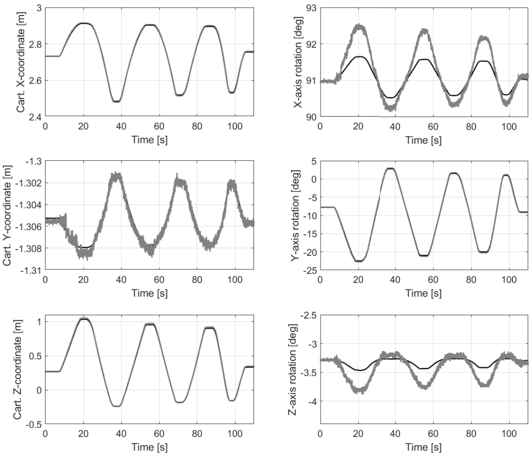


Fig. 4. Results from the first experiment, in which motion was applied only to the lift joint. The black lines denote the ground-truth values obtained with encoders, while the red lines denote the calibrated SLAM estimates.

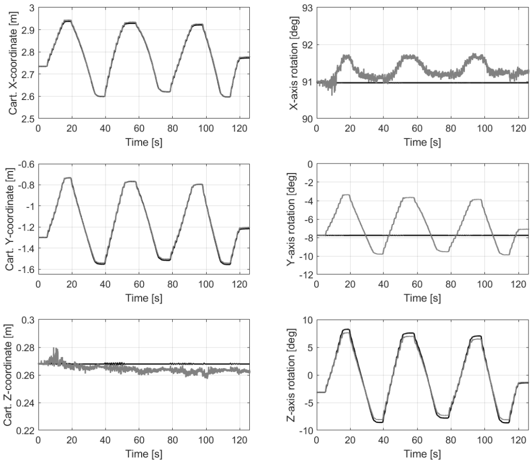


Fig. 5. Results from the second experiment, in which motion was applied only to the base rotation.

TABLE II
ICP Root Mean Square Error in Each Case.

Experiment	1	2	3	4
RMS error	0.0032 (m)	0.0081 (m)	0.0038 (m)	0.0190 (m)

monocular systems. However, stereo implies that the system has a minimum viewing distance required for reliable triangulation of the 3D point features, which is not optimal for confined spaces. A possible solution would be to switch to monocular SLAM when the minimum distance is crossed, while the scale is obtained using stereo data or another sensor.

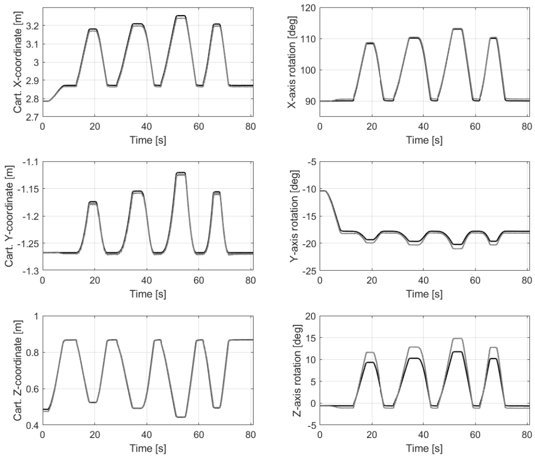


Fig. 6. Results from the third experiment, in which motion was applied mainly to the second wrist joint.

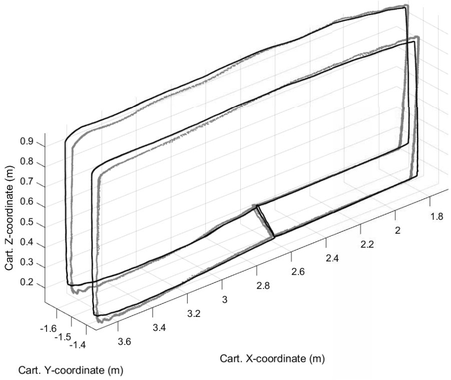


Fig. 7. In the fourth experiment, a longer TCP trajectory was experimented with. The first rectangular part of the path in the XZ plane was completed three times, after which the TCP was driven closer to the wall along the Y-axis. Then, the second rectangular part of the path in the XZ plane was also completed three times. Finally, the TCP was driven to the initial pose along the Y-axis. The black point cloud illustrates the ground-truth poses obtained with encoders, while the red point cloud illustrates the calibrated SLAM poses.

TABLE III
Mean Absolute Errors in Each Measurement.

	Fig. 4	Fig. 5	Fig. 6	Fig. 8
x (m)	0.0032	0.0033	0.0082	0.0338
y (m)	0.0003	0.0073	0.0027	0.0056
z (m)	0.0089	0.0038	0.0032	0.0133
γ_x (deg)	0.2746	0.3352	0.4306	0.8988
γ_y (deg)	0.1565	1.8851	0.4040	0.7108
γ_z (deg)	0.1312	0.3589	1.0950	0.4824

V. Conclusion

In this work, we studied the application and feasibility of SLAM for estimating the TCP pose of a large-scale manipulator in a confined, unknown environment. The SLAM

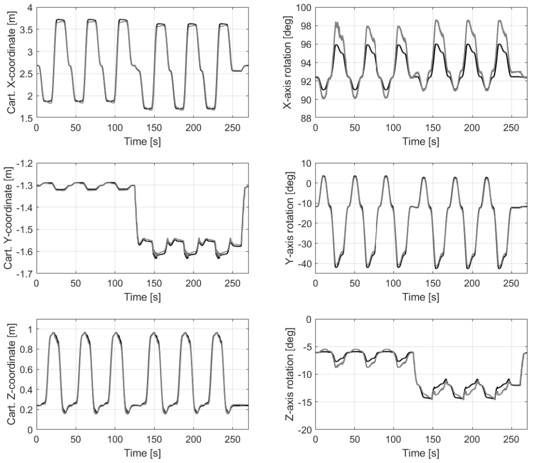


Fig. 8. Results from the fourth experiment.

TABLE IV
MAXIMUM ABSOLUTE ERRORS IN EACH MEASUREMENT.

	Fig. 4	Fig. 5	Fig. 6	Fig. 8
x (m)	0.077	0.0098	0.0203	0.1317
y (m)	0.0015	0.0246	0.0097	0.0229
z (m)	0.0239	0.0116	0.0129	0.0492
γ_x (deg)	0.9166	0.8064	0.9014	2.6606
γ_y (deg)	0.3451	4.4364	0.8373	2.1485
γ_z (deg)	0.4268	0.9145	3.0922	1.4357

algorithm is a key part of the proposed application, in which accuracy, robustness, and real-time performance are all highly important. In the initial results presented in this paper, we were mainly concerned about the potential accuracy. We found that ORB-SLAM2 provided a relatively good performance in the offline data analyses, in which we used an ICP algorithm to extrinsically calibrate the camera. Based on previous research, ORB-SLAM2 is also directly applicable in real time.

Regarding the calibration procedure, the results benefit from running an automatic calibration algorithm by applying ICP to the outputs of each individual experiment. This allowed the effective fine-tuning of the extrinsic calibration parameters of the camera for each test by comparison with the ground truth. In the case of online tests and use of SLAM feedback for control purposes without the ground truth, further investigation of the system calibration is required for online estimation of the extrinsic parameters.

Although the results were obtained specifically with ORB-SLAM2, theoretically the SLAM algorithm itself should not matter as long as the 6 DOF TCP pose can be reliably estimated. To fully localize the TCP of a manipulator with respect to the machine it is attached to, SLAM by itself is not sufficient, as it estimates only the motion of the camera (or the TCP) frame. In this work, the relationship between the frame and the base frame of the manipulator was obtained from ground-truth joint encoder measurements. For future studies, the goal is to omit these sensors completely by developing an alternative method for formulating this correspondence.

REFERENCES

[1] A. H. Kashani, W. S. Owen, N. Himmelman, P. D. Lawrence, and R. A. Hall, “Laser scanner-based end-effector tracking and joint variable extraction for heavy machinery,” *The International Journal of Robotics Research*, vol. 29, no. 10, pp. 1338–1352, 2010.

[2] H. Hyyti, V. V. Lehtola, and A. Visala, “Forestry crane posture estimation with a two-dimensional laser scanner,” *Journal of Field Robotics*, vol. 35, no. 7, pp. 1025–1049, 2018.

[3] M. M. Soltani, Z. Zhu, and A. Hammad, “Skeleton estimation of excavator by detecting its parts,” *Automation in Construction*, vol. 82, pp. 1–15, 2017.

[4] P. Cheng, B. Oelmann, and F. Linnarsson, “A local positioning system for loader cranes based on wireless sensors—a feasibility study,” *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 8, pp. 2881–2893, 2011.

[5] J. Vihonen, J. Mattila, and A. Visa, “Joint-space kinematic model for gravity-referenced joint angle estimation of heavy-duty manipulators,” *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 12, pp. 3280–3288, 2017.

[6] C. Zhang, A. Hammad, and S. Rodriguez, “Crane pose estimation using UWB real-time location system,” *Journal of Computing in Civil Engineering*, vol. 26, no. 5, pp. 625–637, 2011.

[7] K. M. Lundeen, S. Dong, N. Fredricks, M. Akula, J. Seo, and V. R. Kamat, “Optical marker-based end effector pose estimation for articulated excavators,” *Automation in Construction*, vol. 65, pp. 51–64, 2016.

[8] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[9] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, “Visual simultaneous localization and mapping: a survey,” *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[10] C. Kerl, J. Sturm, and D. Cremers, “Dense visual SLAM for RGB-D cameras,” in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2100–2106.

[11] C. Forster, M. Pizzoli, and D. Scaramuzza, “SVO: Fast semi-direct monocular visual odometry,” in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 15–22.

[12] M. Labbé and F. Michaud, “RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation,” *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[13] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, “ORB: An efficient alternative to SIFT or SURF.” in *ICCV*, vol. 11, no. 1. Citeseer, 2011, p. 2.

[14] R. Gomez-Ojeda, F.-A. Moreno, D. Zuñiga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, “Pl-slam: a stereo slam system through the combination of points and line segments,” *IEEE Transactions on Robotics*, 2019.

[15] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[16] R. Mur-Artal and J. D. Tardós, “Visual-inertial monocular slam with map reuse,” *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 796–803, 2017.

[17] D. Gálvez-López and J. D. Tardós, “Bags of binary words for fast place recognition in image sequences,” *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[18] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, “g2o: A general framework for graph optimization,” in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 3607–3613.

[19] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: modelling, planning and control*. Springer Science & Business Media, 2010.

[20] P. J. Besl and N. D. McKay, “Method for registration of 3-D shapes,” in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–607.

PUBLICATION II

Redundancy-based visual tool center point pose estimation for long-reach manipulators

P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila

In *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*,
2020, pp. 1387–1393
DOI: 10.1109/AIM43001.2020.9159022

Redundancy-Based Visual Tool Center Point Pose Estimation for Long-Reach Manipulators

Petri Mäkinen, Pauli Mustalahti, Sirpa Launis and Jouni Mattila

***Abstract*— In this paper, we study a visual sensing scheme for 6 degree-of-freedom (DOF) tool center point (TCP) pose estimation of large-scale, long-reach manipulators. A sensor system is proposed, designed especially for mining manipulators, comprising a stereo camera running a simultaneous localization and mapping (SLAM) algorithm near the TCP and multiple cameras that track a fiducial marker attached near the stereo camera. In essence, the TCP pose is formulated using two different routes in a co-operative (eye-in-hand/eye-to-hand) manner using data fusion, with the goal of increasing the system's fault tolerance and robustness via sensor redundancy. The system is studied in offline data analysis based on real-world measurements recorded using a hydraulic 6 DOF robotic manipulator with a 5 m reach. The SLAM pose trajectory is obtained using the open source ORB-SLAM2 Stereo algorithm, whereas marker-based tracking is realized with a high-end motion capture system. For reference measurements, the pose trajectory is also formulated using joint encoders and a kinematic model of the manipulator. Results of the 6 DOF pose estimation using the proposed sensor system are presented, with future work and key challenges also highlighted.**

I. Introduction

A. Motivation

Mobile working machines represent a significant field in industry, and they come in many different configurations and sizes with respect to their on-board manipulators. In machines designed specifically for mining and construction, the reach of these manipulators can range from approximately 10–15 m in 6 degrees-of-freedom (DOF) tunneling machines to only 1–2 m in small surface drilling platforms. The annual production volume for a specialized machine type can be a few hundred units, while the volume for some production variants can be as low as 1–10 units per year. Therefore, these high-precision, low-volume robotic manipulators call for innovative sensor system solutions that reduce the manufacturing, assembly, and maintenance costs of these machines. The current solution is to fit each joint of a manipulator with a joint sensor, which also requires additional protective housing, mechanical couplings, and cabling that are suitable for the given machine type. Therefore, many components are required to fit all the machine types with mechanical precision sensing, which also results in an overall high cost in the terms of the bill of materials (BOM). The underlying goal of the present research is that all types of mining machines are equipped with a standardized sensor system that is of low cost, easy to install, and scalable to fulfill all requirements across the range of machine types.

Each manipulator should have a sensor system because the 6 DOF tool center point (TCP) pose of the manipulator must be known. In mining machines, knowledge of the joint states and the TCP pose is currently required to carry out automated and semi-automated operations, as a high production rate is very valuable. Due to this, tunneling jumbos, for example, can have up to four drilling booms attached to the same machine for parallel operations. Overall, the goal is to move toward fully automated operations in these hazardous applications, as discussed in, for example, [1]. From a practical point of view, the important factor is being able to accurately measure and control the TCP pose that is expressed in Cartesian work space with respect to a world frame. For this purpose, methods other than the traditional kinematic chain formulation with joint sensors could also be developed. Due to the long reach and high payload-to-own-weight ratios of these manipulators, the traditional method based on a serial kinematic chain structure will always impose significant errors at the end of the chain (the TCP) due to structural flexibilities and calibration uncertainties. Thus, driving these manipulators with external sensor systems, in GPS-denied environments, is of great interest.

Compared to traditional industrial robots, large-scale, long-reach manipulators are often under the radar in research. Whereas an industrial (stationary) robot has a relatively low payload-to-own-weight ratio and precision sensors at each joint providing the manufacturer-quaranteed absolute accuracy and repeatability for the TCP pose in Cartesian space, large-scale manipulators have much higher payload-to-own-weight ratios (e.g., one), with many applications still operated manually, as no sensors are installed due to the harsh working conditions and structural flexibilities that distort the results if basic rigid-body kinematics are applied. This situation is changing, however, as the automation level of these manipulators is increasing, thus requiring sensors. In forestry machines, for example, inertial sensors have been recently introduced commercially to measure the joint angles that are sensitive to gravity, making it possible for the operator to control the TCP directly, instead of controlling each individual joint of the manipulator. A method for computing gravity-sensitive angles using inertial sensors was introduced in [2].

In mining machines, sensors have long been present to measure the joint states, as in this application the TCP pose is required so that drilling plans can be effectively completed.

Petri Mäkinen, Pauli Mustalahti and Jouni Mattila are with the Faculty of Engineering and Natural Sciences, Tampere University, Tampere, Finland. Emails: {petri.makinen,pauli.mustalahti,jouni.mattila}@tuni.fi

Sirpa Launis is with Sandvik Mining and Construction Oy, Tampere, Finland. Email: sirpa.launis@sandvik.com

For example, an orientation error of 5° at the TCP, with a drilling depth of 4 m, will result in a position error of 35 cm at the end of a drill hole. The accuracy of the drill holes with respect to the drilling plan is crucial. In tunneling, inaccurate drilling results in more drilled meters required, along with more blastings required, which slows progress and increases operation costs. Respectively, in long-hole drilling, inaccurate drilling can lead to ore-loss or increased dilution (waste rock) of the product. Overall, straight drill holes result in a better total economy. As rough target values for accuracy in these applications, the positioning error at the TCP should be less than 1 cm and the orientation error less than 1°, respectively.

This paper is the new step after our previous research [3], in which pure visual simultaneous localization and mapping (SLAM)-based TCP pose estimation was studied. In this paper, we extend toward a more complete sensor system concept for the described application. Namely, marker-based TCP pose tracking is combined with the SLAM module in an attempt to obtain a more robust pose estimation. In essence, this corresponds to the so-called eye-in-hand/eye-to-hand co-operation, which is a method used for visual servoing, see e.g. [4], [5]. In this work, the goal of the proposed solution is to increase the system's fault tolerance in the sense of sensor redundancy, while having both measurements (marker tracking and SLAM) available complement each other after data fusion. For marker tracking, we used a commercial OptiTrack motion capture system, which conveniently offered the required functionality for measurements, such as calibration and multi-camera tracking. Although such a high-end system in commercial mining machines is unrealistic, lower-end cameras are becoming more affordable and advanced, not only the hardware but also software. Consequently, multi-camera solutions with redundancy are becoming more viable in cost-sensitive industry applications, and this paper is a step toward this path. For SLAM, we used a Stereolabs ZED camera along with the open source ORB-SLAM2[1] Stereo algorithm [6]. A test case using a laboratory-installed 6 DOF hydraulic crane was designed for a simple practical experiment to study the feasibility and challenges in realizing the conceptual sensor system at full scale. The results of the offline data analysis show that the main challenges lie in the system's calibration (for precise measurements) and in control design. Further practical issues, such as model development for kinematic calibration of flexible robots, are beyond the scope of this study.

B. Brief Literature Review

Due to the harsh and highly varying environmental conditions that large-scale mobile manipulators are exposed to, a wide variety of sensor technologies have been explored. For example, in [7] battery-powered wireless sensors were applied for local positioning of a loader crane. Inertial sensors were used to measure joint angles, and an ultrasound time-of-flight sensor was used to measure the length of a telescopic extension boom. In [8], a laser scanner was used with a customized iterative closest point algorithm to estimate the joint angles.

A marker-based pose estimation method for articulated excavators was presented in [9]. In this case, the camera was installed in the surrounding environment, and several challenges were brought up, such as occlusion and lighting. Thus, a marker-based system is foreseen to work best as an auxiliary sensing method. A marker-less method for the same problem was later presented in [10], in which a deep convolutional network human pose algorithm was used. These studies have in common that they focused on articulated manipulators that have joints only in the vertical plane. Mining manipulators, however, typically also have at least two joints in the horizontal plane, which complicates the pose estimation problem significantly.

The rest of the paper is organized as follows: Section II describes the proposed sensor system, whereas Section III details the experimental setup for measurements. It is followed by Section IV, which contains data analysis and results. Finally, Section V concludes the paper.

II. Conceptual Sensor System for Mining Manipulators

Inertial sensors cannot be effectively utilized in mining manipulators due to the presence of several horizontal joints that are insensitive to gravity. As for visual sensing, such systems are already utilized for collision avoidance [11]. The sensors are installed near the roof of the cabin of a machine. Moreover, unlike with articulated cranes found in, for example, excavators, in mining manipulators the base of the manipulator is typically lower with respect to the cabin, as shown in Fig. 1, giving natural elevation to visual sensors installed near the roof. This results in fewer occlusions that would result due to a part of the manipulator blocking the TCP. However, as discussed in [9], pure marker-based systems can be problematic to realize due to several reasons, which suggests that additional sensors are required for increasing accuracy and robustness.

In the previous study [3] we used SLAM to estimate the TCP pose with good initial results. The idea was that the surrounding mine walls provide enough features, and that the operations conducted by these mining and construction machines are controlled enough for SLAM to be viable. It is also perceived that due to the length of these manipulators (> 10 m), a sensor located near the TCP is required to obtain precise measurements, which is supported by the results of this paper. Thus, a sensor system resembling the eye-in-hand/eye-to-hand co-operative scheme is studied in this work. The SLAM module, located near the TCP of a manipulator, can provide more precision in pose estimation. The marker-based tracking module will be less accurate due to the increased viewing distances, and thus, increased uncertainties, but it will have a better view of the entire scene, which could also be used for calibration and other assistive operations, for example.

[1] https://github.com/raulmur/ORB_SLAM2

Fig. 1. A Sandvik tunneling jumbo with two drilling booms.

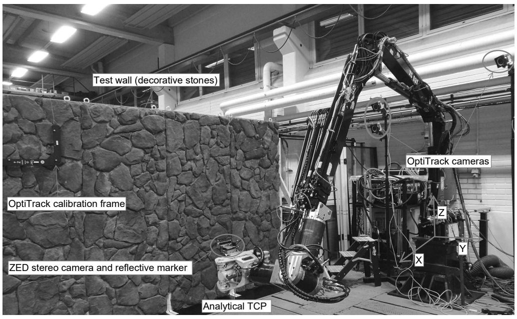

Test wall (decorative stones)
OptiTrack cameras
OptiTrack calibration frame
Z
Y
X
ZED stereo camera and reflective marker
Analytical TCP

Fig. 2. The experimental measurement setup. The goal was to track the TCP pose of the manipulator by using i) SLAM (ZED stereo camera), in which the tracked features were obtained from the test wall, and ii) marker-based tracking, for which a high-end OptiTrack motion capture system was used. For reference measurements, an analytical TCP of the manipulator was also formulated based on forward kinematics and joint encoder measurements. The world coordinate system is also shown.

III. Experimental Setup

To study an application with the proposed sensor system architecture and obtain initial results, a simple use case test bed was designed. The system comprised the following main components:

- A laboratory-installed hydraulic crane with accurate reference sensors and a dSPACE real-time control system.
- A Stereolabs ZED camera for SLAM, along with a textured test wall for feature extraction.
- An OptiTrack motion capture system for marker-based tracking.

The components are detailed further in the next subsections.

A. HIAB033 Hydraulic Crane with Additional 3 DOF Wrist

The target system was a hydraulic lorry crane, HIAB 033, which was located at the heavy laboratory of the Innovative Hydraulics and Automation research unit at Tampere University. The setup is presented in Fig. 2. The manipulator itself had 3 active DOF (rotation, lifting, and tilting). A spherical wrist was also attached at the tip of the structure, adding another 3 DOF to the system. Each of the six active joints was instrumented with an incremental encoder to obtain precision measurements of the joint states.

B. SLAM Module

To estimate the TCP pose with SLAM, a Stereolabs ZED camera was installed near the tip of the manipulator. Grayscale images were captured at 24 ms intervals with a resolution of 672×376 per lens and saved for offline data analysis. As for the SLAM method, we utilized the open source ORB-SLAM2 Stereo algorithm. For the textured environment, from which the feature points for SLAM were to be obtained, a 2.5×4 m test wall was constructed using decorative stones. The underlying goal was to simulate a rock wall, as the target application of this research was underground mining and construction.

C. Marker-Based Tracking Module

As this study was mainly concerned about a conceptual sensor system, we used the most powerful systems available to us. In this case, we used a commercial-off-the-shelf motion capture system to realize high-performance marker tracking. Three OptiTrack Prime 17W cameras were placed around the base pillar of the manipulator: The idea is that the cameras used for marker-based tracking are installed on top of the cabin of a machine. An infrared-reflective (passive) marker was then placed near the tip of the manipulator (next to the ZED camera). Using OptiTrack's Motive motion capture software, the marker's pose was tracked with reference to an OptiTrack L-frame, which was placed next to the test wall and in view of the cameras. With three cameras, the system was able to effectively track the marker, although the boom temporarily occluded the view of the third camera during the measurements.

D. Data Flow

A general depiction of the data flow during the measurements is presented in Fig. 3. The real-time control system of the manipulator was a dSPACE DS1005 PPC controller board, which used a 2 ms sampling period and recorded the encoder and motion capture measurements. The OptiTrack cameras were read using a dedicated laptop running Motive software, from which the measured poses were transmitted at a high frequency to the dSPACE development PC by using Matlab and UDP. The ZED stereo image capture was also realized with a dedicated laptop, which was synchronized with the dSPACE development PC by sending a UDP trigger signal from dSPACE to the dedicated laptop at a time interval of 12×2 ms. The synchronized image sequences were then recorded on the dedicated laptop by using Matlab and the ZED SDK.

E. Ground-Truth TCP Pose

To have a reference pose for the camera measurements, a kinematic model of the manipulator was formulated. This was then used with the encoder measurements to produce the TCP pose based on the model.

Remark 1: Compared to our previous study [3], the Denavit-Hartenberg (DH) parameters of the manipulator were kinematically calibrated using a Sokkia NET05 total

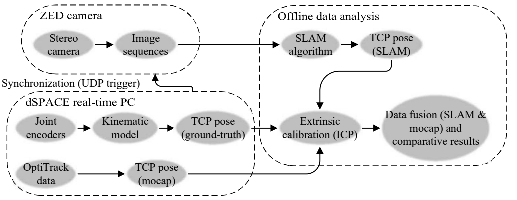


Fig. 3. A general depiction of how the data was obtained and used for offline data analysis.

TABLE I

DH Parameters Of The Manipulator. The System Comprised An Anthropomorphic Arm In Combination With A Spherical Wrist.

Joint	α_i	a_i	θ_i	d_i
Rotation	$\pi/2$	a_1	θ_1	d_1
Lift	0	a_2	θ_2	0
Tilt	$\pi/2$	a_1	$\theta_3 + \pi/2$	d_3
Wrist 1	$\pi/2$	0	θ_4	d_4
Wrist 2	$-\pi/2$	0	θ_5	0
Wrist 3	0	0	θ_6	d_6

station. The resulting Cartesian average error in the calibration was reportedly less than 4 cm.

The symbolic DH parameters are presented in Table I. The forward kinematic relationship between the base and the analytical TCP of the manipulator (see Fig. 2) was then formulated as follows:

$$^{B}\mathbf{T}_{tcp} = \mathbf{T}_r \mathbf{T}_l \mathbf{T}_t \mathbf{T}_{w_1} \mathbf{T}_{w_2} \mathbf{T}_{w_3} \tag{1}$$

where the transformation matrix from the base to the TCP is denoted by $^{B}\mathbf{T}_{tcp}$. Joint transformation matrices $\mathbf{T}_i$, $i \in \{r, l, t, w_1, w_2, w_3\}$ are obtained using the following equation by substituting the respective DH parameters for each joint:

$$\mathbf{T}_i = \begin{bmatrix} c\theta_i & -s\theta_i c\alpha_i & s\theta_i s\alpha_i & a_i c\theta_i \\ s\theta_i & c\theta_i c\alpha_i & -c\theta_i s\alpha_i & a_i s\theta_i \\ 0 & s\alpha_i & c\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where $s = sin$ and $c = cos$. Then, the TCP pose of the manipulator is obtained from $^{B}\mathbf{T}_{tcp}$.

F. Data Fusion of the TCP Pose Estimates

The logic for obtaining the TCP pose estimates based on the available signals was designed as shown in Fig. 4. In the scope of this work, it is assumed that each sensor is either fully operational or not operational (binary), as self-diagnostic systems would be required for more advanced signal analysis. In the event that both TCP pose estimates are available, a data fusion method, confidence-weighted averaging [12], is adopted. This simple, model-free method fuses measurements based on the estimated variance of the measurement error. The advantage is that, assuming that the errors between the sensors are independent and that the expected error equals zero, the variance of the fused output is minimized. The fused 6 DOF TCP pose vector X_{fused} is obtained as follows:

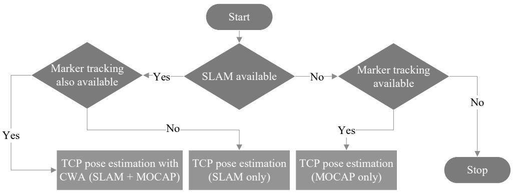

Start
SLAM available
Yes
No
Marker tracking also available
Marker tracking available
Yes
No
No
Yes
TCP pose estimation with CWA (SLAM + MOCAP)
TCP pose estimation (SLAM only)
TCP pose estimation (MOCAP only)
Stop

Fig. 4. A chart illustrating the logic behind utilizing the two TCP pose estimates. In the ideal case that SLAM and marker tracking (mocap) are available, confidence-weighted averaging (CWA) is used for data fusion.

$$X_{fused} = \sum_{i=1}^{N} W_i X_i \tag{3}$$

where N denotes the total number of observations (in this case, SLAM and marker-based tracking), W_i is the weight vector of the ith observation, and X_i is the 6 DOF TCP pose vector of the ith observation. The weights are computed based on the signal variances as follows:

$$W_i = \frac{1}{\sigma_i^2} \sum_{j=1}^{N} \frac{1}{\sigma_j^2} \tag{4}$$

where $\sigma_{i,j}^2$ denotes the variance of a given signal.

Furthermore, if no TCP pose estimate is available, then the manipulator is halted. Matlab Simulink's Stateflow environment was utilized in the experiments.

IV. Data Analysis

A. Calibration of the TCP Frame Correspondences

To obtain comparable results, calibration between the three TCP frames (coordinate systems) is required. The SLAM frame and the marker frame are to be transformed into the analytical TCP frame, which served as the reference. For this purpose, the iterative closest point (ICP) method [13] was employed, which is suitable for offline experiments because the entire pose trajectories from different sources can be matched.

B. Signal Conditioning

The measured TCP orientations using marker-based tracking were conditioned with a geometric moving average (GMA) filter [14] due to noisy data. The equation is given as follows:

$$S_j = (1 - \alpha) S_{j-1} + \alpha s_j, \quad j > 0 \tag{5}$$

where S_j denotes the geometric moving average (conditioned signal) at time j, s_j denotes the unconditioned signal at time j, and $0 < \alpha \leq 1$ denotes the weight coefficient. Note that S_0 is set to the initial value of a given signal. Furthermore, $\alpha = 0.02$ was used.

C. Comparison of Pose Trajectories

First, a test trajectory was designed for the manipulator by using quintic path planning [15]. A rectangular-shaped trajectory was completed three times, after which the TCP was moved closer to the test wall. Then, the rectangle was completed three times again. Finally, the TCP was moved back to the initial position. The trajectories are shown in Fig. 5, which illustrates the three TCP pose trajectories in Cartesian space after point cloud matching using the ICP. The black point cloud represents the analytical reference trajectory, the red point cloud represents the SLAM output poses, and the blue point cloud is associated with the trajectory of the tracked marker. The respective root mean square errors resulting from the ICP matching algorithm are presented in Table II.

The 6 DOF TCP pose estimates are presented based on the chart in Fig. 4. First, only the SLAM pose estimates were used, with the resulting 6 DOF poses shown in Fig. 6. Respectively, the 6 DOF poses from marker-based tracking are shown in Fig. 7. The CWA-fused 6 DOF pose estimates are shown in Fig. 8. The visual measurements in each case were compared with the reference encoder data, with the mean and maximum absolute errors documented in Table IV and Table V. Red lines are associated with the SLAM poses, blue lines denote the marker-based tracking, and black lines represent the encoder computed data. As shown by the measured results, the Cartesian position variables track relatively well over the entire test trajectory, with the mean errors ranging from less than a millimeter to a few centimeters. The orientation variables, however, show less consistent behavior as the amplitudes seize to match well after the TCP is driven closer to the wall. It is suspected this followed from calibration errors, as the uncertainties present in the reference encoder setup and the calibration were quite significant. Furthermore, the visual sensors provided similar behavior, with especially the OptiTrack system being perceived as capable of highly accurate measurements. This emphasizes the challenge of obtaining an accurate 6 DOF pose reference in large-scale, long-reach manipulators.

The two ICP calibrated optical measurements were also compared with each other, see, Fig. 9. Here, the unconditioned marker orientations are also shown with light-blue lines. The results demonstrate a strong correspondence between the SLAM poses and the marker poses. The mean and maximum absolute errors between the optical measurements were also documented in Tables IV-V.

It is evident that the visual estimates of the TCP pose differ from the analytical TCP based on the kinematic model, which suggests that transitioning from the optical measurements to the joint space of the manipulator will be challenging. Thus, alternative, external methods of controlling these manipulators, instead of using the numerical serial kinematic chain structure, should be explored. For the offline data analysis, the weights of the data fusion were obtained using the variances computed over the entire test trajectory by using the encoder reference measurements as the ground-truth values. The weights used are presented in Table III. As it shows, the marker-based tracking was emphasized in fusing the positions which, in this case, was logical due to the reduced ICP calibration error. The orientations were weighted quite evenly. However, the orientations from the marker-based tracking module were GMA-filtered before weights were computed. Consequently, the resulting data fusion is, in this case, optimal in the sense that the fused variance was minimized. However, for future online experiments, a method for determining the weights in real-time is required.

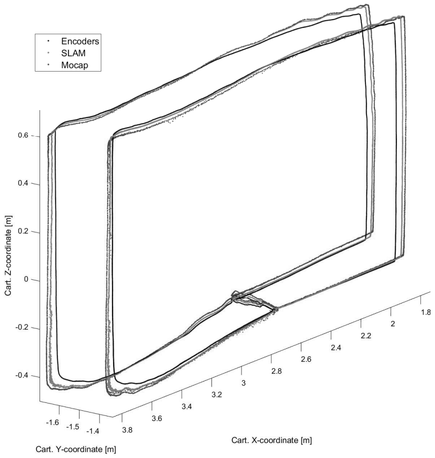


Fig. 5. The pose trajectories after ICP registration and calibration. A rectangular trajectory was first completed three times, after which the TCP moved closer to the test wall and completed another three laps on a rectangular trajectory. Finally, the TCP was moved back into the initial position. The black point cloud represents the analytical TCP, the red point cloud represents the SLAM TCP, and the blue point cloud represents the tracked marker TCP.

TABLE II

ROOT MEAN SQUARE ERRORS RESULTING FROM THE ICP ALGORITHM.

Coord. transf.	SLAM→Analytical TCP	Marker→Analytical TCP
RMS error	0.0371 [m]	0.0253 [m]

TABLE III

WEIGHTS USED IN THE CWA DATA FUSION.

W_i	x	y	z	γ_x	γ_y	γ_z
SLAM	0.2828	0.4728	0.1311	0.4767	0.6613	0.4948
Mocap	0.7172	0.5272	0.8689	0.5233	0.3387	0.5052

V. DISCUSSION AND CONCLUSION

This work presented a new sensor system concept designed especially for large-scale, long-reach mining and construction manipulators used underground, in which an eye-in-hand/eye-to-hand co-operative scheme is utilized by

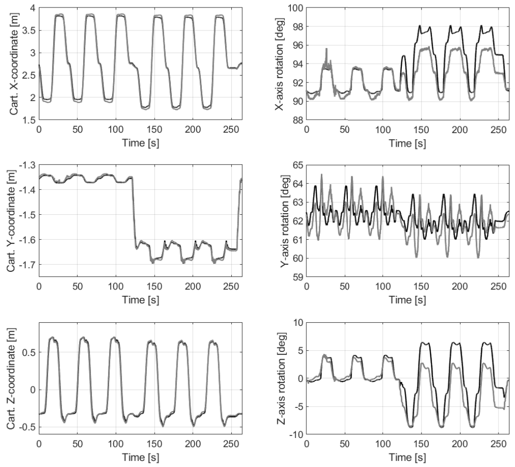


Fig. 6. The estimated 6 DOF TCP pose variables, when only SLAM is available. The black lines denote the reference values using encoder measurements, and the red lines denote the SLAM pose variables.

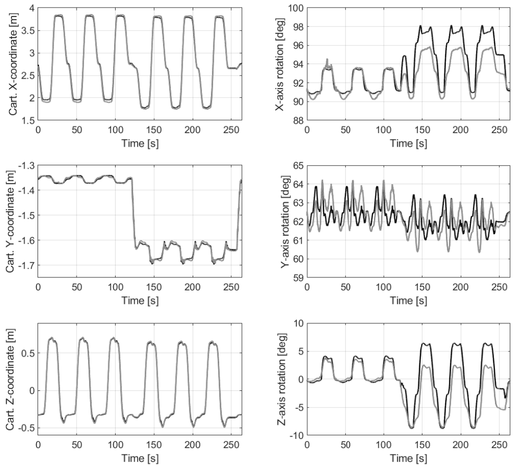


Fig. 8. The estimated 6 DOF TCP pose variables, with signals from SLAM and marker tracking fused with CWA. The black lines denote the reference values using encoder measurements, and the magenta lines denote the pose variables after data fusion.

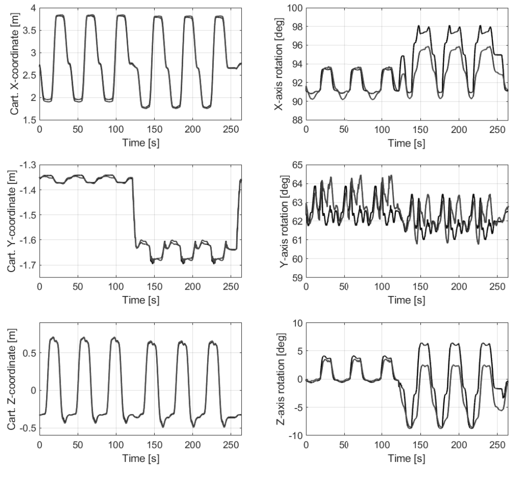


Fig. 7. The estimated 6 DOF TCP pose variables, when only marker-based tracking is available. The black lines denote the reference values using encoder measurements, and the blue lines denote the values from marker tracking, with GMA-filtered orientations.

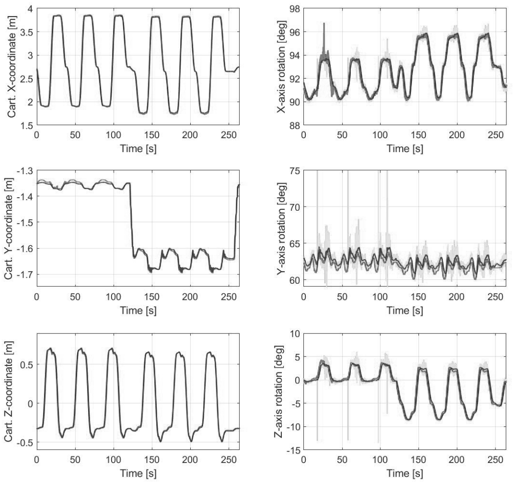


Fig. 9. Comparison of the optically measured 6 DOF TCP poses (with ICP calibration). The red lines are associated with SLAM, the blue lines represent the marker-based tracking, and the light blue orientations denote the unconditioned signals.

combining marker-based tracking with SLAM pose estimation. The test case using a 6 DOF hydraulic manipulator, with a reach of approximately 5 m, assumed equal availability (same frequency) of the SLAM and marker poses. However, it was shown that the SLAM camera, located near the TCP, provided higher quality orientation measurements in relation to the marker-based orientation measurements. In reality, it is foreseen that the SLAM module is required to do the majority of the work in the TCP pose estimation, as the distances between the TCP and the base of the manipulator are quite large in actual mining manipulators, which will degrade the accuracy of any marker-based tracking system. In addition, occlusions will be a challenge in mining manipulators that can rotate approximately 360°. Thus, marker-based tracking is likely to be more useful as a secondary sensor module, which can be realized, for example, with the CWA data fusion method by tuning the weights appropriately. The utilization of marker-based tracking for calibration, and for example, condition monitoring, should be explored in the

TABLE IV
Mean Absolute Errors In Each Case.

	Fig. 6	Fig. 7	Fig. 8	Fig. 9
x [m]	0.0415	0.0262	0.0292	0.0216
y [m]	0.0052	0.0046	0.0044	0.0046
z [m]	0.0210	0.0081	0.0095	0.0142
γ_x [deg]	1.1221	1.0961	1.0924	0.2775
γ_y [deg]	0.6462	0.6429	0.6079	0.6423
γ_z [deg]	1.8541	1.8277	1.8171	0.6104

TABLE V
Maximum Absolute Errors In Each Case.

	Fig. 6	Fig. 7	Fig. 8	Fig. 9
x [m]	0.1195	0.0615	0.0619	0.1457
y [m]	0.0288	0.0163	0.0214	0.0174
z [m]	0.0894	0.0318	0.0347	0.0815
γ_x [deg]	3.4733	3.5547	3.4241	3.1852
γ_y [deg]	1.8698	2.1864	1.3959	1.9344
γ_z [deg]	5.3777	7.1867	6.0092	2.7358

future.

Although the two visual sensor modules produced seemingly high performance, the challenge lies in the numerous uncertainties present in the system. These follow especially from the calibration that, even in the case of offline data analysis, resulted in relatively considerable errors. Furthermore, a new calibration method is required for future online experiments. It is also perceived that the encoder setup used for reference measurements may provide the least accurate TCP pose measurement. Being able to match the visual pose measurements to the analytical TCP pose is desirable in the sense that to control the manipulator, knowledge of the joint states is usually required, which could be achieved by using an inverse kinematic model. In addition, although the applied kinematic model is based on the rigidity assumption, these types of long-reach manipulators are very flexible due to their length and high payload-to-own-weight ratio. This flexibility and the following non-rigid kinematics are also a central research problem related to the TCP pose estimation in long-reach manipulators, and solving this with external sensors is a long-term goal of this research. Thus, alternative control methods that are not directly based on the joint states should be pursued.

Acknowledgment

This work was supported by the Doctoral School of Industry Innovations (DSII) of Tampere University. This work was also carried out with the support of Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland.

References

[1] P. Corke, J. Roberts, and G. Winstanley, “Vision-based control for mining automation,” *IEEE Robot. Automat. Mag.*, vol. 5, no. 4, pp. 44–49, 1998.

[2] J. Vihonen, J. Mattila, and A. Visa, “Joint-space kinematic model for gravity-referenced joint angle estimation of heavy-duty manipulators,” *IEEE Trans. Instrum. Meas.*, vol. 66, no. 12, pp. 3280–3288, 2017.

[3] P. Mäkinen, M. M. Aref, J. Mattila, and S. Launis, “Application of simultaneous localization and mapping for large-scale manipulators in unknown environments,” in *Cybernetics and Intelligent Systems (CIS) and IEEE Conf. Robotics, Automation and Mechatronics (RAM), 2019 IEEE 9th Inter. Conf.* IEEE, 2019.

[4] G. Flandin, F. Chaumette, and E. Marchand, “Eye-in-hand/eye-to-hand cooperation for visual servoing,” in *Proc. 2000 ICRA. Millennium Conf. IEEE Int. Conf. Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 3. IEEE, 2000, pp. 2741–2746.

[5] V. Lippiello, B. Siciliano, and L. Villani, “Eye-in-hand/eye-to-hand multi-camera visual servoing,” in *Proc. 44th IEEE Conf. Decision and Control.* IEEE, 2005, pp. 5354–5359.

[6] R. Mur-Artal and J. D. Tardós, “ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras,” *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[7] P. Cheng, B. Oelmann, and F. Linnarsson, “A local positioning system for loader cranes based on wireless sensors—a feasibility study,” *IEEE Trans. Instrum. Meas.*, vol. 60, no. 8, pp. 2881–2893, 2011.

[8] A. H. Kashani, W. S. Owen, N. Himmelman, P. D. Lawrence, and R. A. Hall, “Laser scanner-based end-effector tracking and joint variable extraction for heavy machinery,” *Inter. J. Robot. Res.*, vol. 29, no. 10, pp. 1338–1352, 2010.

[9] K. M. Lundeen, S. Dong, N. Fredricks, M. Akula, J. Seo, and V. R. Kamat, “Optical marker-based end effector pose estimation for articulated excavators,” *Autom. Construction*, vol. 65, pp. 51–64, 2016.

[10] C.-J. Liang, K. M. Lundeen, W. McGee, C. C. Menassa, S. Lee, and V. R. Kamat, “A vision-based marker-less pose estimation system for articulated construction robots,” *Autom. Construction*, vol. 104, pp. 80–94, 2019.

[11] T. Kivelä, J. Mattila, J. Puura, and S. Launis, “Redundant robotic manipulator path planning for real-time obstacle and self-collision avoidance,” in *Int. Conf. Robotics in Alpe-Adria Danube Region*. Springer, 2017, pp. 208–216.

[12] W. Elmenreich, “Fusion of continuous-valued sensor measurements using confidence-weighted averaging,” *J. Vib. Control*, vol. 13, no. 9-10, pp. 1303–1312, 2007.

[13] P. J. Besl and N. D. McKay, “Method for registration of 3-D shapes,” in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–607.

[14] S. W. Roberts, “Control chart tests based on geometric moving averages,” *Technometrics*, vol. 42, no. 1, pp. 97–101, 2000.

[15] R. N. Jazar, *Theory of Applied Robotics - Kinematics, Dynamics, and Control.* Dordrecht, the Netherlands: Springer, 2010.

PUBLICATION
III

Probabilistic camera-to-kinematic model calibration for long-reach robotic manipulators in unknown environments

P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila

In *2022 IEEE 17th International Conference on Advanced Motion Control (AMC)*, 2022, pp. 48–55
DOI: 10.1109/AMC51637.2022.9729259

Probabilistic Camera-to-Kinematic Model Calibration for Long-Reach Robotic Manipulators in Unknown Environments

1st Petri Mäkinen
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
petri.makinen@tuni.fi

2nd Pauli Mustalahti
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
pauli.mustalahti@tuni.fi

3rd Sirpa Launis
Rock Technologies and Drilling
Sandvik Mining and Construction Oy
Tampere, Finland
sirpa.launis@sandvik.com

4th Jouni Mattila
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
jouni.mattila@tuni.fi

***Abstract*—In this paper, we present a methodology for extrinsic calibration of a camera attached to a long-reach manipulator in an unknown environment. The methodology comprises coarse frame alignment and fine matching based on probabilistic point set registration. The coarse frame alignment is based on the known initial pose and assists in the fine matching step, which is based on robust generalized point set registration that utilizes position and orientation data. Comparison with other methods utilizing only position data is provided. The first 6 DOF point set is obtained using a SLAM algorithm running on a camera attached near the tip of a manipulator, whereas the second point set is obtained using a kinematic model and joint encoders. Real-time experiments and a use case are presented. The results demonstrate that the proposed methodology is suited for the application, and that it can be useful in operations requiring precise visual measurements obtained near the tip of the manipulator.**

***Index Terms*—robot vision systems, simultaneous localization and mapping, iterative methods**

I. Introduction

Visual sensors, such as different types of cameras and laser scanners, have seen some significant technological advances in hardware and software in the past decades. These types of sensors are able to provide large amounts of information related to the surroundings, which, due to more affordable and increasing processing power, have been widely adopted in numerous applications, especially in the manufacturing industry in controlled factory environments. However, visual sensors are challenging to utilize in harsh working environments, as the sensors often lack the robustness and reliability required in, for example, mobile work machines that do not operate in strictly controlled environments. Despite this problem, the current direction in the heavy-duty mobile machine industry is toward autonomous systems, where an affordable perception system is essential. This calls for new technologies for perception systems that perform in a robust manner under uncertainties arising from inconsistent working environments and the characteristics of robotic manipulators used in mobile machinery, such as structural flexibility and actuator backlash.

The extrinsic camera calibration problem arises when information measured in a camera's coordinate system needs to be expressed with respect to another sensor's coordinate system. For example, mounting a visual sensor on a robotic manipulator and using the sensor data for control purposes requires determining the sensor's position and orientation in relation to the manipulator's coordinate system, typically defined by its set kinematic model and joint encoders. This is known as the eye-in-hand calibration problem. Finding this extrinsic calibration has been examined in, for example, [1] and [2], where three separate methods were presented. However, each method relied on a visible reference object or point, which is problematic to realize outside controlled environments, such as factories employing stationary industrial robots. Few studies exist for large-scale manipulators working in unstructured or unknown environments, where predefined objects for extrinsic calibration are not practical or available.

Point sets, or point clouds, are a common method of processing and visualizing 3D vision data. These point sets can be used for several applications, such as map building, searching for and tracking known objects, or extrinsic camera calibration by utilizing point sets obtained from two sources. In point set registration, the goal is to find the correspondence between a measured point set and a reference point set. The correspondence between the two point sets is described by a transformation comprising rotation and translation components. Many methods exist for point set registration; the most well-known is the iterative closest point (ICP) algorithm [3] and its numerous variants. In [4], an ICP-based method for extrinsic calibration of an eye-in-hand 2D LiDAR sensor in unstructured environments was presented. A small-scale industrial robot was used in the experiments. Other types of more sophisticated algorithms utilizing 3 degrees of freedom (DOF) position data have also been proposed, such as coherent point drift (CPD) [5] that adopts a probabilistic approach using a Gaussian mixture model (GMM). However, in robotic applications, pose data (3 DOF position and 3 DOF orientation) are readily available. Until recently, point set registration methods utilized only 3 DOF position data, which may not be optimal for robotic applications, as half of the available data is not utilized in the registration process. However, a robust generalized point set registration method was proposed in [6], which builds on the CPD algorithm by incorporating orientation data via the von Mises-Fisher mixture model (FMM) [7]. The resulting hybrid mixture model (HMM) comprises a GMM for position data and an FMM for orientation data, which is perceived as useful in robotic applications especially due to the availability of 6 DOF pose data.

This paper is a continuation of our previous research [8], [9], in which new visual sensor system solutions were investigated for long-reach robotic manipulators in unknown environments, especially underground. In this paper, we focus on the development of a robust, generalized methodology for on-site extrinsic camera-to-kinematic model calibration in such applications. Specifically, 1) an outline for optimal extrinsic camera calibration for long-reach robotic manipulators in unknown environments is presented, with 2) comparison to other similar methods, and 3) real-time experiments with a visual servo use case is discussed.

A two-step methodology is proposed, in which the first step is coarse alignment of the camera frame (or coordinate system) by utilizing a kinematic model of the manipulator and the known initial pose. The second step is fine matching of pose data sequences using robust generalized point set registration [6], a method that benefits not only from position data but also from orientation data and is robust against noise and outliers that can be an encumbrance in visual measurements. For comparison, the fine matching step is also realized with the CPD algorithm [5] and a least-squares-based estimation method [10] that only utilize 3 DOF position data. It is assumed that the intrinsic parameters of the camera are pre-calibrated. Real-time experiments are presented using a laboratory-installed hydraulic crane with 5 m reach. For fine matching, the pose trajectory data are obtained using a camera located near the tip of the manipulator, with a simultaneous localization and mapping (SLAM) algorithm providing the pose estimates. A kinematic model with joint encoders is used to obtain the second set of pose trajectory data. After computing the optimal extrinsic calibration matrix, we apply it to a use case of driving the manipulator to a specific feature detected with the camera. This type of operation is very common in mining and is relatable to bolting, for example, in which supportive rods are inserted into drill holes. In this paper, only a planar case was examined, and ArUco markers [11] were used as the specific features to detect.

The paper is organized as follows: In Section II, we describe the methodology of 6 DOF pose trajectory registration; in Section III, we present the experimental setup, which was used in the real-time experiments; in Section IV, we present the measurements and results; and finally, in Section V, we conclude the paper.

II. Methodology

A. Coarse Frame Alignment

A coarse frame alignment between the camera frame and the encoder-based tool center point (TCP) frame is required for initialization. This alignment reduces the number of iterations in the fine matching step, while also reducing the possibility of the registration algorithm converging to local minima that do not produce correct matching results.

The coarse frame alignment is performed based on the known initial pose of the encoder-based TCP and applying a rigid transformation to the camera frame to roughly align the axes with the encoder-based frame axes. This step must be carefully performed to avoid issues when employing Euler angles.

B. Robust Generalized Point Set Registration

The fine matching of the 6 DOF point sets is based on a probabilistic hybrid mixture model (HMM) [6], [12] that utilizes position and orientation data. Specifically, a GMM is used to model positional uncertainties, whereas an FMM is used to model the orientation uncertainties. The optimal (rigid) transformation between two point sets is solved iteratively using the expectation-maximization (EM) algorithm [13]. The notations used in the HMM formulation are as follows:

- M – Number of points in the encoder-based point set,
- N – Number of points in the SLAM-based point set,
- $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{3\times M}$ – encoder-based TCP position vector set,
- $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M] \in \mathbb{R}^{3\times M}$ – encoder-based TCP orientation unit vector set,
- $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{3\times N}$ – SLAM-based position vector set,
- $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_N] \in \mathbb{R}^{3\times N}$ – SLAM-based orientation unit vector set.

The encoder-based points in $\mathbf{Y}$ are considered the GMM centroids, and the respective unit orientation vectors in $\hat{\mathbf{Y}}$ are considered the mean directions of the FMM. The SLAM-based points in $\mathbf{X}$ are generated by the GMM, and the respective orientation unit vectors in $\hat{\mathbf{X}}$ are generated by the FMM. The goal is to find the optimal rigid transformation (rotation and translation) between the two pose trajectory data sequences $(\mathbf{X}, \hat{\mathbf{X}})$ and $(\mathbf{Y}, \hat{\mathbf{Y}})$. The probability density function of the HMM is expressed as follows:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n) = \sum_{m=1}^{M+1} P(m)p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m), \tag{1}$$

where

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m) = \frac{\kappa}{(2\pi\sigma^2)^{\frac{3}{2}} 2\pi(e^{\kappa} - e^{-\kappa})} e^{\kappa(\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n - \frac{1}{2\sigma^2}||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2}. \quad (2)$$

The variance parameter of the GMM is denoted by $\sigma^2 \in \mathbb{R}$, the concentration parameter of the FMM is denoted by κ, $(\mathbf{x}_n, \hat{\mathbf{x}}_n)$, $(\mathbf{y}_m, \hat{\mathbf{y}}_m)$ denote arbitrary data points in the point sets, and $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$ denote the rotation and translation transformations applied to $(\mathbf{Y}, \hat{\mathbf{Y}})$, respectively. The assumption is made that the position and orientation data are independent.

To account for noise and outliers in the SLAM-based pose data, an additional uniform distribution is added to the model:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n|M+1) = \frac{1}{N} \quad (3)$$

with equal membership probabilities $P(m) = \frac{1}{M}$ assumed for the GMM components. The complete HMM is now as follows:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n) = w\frac{1}{N} + (1-w)\sum_{m=1}^{M}\frac{1}{M}p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m), \quad (4)$$

where $w \in [0, 1]$ denotes the weight of the uniform distribution. To find the optimal set of parameter estimates $\mathbf{R}, \mathbf{t}, \kappa$, and σ^2, the following negative log-likelihood function is to be minimized:

$$E(\mathbf{R}, \mathbf{t}, \kappa, \sigma^2) = -\sum_{n=1}^{N}\log\sum_{m=1}^{M+1}P(m)p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m). \quad (5)$$

The EM algorithm is used to obtain the parameter estimates in an iterative manner. New parameters are found by minimizing the complete negative log-likelihood function:

$$Q = -\sum_{n=1}^{N}\sum_{m=1}^{M+1}P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n)\log(P^{new}(m)p^{new}(\mathbf{x}_n, \hat{\mathbf{x}}_n|m)). \quad (6)$$

Then, the encoder-based TCP data $(\mathbf{Y}, \hat{\mathbf{Y}})$ are transformed by applying $\mathbf{R}$ and $\mathbf{t}$. Ignoring constants independent of $\mathbf{R}, \mathbf{t}, \kappa$, and σ^2, (6) is reformulated as follows:

$$Q(\mathbf{R}, \mathbf{t}, \kappa, \sigma^2) = \sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}\left(\frac{1}{2\sigma^2}||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2 - \kappa((\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n)\right) + \frac{3}{2}N_{\mathbf{P}}\log\sigma^2 + N_{\mathbf{P}}\log(e^{\kappa} - e^{-\kappa}) - N_{\mathbf{P}}\log\kappa, \quad (7)$$

where $p_{mn} = P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n)$, $N_{\mathbf{P}} = \sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}$. The Bayes theorem is used to compute the posterior probabilities p_{mn} as follows:

$$P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n) = \frac{P(m)p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m)}{p(\mathbf{x}_n, \hat{\mathbf{x}}_n)}. \quad (8)$$

According to the EM algorithm, the parameters $\mathbf{R}, \mathbf{t}, \kappa$ and σ^2 are updated in an iterative manner until convergence. The optimal translation $\mathbf{t}^*$ is obtained by minimizing (7) with respect to $\mathbf{t}$, whereas the optimal rotation matrix $\mathbf{R}^*$ is obtained by minimizing (7) with respect to $\mathbf{R}$, respectively. The resulting solutions are as follows:

$$\mathbf{R}^* = \mathbf{V}\ \mathrm{diag}([1, 1, \det(\mathbf{V}\mathbf{U}^{\mathrm{T}})])\ \mathbf{U}^{\mathrm{T}} \quad (9)$$

$$\mathbf{t}^* = \boldsymbol{\mu}_x - \mathbf{R}^*\boldsymbol{\mu}_y, \quad (10)$$

where the mean positional vectors for each point set are defined as follows:

$$\boldsymbol{\mu}_x = \frac{1}{N_{\mathbf{P}}}\mathbf{X}\mathbf{P}^{\mathrm{T}}\mathbf{1}, \quad \boldsymbol{\mu}_y = \frac{1}{N_{\mathbf{P}}}\mathbf{Y}\mathbf{P}\mathbf{1}, \quad (11)$$

$\mathbf{P} \in \mathbb{R}^{M\times N}$ has elements p_{mn} in (8), and $\mathbf{1}$ is a vector of ones. The singular value decomposition (SVD) of $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ is used to obtain $\mathbf{V}$ and $\mathbf{U}$, where $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$, $\mathbf{H} \in \mathbb{R}^{3\times 3}$ and

$$\mathbf{H}_1 = \mathbf{Y}'\mathbf{P}\mathbf{X}', \quad \mathbf{H}_2 = \hat{\mathbf{Y}}\mathbf{P}\hat{\mathbf{X}}^{\mathrm{T}}. \quad (12)$$

The matrices $\mathbf{Y}'$ and $\mathbf{X}'$ contain de-meaned positional data $\mathbf{y}'_m = \mathbf{y}_m - \boldsymbol{\mu}_y$ and $\mathbf{x}'_n = \mathbf{x}_n - \boldsymbol{\mu}_x$.

The variance parameter of the GMM is updated by minimizing (7) with respect to σ^2:

$$\sigma^2 = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}(||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2)}{3N_{\mathbf{P}}}. \quad (13)$$

The concentration parameter (κ) of the FMM is updated using two parts [7]. The first part r_1 results from orientation error and is computed as follows:

$$r_1 = \frac{1}{N_{\mathbf{P}}}\sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}(\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n. \quad (14)$$

The second part r_2 is caused by positional error and is computed as follows:

$$r_2 = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}{\mathbf{x}'_n}^{\mathrm{T}}\mathbf{R}\mathbf{y}'_m}{\sum_{n=1}^{N}\sum_{m=1}^{M}p_{mn}||\mathbf{R}\mathbf{y}'_m||\ ||\mathbf{x}'_n||}. \quad (15)$$

Then, κ is updated with $\kappa = r(3 - r^2)/(1 - r^2)$, where $r = vr_1 + (1-v)r_2$, in which $v = 0.5$.

After successful convergence, the optimal calibration matrix for fine matching is written as follows:

$$\mathbf{T}_{fm} = \begin{bmatrix} & \mathbf{R}^* & & \mathbf{t}^* \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (16)$$

During iteration, the algorithm was stopped if one of the following conditions was met: $\sigma^2 < 10^{-6}$, $|\sigma_{i+1}^2 - \sigma_i^2| < 10^{-6}$, or 100 iterations were reached. The maximum concentration parameter was also set as $\kappa_{max} = 100$ to avoid computational issues.

The initial iteration parameters were set as follows: $\mathbf{R} = \mathbf{I} \in \mathbb{R}^{3\times 3}$, $\mathbf{t} = \mathbf{0}$, $\sigma_0^2 = \sum_{n+1}^{N}\sum_{m+1}^{M}||\mathbf{x}_n - \mathbf{y}_m||^2/(3MN)$, and $\kappa = 1$.

Finally, the extrinsic camera-to-kinematic model calibration matrix is formulated as follows:

$$\mathbf{T} = \mathbf{T}_{fm}^{-1}\mathbf{T}_{cfa}\mathbf{T}_{slam}, \quad (17)$$

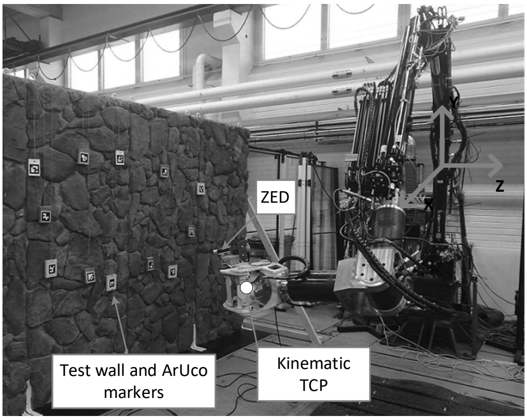


Fig. 1. The experimental setup showing the manipulator, the ZED attached to the claw, a test wall, and 12 ArUco markers placed in the workspace. The base frame of the manipulator is also (roughly) shown.

where $\mathbf{T}_{cfa} \in \mathbb{R}^{4\times4}$ denotes the coarse frame alignment homogeneous transformation matrix, and $\mathbf{T}_{slam} \in \mathbb{R}^{4\times4}$ denotes a single SLAM pose expressed with a homogeneous transformation matrix.

C. Orientation Magnitude Correction

As the FMM employs orientation *unit* vectors, the computed transformation matrix (16) cannot directly produce transformed orientations with true magnitudes. This is resolved by using the encoder measured magnitudes as references. The mathematical expression is as follows:

$$\theta_{corr}^{slam} = \begin{cases} \theta^{slam} - |\theta_f^{slam} - \theta_f^{enc}|, \text{ if } \theta_f^{slam} > \theta_f^{enc} \\ \theta^{slam} + |\theta_f^{slam} - \theta_f^{enc}|, \text{ else} \end{cases}, \tag{18}$$

where θ represents a current Euler angle, and θ_f denotes the final value of the respective variable in a calibration data sequence.

III. Experimental Setup

The experimental setup is shown in Fig. 1. The main components and systems as follows:

- HIAB033 hydraulic crane with an additional 3 DOF wrist, and each joint was equipped with an incremental encoder,
- ZED stereo camera running a SLAM algorithm,
- A dSPACE real-time control platform,
- A test wall comprising decorative stones to simulate a mine and provide visual features,
- Markers attached to the wall acting as specific features.

A dSPACE DS1005 PPC controller board served as the real-time control system, and a 2 ms sampling period was used in the experiments.

TABLE I
DH Parameters of HIAB033 With a 3 DOF Wrist

Joint	α_i	a_i	θ_i	d_i
Rotation	$\pi/2$	a_1	θ_1	d_1
Lift	0	a_2	θ_2	0
Tilt	$\pi/2$	a_3	$\theta_3 + \pi/2$	d_3
Wrist 1	$\pi/2$	0	θ_4	d_4
Wrist 2	$-\pi/2$	0	θ_5	0
Wrist 3	0	0	θ_6	d_6

A. HIAB033 Hydraulic Crane With 3 DOF Wrist

A forward kinematic representation of the manipulator is formulated using the Denavit-Hartenberg (DH) parameters, which are presented in Table I in symbolic form. The rigid transform from the base frame to the TCP frame, $\mathbf{T}_{enc}$, is formulated as follows:

$$\mathbf{T}_{enc} = \mathbf{T}_{j1}\mathbf{T}_{j2}\mathbf{T}_{j3}\mathbf{T}_{j4}\mathbf{T}_{j5}\mathbf{T}_{j6}, \tag{19}$$

where joint specific transforms $\mathbf{T}_{ji},\ i \in \{1, ..., 6\}$ are computed with

$$\mathbf{T}_i = \begin{bmatrix} c\theta_i & -s\theta_i c\alpha_i & s\theta_i s\alpha_i & a_i c\theta_i \\ s\theta_i & c\theta_i c\alpha_i & -c\theta_i s\alpha_i & a_i s\theta_i \\ 0 & s\alpha_i & c\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

while applying the respective DH parameters for each joint. Additionally, $s = sin$ and $c = cos$.

B. Visual Measurements

A Stereolabs ZED stereo camera was used in the experiments. It was installed near the tip of the manipulator, and the ROS node provided by the manufacturer was used to publish 720p images.

For SLAM, the open-source version of ORB-SLAM2 Stereo [14] was utilized. The algorithm ran in real time and the pose data were transmitted to the dSPACE controller board via UDP. A 2.5×4 m textured wall served as the main feature extraction area for the SLAM algorithm, because the main focus of this research is underground applications.

For detecting specific markers on the wall, the OpenCV ArUco detection library was used. Twelve ArUco markers were placed around the workspace of the manipulator, as in Fig. 1.

C. Robot Control

Quintic polynomial path planning [15] was used to generate trajectories, and a P-controller with a first-order time delay (PT1 control) was used on the actuator level. The controller's transfer function is described as follows:

$$G(s) = \frac{K_P}{\tau s + 1}. \tag{21}$$

The time delay term (τ) enables larger proportional gain (K_P) values, which reduces static positioning errors when driving to a specific point.

Furthermore, the manipulator was constrained so that only the first three joints (rotation, lift, and tilt) were used for

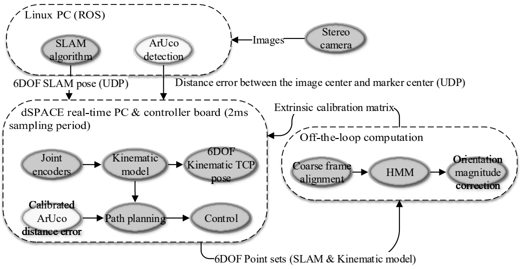

Linux PC (ROS)
SLAM algorithm
ArUco detection
Images
Stereo camera
6DOF SLAM pose (UDP)
Distance error between the image center and marker center (UDP)
dSPACE real-time PC & controller board (2ms sampling period)
Extrinsic calibration matrix
Off-the-loop computation
Joint encoders
Kinematic model
6DOF Kinematic TCP pose
Coarse frame alignment
HMM
Orientation magnitude correction
Calibrated ArUco distance error
Path planning
Control
6DOF Point sets (SLAM & Kinematic model)

Fig. 2. A simplified diagram of the experimental setup: The camera algorithms were processed on a dedicated Linux PC running ROS and the desired camera measurements were sent to the dSPACE real-time control PC via UDP. The camera-to-kinematic model calibration was processed outside the 2 ms control loop. The resulting extrinsic calibration matrix, computed using the two point sets, was then updated in the main control system.

motion, whereas the wrist joints moved only to keep the orientation of the wrist constant.

A simplified diagram of the experimental setup is shown in Fig. 2, in which the orange blocks are related to the overall system, whereas the yellow blocks are related to the use case regarding the ArUco markers.

IV. Measurements and Results

First, a calibration measurement was conducted, in which the manipulator was arbitrarily moved around the workspace to obtain pose data sequences using SLAM and encoder measurements. The recorded data were used to compute the optimal calibration matrix by first applying coarse alignment transform to the SLAM-based pose data by using (17). Then, the coarse frame aligned pose data were used for fine matching, i.e., point set registration with the encoder-based pose data by using the robust generalized point set registration algorithm (4)–(16). The three point sets (encoder-based, SLAM-based with coarse frame alignment, and SLAM-based after fine matching) are shown in Fig. 3. The black points represent the encoder-based TCP position data, whereas the red point sets represent the SLAM-based position data before and after fine matching. The individual pose variables are presented in Fig. 4, where the black lines denote the encoder-based pose variables, and the red lines represent the calibrated SLAM-based pose variables. As illustrated, the algorithm was able to accurately match the pose trajectories resulting from arbitrary motions. Two additional separate calibrations were performed, using the same coarse frame alignment transform, for which the results are shown in Fig. 5 and Fig. 6. The number of iterations required for the fine matching varied between 20 and 25.

After each calibration, the manipulator was driven to 12 different ArUco markers that were placed around the workspace. Monocular detection was employed by using the left lens of ZED, and the middle of the image was treated as the TCP that was to be driven to a marker center. An example image of the left camera view is shown in Fig. 7. The metric distance between the camera center and a marker center was computed based on the known marker size from the image. Then, the point distance was calibrated with the camera-to-kinematic model calibration. Only the rotation part of (16) was required to transform the camera reference to the kinematic frame, meaning this use case does not suffer from the larger position errors in the calibration. The Euclidean distance errors for each marker, for each of the three calibrations, are documented in Table II. The errors were measured from the images. The average positioning error in each measurement was less than 1.0 cm, which is acceptable for this type of application. Only planar results are presented, as the ZED camera was not able to provide reliable depth measurements.

To compare the HMM-based 6 DOF point set registration method with methods utilizing only 3 DOF position data, offline data analysis was conducted for the three measurements. Namely, the very similar CPD algorithm [5] employing a GMM and a simple pairwise least-squares-based estimation algorithm [10] were chosen for comparison purposes. The coarse frame alignment and the orientation magnitude correction steps were performed identically in each case, with only the fine matching step changing. Furthermore, to test the robustness of the three algorithms, Gaussian noise was injected to the X-axis position with varying signal-to-noise ratios (SNR). The SNRs tested were 10, 20, and 30 dB. The effect of added noise to the signal is illustrated in Fig. 8. The root mean square errors (RMSE) for 3 DOF position and 3 DOF orientation in each measured case are presented in Tables III-V. As shown, despite the arbitrary motions in each measurement, the resulting errors are very similar. The position errors are on the centimeter range, whereas the orientation errors are less than 2°. The errors follow from kinematic inaccuracies, for example, due to flexibility, which makes perfect pose trajectory matching practically impossible. The bending of the manipulator is witnessed especially in the X-axis orientations estimated with the SLAM algorithm. Visual measurements are also susceptible to outliers and errors, however, they performed well in the experiments.

As seen from Tables III-V, minimal differences can be found between the fine matching algorithms. When Gaussian noise is added to the X-axis position signal, however, the utilization of orientation data in the HMM appears to slightly improve the matching result compared to the CPD, which only incorporates position data via a GMM. We also experimented by adding similar noise to the other signals, including orientations, which showed uniform results with the presented case. However, to obtain the best result, the weight of the uniform distribution in (4) should be tuned. Same values were used for both the HMM-based method and the CPD. In the cases where noise was added, the least-squares-based algorithm provided the least accurate results. It is worth noting the least-squares-based method required pairwise point sets, whereas the HMM-based method and the CPD are able to process point sets that do not match in length.

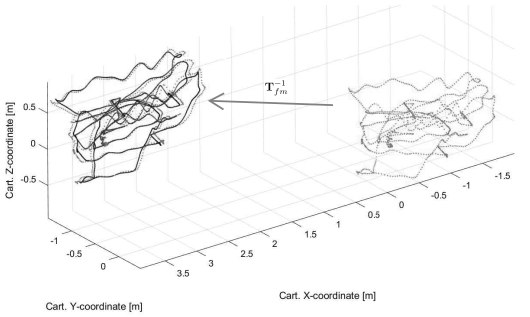


Fig. 3. For computing the calibration matrix, the manipulator was moved arbitrarily around the workspace, while the pose trajectories were recorded for point set registration. The black points represent the encoder-based point set. The right side red points represent the coarse frame aligned SLAM-based point set, whereas the left side red points represent the same SLAM-based point set after fine matching using the HMM.

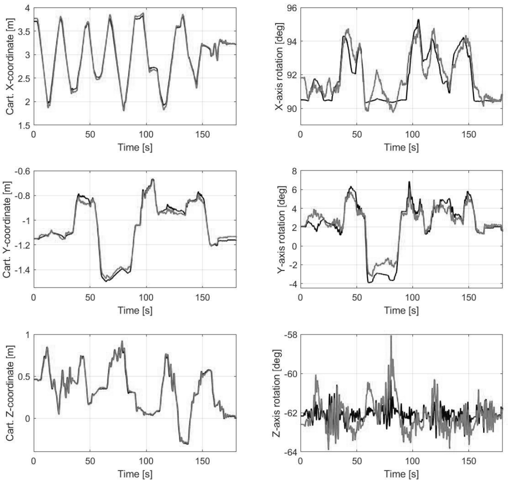


Fig. 4. The first calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

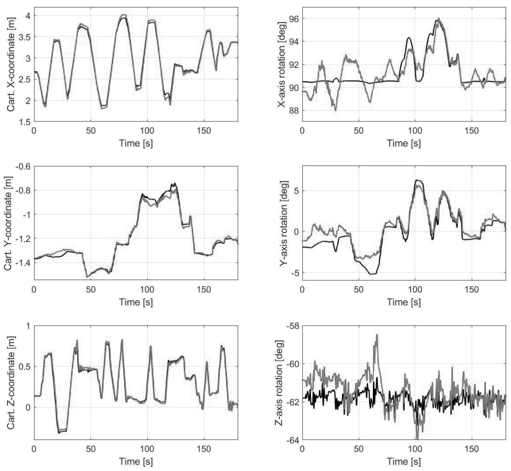


Fig. 5. The second calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

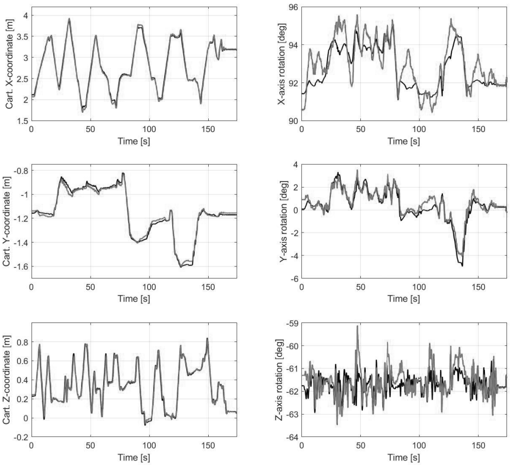


Fig. 6. The third calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

V. Discussion and Conclusion

In this paper, a methodology for camera-to-kinematic model calibration was proposed, with camera-aided operations for long-reach manipulators in unknown environments as motivation. The goal of this method is to be able to perform fast extrinsic camera calibration easily on the worksite, with arbitrary manipulator motions and in unknown environments. The methodology comprised coarse frame alignment based on the known initial pose of the manipulator and fine matching based on the robust generalized point set registration that benefits not only from position data but also from orientation data, which is perceived as optimal for robotic applications that have complete pose data. Comparison with two other methods utilizing only position data was conducted in offline data analysis, with the results suggesting that utilizing both the orientation and position data is most efficient. As the FMM resolves orientation using unit vectors, a simple solution for correcting the transformed orientation magnitudes using the joint sensors present in the system was shown.

Real time experiments were conducted using a hydraulic

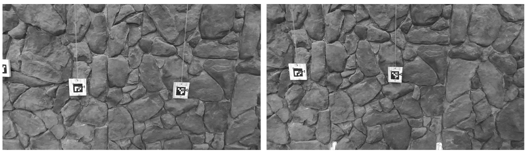


Fig. 7. The left image shows the initial pose of the camera, whereas the right image shows a control result of driving the TCP (image center) to a specific marker.

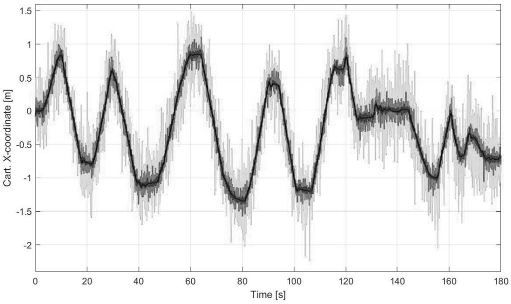


Fig. 8. Gaussian noise added to the X-axis position signal. The black line denotes the raw signal, the green line denotes SNR 10 dB, the red line denotes SNR 20 dB, and the blue line denotes SNR 30 dB.

TABLE II
Euclidean Distance Errors Between the Image Center and Marker Centers

	Meas. 1 [m]	Meas. 2	Meas. 3
ArUco#1	0.0085	0.0073	0.0082
ArUco#2	0.0075	0.0066	0.0095
ArUco#3	0.0047	0.0077	0.0031
ArUco#4	0.0052	0.0079	0.0072
ArUco#5	0.0063	0.0097	0.0095
ArUco#6	0.0077	0.0100	0.0083
ArUco#7	0.0103	0.0127	0.0096
ArUco#8	0.0108	0.0100	0.0110
ArUco#9	0.0104	0.0088	0.0104
ArUco#10	0.0105	0.0114	0.0102
ArUco#11	0.0114	0.0104	0.0112
ArUco#12	0.0065	0.0092	0.0083
Avg.	0.0083	0.0093	0.0089

TABLE III
Root Mean Square Errors for 3 DOF Position and 3 DOF Orientation in the First Measurement

Fig. 4 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0331	0.0331	0.0332
Raw signals [deg]	1.1864	1.1864	1.1849
SNR 10 dB [m]	0.0337	0.0339	0.0340
SNR 10 dB [deg]	1.1894	1.1998	1.1867
SNR 20 dB [m]	0.0272	0.0273	0.0274
SNR 20 dB [deg]	1.1876	1.1882	1.1844
SNR 30 dB [m]	0.0259	0.0259	0.0260
SNR 30 dB [deg]	1.1858	1.1863	1.1848

TABLE IV
Root Mean Square Errors for 3 DOF Position and 3 DOF Orientation in the Second Measurement

Fig. 5 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0375	0.0375	0.0393
Raw signals [deg]	1.5895	1.5895	1.5926
SNR 10 dB [m]	0.0335	0.0351	0.0335
SNR 10 dB [deg]	1.5896	1.5896	1.5931
SNR 20 dB [m]	0.0286	0.0287	0.0289
SNR 20 dB [deg]	1.5925	1.5925	1.5932
SNR 30 dB [m]	0.0292	0.0292	0.0302
SNR 30 dB [deg]	1.5897	1.5897	1.5924

TABLE V
Root Mean Square Errors for 3 DOF Position and 3 DOF Orientation in the Third Measurement

Fig. 6 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0287	0.0287	0.0296
Raw signals [deg]	0.9665	0.9665	0.9679
SNR 10 dB [m]	0.0239	0.0240	0.0248
SNR 10 dB [deg]	0.9663	0.9665	0.9678
SNR 20 dB [m]	0.0258	0.0259	0.0267
SNR 20 dB [deg]	0.9677	0.9680	0.9680
SNR 30 dB [m]	0.0239	0.0240	0.0248
SNR 30 dB [deg]	0.9663	0.9665	0.9678

manipulator with three moving joints. The results showed that the proposed methodology was able to match the pose variables sufficiently in each measured case. Inaccuracies in the matching result were caused by, for example, the rigidity assumption in the kinematic formulation. Furthermore, in the use case, the results were promising for visually assisted operations in applications involving long-reach manipulators with uncertainties, as an acceptable average positioning error was achieved. Some challenges include reliance on the performance of the SLAM algorithm in the sense that the variables may drift during the calibration sequence, for example. Another challenge is that if the camera and the kinematic TCP are on different rotation axes, the two point sets cannot be matched with good accuracy due to the camera's offset.

Acknowledgment

This work was supported by the Doctoral School of Industry Innovations (DSII) of Tampere University.

References

[1] R. Y. Tsai, R. K. Lenz *et al.*, "A new technique for fully autonomous and efficient 3 d robotics hand/eye calibration," *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 345–358, 1989.

[2] C.-C. Wang, "Extrinsic calibration of a vision sensor mounted on a robot," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 2, pp. 161–175, 1992.

[3] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–607.

[4] A. Peters, A. Schmidt, and A. C. Knoll, "Extrinsic calibration of an eye-in-hand 2d lidar sensor in unstructured environments using icp," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 929–936, 2020.

[5] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[6] Z. Min, J. Wang, and M. Q.-H. Meng, "Robust generalized point cloud registration using hybrid mixture model," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4812–4818.

[7] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway, "Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research*, vol. 6, no. 9, 2005.

[8] P. Mäkinen, M. M. Aref, J. Mattila, and S. Launis, "Application of simultaneous localization and mapping for large-scale manipulators in unknown environments," in *Cybernetics and Intelligent Systems (CIS) and IEEE Conf. Robotics, Automation and Mechatronics (RAM), 2019 IEEE 9th Inter. Conf.* IEEE, 2019.

[9] P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Redundancy-based visual tool center point pose estimation for long-reach manipulators," in *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM).* IEEE, 2020, pp. 1387–1393.

[10] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

[11] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[12] Z. Min, J. Wang, and M. Q.-H. Meng, "Robust generalized point cloud registration with orientational data based on expectation maximization," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 1, pp. 207–221, 2019.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[14] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[15] R. N. Jazar, *Theory of Applied Robotics - Kinematics, Dynamics, and Control*. Dordrecht, the Netherlands: Springer, 2010.

PUBLICATION IV

Model-free sensor fusion for redundant measurements using sliding window variance

P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila

In *2022 22st International Conference on Control, Automation and Systems (ICCAS)*, 2022, pp. 1481–1487
DOI: 10.23919/ICCAS55662.2022.10003720

2022 The 22st International Conference on Control, Automation and Systems (ICCAS 2022)
BEXCO, Busan, Korea, Nov. 27∼Dec. 01, 2022

Model-Free Sensor Fusion for Redundant Measurements Using Sliding Window Variance

P. Mäkinen[1*], P. Mustalahti[1], S. Launis[2] and J. Mattila[1]

[1]Automation Technology and Mechanical Engineering, Tampere University, Tampere, Finland,
(petri.makinen@tuni.fi, pauli.mustalahti@tuni.fi, jouni.mattila@tuni.fi) * Corresponding author
[2]Rock Technologies and Drilling, Sandvik Mining and Construction Oy,
Tampere, Finland (sirpa.launis@sandvik.com)

Abstract: In this paper, a model-free data fusion method for combining redundant sensor data is presented. The objective is to maintain a reliable tool center point pose measurement of a long-reach robotic manipulator using a visual sensor system with multiple cameras. The fusion method is based on weighted averaging. The weight parameter for each variable is computed using the sliding window variance with N latest observations. After each sliding window, the window length N is updated, and simple transition smoothing is included. For experimental validation, two sets of pose trajectory data from redundant visual sensors were obtained: 1) using a camera located near the tip of a long-reach manipulator running a simultaneous localization and mapping (SLAM) algorithm and 2) marker-based tracking with cameras located near the base of the manipulator. For pose tracking, a fiducial marker was attached near the SLAM camera. The proposed methodology was examined using a real-time measurement setup and offline data analysis using the recorded data. The results demonstrate that the proposed system can increase the overall robustness and fault tolerance of the system, which are desired features for future autonomous field robotic machines.

Keywords: Sensor Fusion, Machine Vision, Sensor Systems and Applications

1. INTRODUCTION

The heavy machinery industry is taking major leaps toward electrification and autonomous systems. These heavy-duty mobile machines require new intelligent algorithms and sophisticated sensors [1] in order to work independently in harsh environments, such as mines. A variety of long-reach robotic manipulators are found in such machines to perform various tasks related to mining and heavy lifting, for example [2]. One of the key challenges is replacing human vision and decision making with sensors and computerized algorithms in order to perform work tasks autonomously. For this purpose, measurement information about the manipulator's tool center point (TCP) is typically required. The TCP pose (3 degrees-of-freedom (DOF) position and 3 DOF orientation) can be obtained using forward kinematics with joint encoders. For a small-scale, ideal industrial robot, an accurate forward kinematic model can be obtained. However, this is not the case for long-reach manipulators, as they have significant structural flexibilities that are not considered by traditional rigid-body kinematics. Consequently, visual servoing methods have been well established for small-scale industrial robots. In contrast, for long-reach manipulators working in unstructured environments, there are challenges with visual sensing related to camera calibration, view distance, field of view, and occlusions, for example [3].

In an attempt to replace human vision in applications striving toward autonomy, a wide variety of visual sensors have been investigated, including radar technology and optical methods, such as laser scanners and camera systems [4]. Data provided by different proprioceptive and exteroceptive sensors can be combined to obtain a more accurate or robust picture of an observed system. This process of combining sensor information is called multi-sensor fusion, or simply sensor fusion. Based on how sensor information is utilized, fusion methods are usually classified as competitive, complementary, or cooperative systems [5]. For sensor fusion, the Kalman filter and its nonlinear variants are popular methods [6-7]. Neural networks and fuzzy set theory have also been investigated [8-9]. The most mature branch of sensor fusion is perhaps related to self-driving vehicles, which have been avidly examined [10-11] and with, for example, Tesla Autopilot available for consumer vehicles. These systems are built on deep learning algorithms, requiring massive amounts of training data, which, scale-wise, are not feasible for the low-volume heavy machinery industry.

One of the previous studies on multi-sensor integration [5] argued that the key to intelligent fusion of disparate sensory information is to provide an effective model of sensor capabilities. However, in some cases, finding a sufficient sensor model may not be possible. Research on such model-free sensor data fusion methods is very limited and restricted to simple, albeit potentially effective, methods. For example, in [12], fusion was carried out using confidence weighted averaging. However, determining the confidence functions for weight computations in dynamic, online scenarios has not been well established.

In this paper, we focus on combining continuous measurements of the same variables from different sensors in an attempt to increase the system's robustness and reliability. The fusion method is a statistical approach based on confidence weighted averaging [12], with our contributions including determining the weight parameters in a dynamic manner using sliding window variance (sample variance) instead of a specified confidence function. The sliding window length is updated after each individual window, and a simple approach for transition smoothing is also presented. The proposed methodology was investigated using a real-time setup comprising a long-reach hydraulic manipulator. The objective was to estimate the end effector's pose with visual sensors [13]. The first pose estimate was obtained using a camera running a simultaneous localization and mapping (SLAM) algorithm, with the camera attached near the tip of the manipulator. The second pose estimate was obtained using marker-based tracking, with a fiducial marker attached near the SLAM camera. Before sensor fusion was performed, the pose variables were extrinsically calibrated to a concurrent coordinate system according to [14].

The underlying motivation with this configuration is that the SLAM algorithm has a narrow but accurate view, whereas the marker-tracking cameras are placed on top of a machine and provide a wider view but a less accurate pose measurement. A conceptual example is shown in Fig. 1. For this application, a marker provides a fixed and repeatable target. Consequently, the visual sensors are able to complement each other, but also provide the necessary redundancy, as both sensing methods are susceptible to faulty situations. These include, for example, marker occlusions and insufficient feature extraction for SLAM.

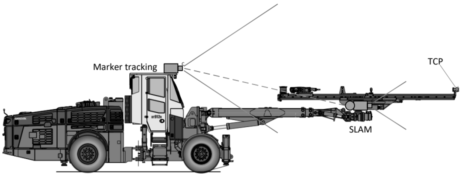


Fig. 1. The overall conceptual design: The TCP is observed using a visual sensor system comprising marker tracking and SLAM modules. The Sandvik DT912D single boom tunneling jumbo is shown as an example.

The remainder of the paper is organized as follows: The data fusion methodology is described in Section 2, the experimental setup is detailed in Section 3, the results are discussed in Section 4, and finally Section 5 concludes the paper.

2. METHODOLOGY

2.1 Data fusion using sliding window variance

A fused sensor signal can be formulated by taking the weighted average of all the sensor signals that estimate the same variable:

$$x_F = \sum_{i=1}^{n} w_i x_i, \tag{1}$$

where x_F is the fused signal, w_i denotes the weight parameters, x_i denotes the redundant sensor data, and n is the number of sensors. The fused variance can be written as

$$\sigma_F^2 = \sum_{i=1}^{n} w_i \sigma_i^2, \tag{2}$$

where σ_F^2 is the fused variance, and σ_i^2 denotes the input signal variance.

To obtain the optimal fused measurement, the weight parameters should be chosen so that the fused variance is minimized, which can be achieved by solving the following minimization problem:

$$\arg\min_{w_i} \sum_{i=1}^{n} w_i^2 \sigma_i^2, \tag{3}$$

with the sum of all weights w_i equal to 1. Solving the minimization problem results in the following equation to compute the weights:

$$w_i = \frac{1}{\sigma_i^2} \sum_{j=1}^{n} \frac{1}{\sigma_j^2}. \tag{4}$$

Substituting Eq. (4) into Eq. (2) shows that for $n \geq 2$, the fused variance is always smaller than the input variances.

The variance for a given measurement signal is computed over a sliding window of length N data points:

$$\sigma_i^2 = \frac{1}{N-1} \sum_{j=1}^{N} |x_j - \mu_i|^2, \tag{5}$$

where μ_i denotes the mean of x_j over the sliding window of N observations and is computed as

$$\mu_i = \frac{1}{N} \sum_{i=j}^{N} x_j. \tag{6}$$

Then, the sliding window variances are used to compute the weight parameters using Eq. (4) for data fusion.

The rationale is that computing the weights based on sliding window variances of redundant measurements will emphasize better-quality signals, as it is expected that a signal with less variance is more accurate. This derives from the assumption that the measurements are reliable, and grossly faulty measurements are detected and discarded before the data fusion procedure.

2.2 Updating the sliding window length

The length of the sliding window, denoted by N, is updated at the end of each window, which is conducted as follows:

$$N = k_N \max|\boldsymbol{\mu}_{x_1} - \boldsymbol{\mu}_{x_j}|, \tag{7}$$

where $\boldsymbol{\mu}_{x_1}$ and $\boldsymbol{\mu}_{x_j}$, $j \in \{2, ..., n\}$ denote vectors of the mean values of each sensor measurement, computed over the entire current sliding window. The largest absolute mean difference is used to compute the next window length, and the coefficient k_N is used to tune the window length to a desired scale. Note that N is rounded to an integer value. The sliding window length should be constrained between the minimum N_{min} and maximum N_{max} values to avoid computational issues.

2.3 Transition smoothing

Raw sensor data can contain occasional outliers, or some sensors may cease to operate, which requires smoothly and safely transitioning the fused sensor signal to exclude the unavailable sensor measurement. For this purpose, we use a simple transition smoothing method. The fused sensor signal $\hat{x}_F$ is computed using the following condition:

$$\hat{x}_{F,new} = \begin{cases} x_F & \text{if } |\hat{x}_{F,previous} - \hat{x}_{F,current}| < \epsilon \\ x_{F_{corr}} & \text{otherwise} \end{cases}. \quad (8)$$

The error coefficient ϵ determines the limit after which the transition smoothing is applied. If the absolute difference between the previous and current fused values is less than the designated error coefficient, the next fused value is computed normally using Eq. (1). If the condition is not met, the next fused value is predicted in a naïve manner by using linear interpolation. A polynomial function $p(x)$ of k degree is written as:

$$p(x) = p_1 x^k + p_2 x^{k-1} + ... + p_k x + p_{k+1}. \quad (9)$$

Considering a first-order polynomial, the linear system can be presented as follows:

$$\begin{bmatrix} t_1 & 1 \\ t_2 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \hat{x}_{F,previous} \\ \hat{x}_{F,current} \end{bmatrix}, \quad (10)$$

where $\{t_1, t_2\}$ are time stamps dictating the rate of the desired transition smoothing, and $\{p_1, p_2\}$ are polynomial coefficients to be solved. Then, the new corrected fused value for the next time step is obtained using Eq. (9):

$$x_{F_{corr}} = p_1 * (t_1 + Ts) + p_2, \quad (11)$$

where Ts is the sampling period.

3. EXPERIMENTAL SETUP

For validating the proposed model-free sensor fusion pipeline, two sets of 6-DOF pose trajectories were obtained from two redundant visual sensing methods. The experimental system is illustrated in Fig. 2, and it comprised a hydraulic manipulator, a stereo camera for SLAM, and a motion capture system for marker tracking.

The real-time system controlling the hydraulic manipulator was a Beckhoff CX2030 industrial PC. During the experiments, the manipulator was moved arbitrarily around its workspace. All the measurement data were collected by the Beckhoff PC to ensure time synchronization.

3.1 SLAM module

A ZED2 stereo camera was attached near the tip of the hydraulic manipulator. The camera was connected to a dedicated Linux PC running ROS (the robot operating system), and 720p images were published using the manufacturer-provided ROS node.

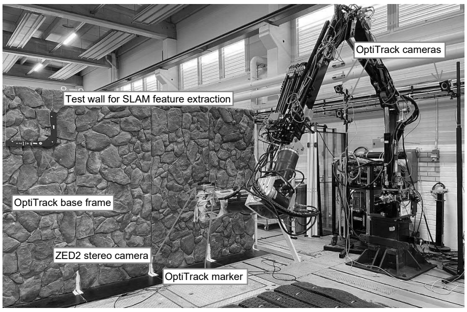


Fig. 2. The experimental setup: The hydraulic manipulator was moved arbitrarily around its workspace, and two sets of 6-DOF pose trajectory data were obtained using 1) the ZED2 stereo camera for SLAM and 2) the OptiTrack motion capture system for tracking the marker pose with respect to the OptiTrack L-frame.

For SLAM, we used the open-source ORB-SLAM2 Stereo[1] algorithm [15]. The algorithm ran on the dedicated Linux PC in real time using the images published by the ZED2 ROS node, and the 6-DOF pose trajectory data were transmitted to the Beckhoff industrial PC via UDP (user datagram protocol).

3.2 Marker tracking module

The marker-tracking module comprised three OptiTrack Prime 17W wide angle coverage cameras, a passive marker, and a base frame. The cameras were placed on high pillars around the base of the manipulator. The base frame (or OptiTrack L-frame) was placed in view of the cameras, and the marker was attached near the tip of the manipulator. The system then tracked the 6-DOF marker pose with reference to the L-frame.

A dedicated laptop with OptiTrack's Motive software was used to set up the marker-tracking module. A MATLAB plugin was configured to transmit the 6-DOF pose trajectory data to the Beckhoff industrial PC.

3.3 Signal calibration

Sensor fusion requires variables that represent the same information. A requirement before fusion is that the measured variables are transformed from each sensor's local coordinate system to a common one [16]. For pose estimates, this implies that the poses must be expressed with respect to a concurrent coordinate system. This extrinsic calibration is defined as a rigid transformation, comprising a rotation matrix and a translation vector, from one coordinate system to another.

Obtaining this rigid relationship can be a challenging task especially with field robotic systems due to the unstructured and unknown environments. In this work, we used a probabilistic point set matching-based methodology [14] to find the transformation between the two visual sensor coordinate systems. Specifically, the SLAM poses

[1] https://github.com/raulmur/ORB_SLAM2

were calibrated to the OptiTrack's base frame.

3.4 Real-time implementation

After the extrinsic calibration, the two pose measurements were expressed in the same coordinate system and fused according to the methodology detailed in Section 2. The error limit for transition smoothing in Eq. (8) was computed using the maximum absolute errors resulting from the extrinsic calibration as follows: $\epsilon = k_\epsilon \epsilon_{calib}$, where k_ϵ is a multiplication factor used to tune the transition smoothing. The relevant parameters applied in the experiments are shown in Table 1 and they were manually optimized for the investigated application.

Table 1. Applied parameters.

N_{min}	N_{max}	k_N	k_ϵ	t_1	t_2	Ts
200	4000	$30 * 10^4$	1.0	0 s	0.5 s	1 ms

The data fusion algorithm was initialized using the set maximum window size, which took 4 s with the applied parameters. Normal operation was commenced only after the initialization. The sliding window was implemented so that the overall window size was constant (the set maximum size), with the unused elements set to zero. The variance and mean value Eqs. (5)–(6) related to each measured variable were computed over the nonzero elements, with the length of the nonzero variables depending on the current window size N.

In the case of noisy data, we suggest using a geometric moving average filter [17] before data fusion for improved signal quality. The filter is formulated as follows:

$$X_i = (1 - \alpha)X_{i-1} + \alpha x_i, \quad i > 0, \tag{12}$$

where X_i is the conditioned output signal at time i, x_i is the unconditioned input signal at time i, and $0 < \alpha \leq 1$ is the filter gain, for which a low value is advised. The results presented in this paper, however, were unfiltered.

4. RESULTS AND DISCUSSION

The data fusion algorithm was tested online on the real-time system, and data were recorded for further offline data analysis. The results presented here were obtained in MATLAB's Simulink environment using the recorded data. Three cases were studied: Case 1: normal operation, Case 2: updating the sliding window length, and Case 3: transition smoothing.

For Case 1, Figs. 3–6 illustrate the poses, weight parameters, variances, and sliding window lengths obtained from the same measurement. In general, the red lines denote SLAM signals, the green lines denote marker-tracking signals, and the black lines denote fused signals. Fig. 3 shows the poses for which the positional components of both measurements perform similarly. Regarding the orientation measurements, the marker-tracking signals were noisy, whereas SLAM provided better-quality signals. Thus, the fused signals emphasized the orientations provided by the SLAM module. The difference between the position and orientation measurements is also demonstrated in the weight parameters in Fig. 4: The position signals had similar qualities, resulting in uniform weight parameter distributions. The SLAM orientation measurements had better qualities in the sense that the sliding window variances were smaller, resulting in larger weight parameters for the SLAM orientations. Note that the weight parameters were set to zeros during the initialization, but during normal operation the total sum of the weights is equal to 1. The respective signal variances are shown in Fig. 5. As discussed, the fused variances are always smaller than the input variances as a result from Eq. (3). Finally, Fig. 6 illustrates the sliding window lengths for each variable. As shown, the fusion algorithm was initialized with the set maximum window length $N_{max} = 4000$, after which the lengths were updated after each sliding window according to Eq. (7).

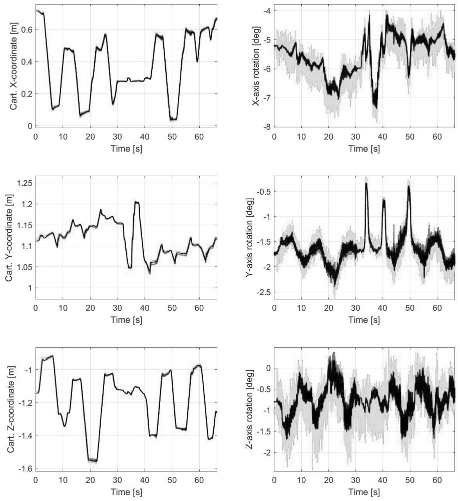


Fig. 3. Case 1: All six pose variables are shown. The red lines are the SLAM signals, the green lines are the marker-tracking signals, and the black lines are the fused signals.

For Case 2, the goal was to demonstrate the impact of the sliding window length on the fusion output. Figs. 7–9 show the fused poses for three instances: constant (minimum) sliding window length, constant (maximum) sliding window length, and variable sliding window length (same as in Case 1). For clearer visualization, at 30 s, the SLAM position signals were artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°. The left figures show the pose signals, and the right figures show the respective sliding window lengths. On the left, the red lines are the SLAM signals, the green lines are the marker-tracking signals, and the black lines are the fused signals. Fig. 7 shows the results when the set minimum sliding window length of $N_{min} = 200$ is

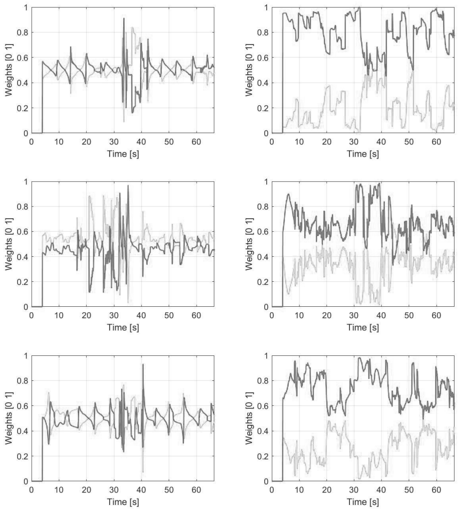


Fig. 4. Case 1: The weight parameters for each fused signal at the given time stamps. The red lines represent the SLAM weight parameters, whereas the green lines represent the marker-tracking weight parameters.

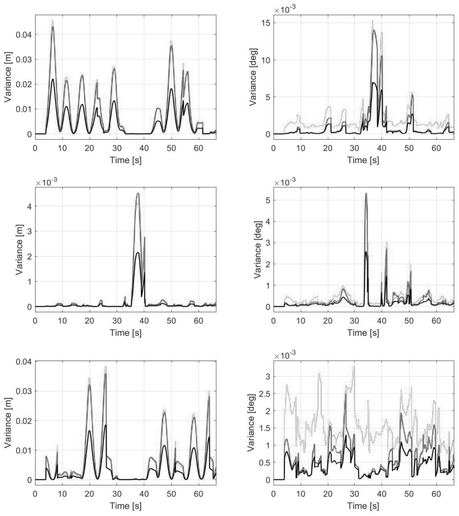


Fig. 5. Case 1: Computed variances over the sliding window for each signal. The red lines are the SLAM signal variances, the green lines are the marker-tracking signal variances, and the black lines are the fused variances.

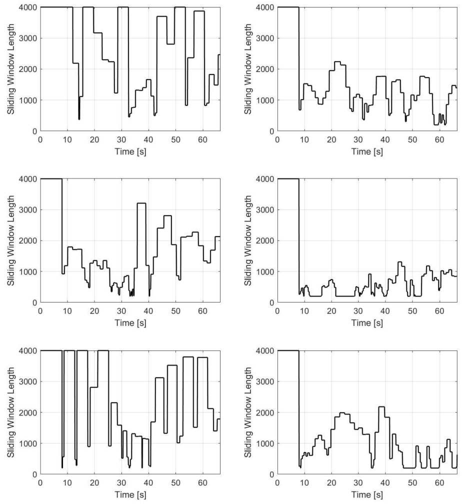


Fig. 6. Case 1: The sliding window lengths for each signal at the given time stamps.

used. Applying a small window length for signals with approximately equal variances results in poor fused signal quality, when the difference between the input signals increases. This is shown in the position signals. For orientation signals with clearly different variances (due to the noise level), the fused output strictly emphasizes the better-quality SLAM orientation signal. However, the small window length still induces some of the noise in the fused output. Fig. 8 shows the same results, while the set maximum window length of $N_{max} = 4000$ is used. In this case, the resulting fused signals are more restrained. Finally, Fig. 9 shows the results for the variable sliding window lengths. As illustrated, the window lengths first float between the set minimum and maximum values. However, after the SLAM signals are artificially increased, the window lengths jump to the set maximum values. The difference compared with the previous case of using the set maximum window length is that with variable lengths the computations can be performed over a smaller number of elements. Moreover, as demonstrated in Figs. 7–9 before the 30 s marks, the fused signals are slightly better with smaller window lengths due to the small difference between the input measurements. Thus, the absolute mean difference between the signals, computed over the current sliding window, was chosen as the basis for updating the window length in Eq. (7).

For Case 3, the aim was to assess the fusion algorithm's performance when a measured signal is lost, and transition smoothing is required. Smoothing is applied to avoid sudden, undesired changes with large amplitudes in the fused output signal. The results are illustrated in Fig. 10, in which, at 30 s, the SLAM position signals were again artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°, respectively. Then, the SLAM measurement was switched off, after which the fusion algorithm switched to utilize only the marker tracking-based measurement. Then, the SLAM measurement was switched back on, and the fusion algorithm resumed to utilize both pose measurements. For comparison, the same procedure was repeated by switching the marker tracking off and on. The red lines are the

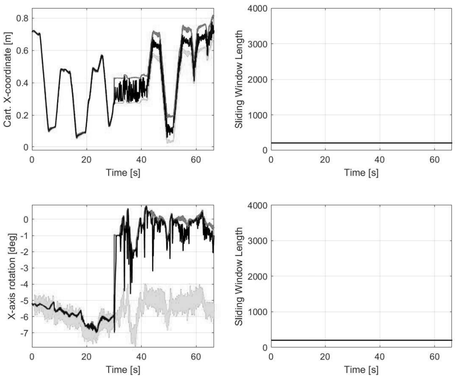


Fig. 7. Case 2: A constant sliding window length of 200 was used: At 30 s, the SLAM position signals were artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°. The left figures show the signals, and the right figures show the respective sliding window lengths. On the left, the red lines are the SLAM signals, the green lines are the marker-tracking signals, and the black lines are the fused signals.

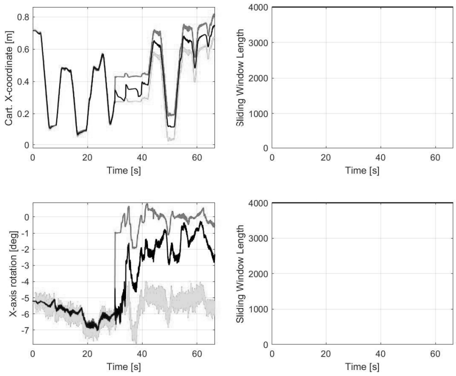


Fig. 8. Case 2: A constant sliding window length of 4000 was used: At 30 s, the SLAM position signals were artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°. The left figures show the signals, and the right figures show the respective sliding window lengths. On the left, the red lines are the SLAM signals, the green lines are the marker-tracking signals, and the black lines are the fused signals.

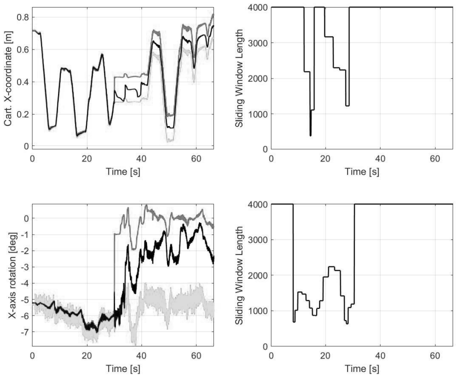


Fig. 9. Case 2: Variable sliding window lengths were used: At 30 s, the SLAM position signals were artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°. The left figures show the signals, and the right figures show the respective sliding window lengths. On the left, the red lines are the SLAM signals, the green lines are the marker-tracking signals, and the black lines are the fused signals.

SLAM signals, the green lines are the marker-tracking signals, the black lines are the fused signals with transition smoothing, and the magenta lines are the fused signals without transition smoothing. The switching time $t_2 - t_1$ in Eq. (10) was 0.5 s, which dictated the desired convergence time toward the available pose measurement, and the "jump" occurs when the error coefficient ϵ in Eq. (8) is reached. As shown, the transitions are appropriately smoothed, and the effectiveness can be tuned by adjusting the parameters.

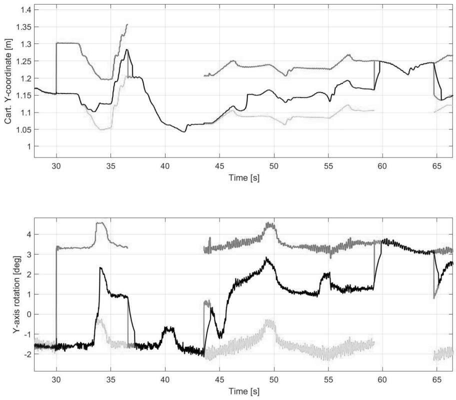


Fig. 10. Case 3: At 30 s, the SLAM position signals were artificially increased by 0.15 m, and the SLAM orientation signals were increased by 5°. The transitioning of the fused signal is demonstrated when the other measurement signal is lost. The red lines are the SLAM signals, the green lines are the marker-tracking signals, the black lines are the fused signals with transition smoothing, and the magenta lines show the fused signals without transition smoothing.

The error coefficients have to be carefully set, as values that are too low will have a deteriorating effect on the fusion output; the transition smoothing should enable only when sudden, undesired changes occur in the input measurements. However, values that are too large will render the smoothing ineffective.

5. CONCLUSION

In this paper, we examined the problem of directly fusing continuous sensor data in a real-time setting. The presented model-free pipeline is a statistical approach based on weighted averaging, in which the weight parameters are constantly updated using the sliding window variances of the respective signals to be fused. A method for updating the window length was shown, along with a simple transition smoothing design.

Results based on real-time experiments were presented: 6-DOF pose trajectory data from two independent, redundant visual sensors were fused in an optimal manner in the sense that the variances of the fused signals were minimized with respect to the input variances, computed over the current sliding windows. The experimental results demonstrated that the proposed methodology can increase the system's robustness and fault tolerance, which are the desired features for future autonomous field robotic machines.

Some challenges of this methodology include the lack of a model, which makes the system rely more on sophisticated sensor self-diagnostics before fusion occurs, so that faulty measurements are detected and discarded before fusion is executed.

ACKNOWLEDGEMENT

This work was supported by the Doctoral School of Industry Innovations (DSII) of Tampere University. This work was also carried out with the support of the Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland.

REFERENCES

[1] T. Machado, D. Fassbender, A. Taheri, D. Eriksson, H. Gupta, A. Molaei, P. Forte, P. K. Rai, R. Ghabcheloo, S. Mäkinen, A. J. Lilienthal, H. Andreasson, and M. Geimer, "Autonomous heavy-duty mobile machinery: A multidisciplinary collaborative challenge," in *2021 IEEE International Conference on Technology and Entrepreneurship (ICTE)*, pp. 1–8, 2021.

[2] L. Lopes, T. Miklovicz, E. Bakker, and Z. Milosevic, "The benefits and challenges of robotics in the mineral raw materials sector-an overview," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1507–1512, 2018.

[3] J. A. Marshall, A. Bonchis, E. Nebot, and S. Scheding, "Robotics in mining," in *Springer handbook of robotics*, Springer, pp. 1549–1576, 2016.

[4] J. Kocić, N. Jovičić, and V. Drndarević, "Sensors and sensor fusion in autonomous vehicles," in *2018 26th Telecommunications Forum (TELFOR)*, IEEE, pp. 420–425, 2018.

[5] H. F. Durrant-Whyte, "Sensor models and multisensor integration," in *Autonomous Robot Vehicles*, Springer, New York, pp. 73–89, 1990.

[6] S.-L. Sun and Z.-L. Deng, "Multi-sensor optimal information fusion Kalman filter," *Automatica*, vol. 40, no. 6, pp. 1017–1023, 2004.

[7] J. Gross, Y. Gu, S. Gururajan, B. Seanor, and M. Napolitano, "A comparison of extended Kalman filter, sigma-point Kalman filter, and particle filter in GPS/INS sensor fusion," in *AIAA Guidance, Navigation, and Control Conference*, p. 8332, 2010.

[8] E. Bostanci, B. Bostanci, N. Kanwal, and A. F. Clark, "Sensor fusion of camera, GPS and IMU using fuzzy adaptive multiple motion models," *Soft Computing*, vol. 22, no. 8, pp. 2619–2632, 2018.

[9] S. Yazdkhasti and J. Z. Sasiadek, "Multi sensor fusion based on adaptive Kalman filtering," in *Advances in Aerospace Guidance, Navigation and Control*, Springer, Cham, pp. 317–333, 2018.

[10] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[11] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, p. 4220, 2020.

[12] W. Elmenreich, "Fusion of continuous-valued sensor measurements using confidence-weighted averaging," *Journal of Vibration and Control*, vol. 13, no. 9–10, pp. 1303–1312, 2007.

[13] P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Redundancy-based visual tool center point pose estimation for long-reach manipulators," in *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 1387–1393, 2020.

[14] P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Probabilistic camera-to-kinematic model calibration for long-reach robotic manipulators in unknown environments," in *2022 IEEE 17th International Conference on Advanced Motion Control (AMC)*, pp. 48–55, 2022.

[15] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[16] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[17] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 42, no. 1, pp. 97–101, 2000.

UNPUBLISHED MANUSCRIPT
V

Vision-aided precise positioning for long-reach robotic manipulators using local calibration

P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila