Tampere University

LONGCHUAN NIU

Improving the Visual Perception of Heavy Duty Manipulators in Challenging Scenarios

Tampere University Dissertations 355

Tampere University Dissertations 355

LONGCHUAN NIU

Improving the Visual Perception of Heavy Duty Manipulators in Challenging Scenarios

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Engineering and Natural Sciences
of Tampere University,
for public discussion in the K1702
of the Konetalo, Korkeakoulunkatu 6, Tampere,
on 4th of December, 2020, 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 Lasse Lensu LUT University Finland	Dr. Mika Vainio Aalto University Finland
Opponent	Professor Arto Visala Aalto University Finland	

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

Cover design: Roihu Inc.

ISBN 978-952-03-1796-6 (print)
ISBN 978-952-03-1797-3 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1797-3

PunaMusta Oy – Yliopistopaino
Vantaa 2020

PREFACE

The study presented in this dissertation was carried out from 2016 to 2020 within the Unit of Automation Technology and Mechanical Engineering of the Faculty of Engineering and Natural Sciences at Tampere University.

I would like to express my deepest gratitude to my supervisor, Prof. Jouni Mattila, for his encouragement and for providing the opportunity to work on this dissertation.

I want to thank the preliminary examiners Prof. Lasse Lensu and Dr. Mika Vainio for their evaluation of this dissertation and for their professional feedback. I am also grateful to Prof. Arto Visala for agreeing to be my opponent.

I would like to acknowledge all staff at the Faculty of Engineering and Natural Sciences and the Faculty of Information Technology and Communications. Special thanks to Prof. Ke Chen, Dr. Mohammad M. Aref, M.Sc. Olli Suominen, M.Sc. Sergey Smirnov, M.Sc. Santeri Lampinen and M.Sc. Liisa Aha, I enjoyed the moment working with them.

The funding from the EUROfusion foundation and the Faculty of Engineering and Natural Sciences is greatly appreciated; this study could not have been completed without it.

Finally, I would like to thank my parents for their continuous support and express my wishes for my mother's swift recovery from a cerebral infarction. I love you.

Longchuan Niu
November 2020

ABSTRACT

Robotic vision is a subfield of computer vision intended to provide robots with the capability to visually perceive the surrounding environment. For example, a robotic manipulator leverages its visual perception system to gather visual data through cameras and other sensors, then uses that input to recognize different objects in order to safely perform an autonomous operation.

However, in many robotics applications, robots have to face a cluttered and dynamic scene, where classic computer vision algorithms show the limitation of tackling the environmental uncertainty. Such scene understanding requires a fusion of traditional and modern approaches involving classic computer vision, machine learning and deep learning methods.

This thesis examines visual perception challenges in remote handling and the mining industry. It begins with two research questions: Can the robustness of target-object pose estimation be improved in challenging real-world, heavy-duty robotic scenarios? Can fast detection and localization for objects be obtained without prior known geometry in a scenario with piles of overlapping objects? Six publications[1] cover the methods from algorithm design to system-level integration used to solve real-world problems.

In the ITER fusion reactor, the operator teleoperates a robotic manipulator to perform maintenance tasks amidst a high level of noise and erosion. The operator cannot fully rely on the virtual reality (VR) system, which may not reflect the current scene accurately, as physical conditions may have changed in the harsh environment. Meanwhile, every operation inside the reactor requires robust, millimeter-level accuracy. This thesis analyzes research questions and presents a novel edge-point iterative closest point (ICP) method as a solution for target-object detection, tracking and pose estimation. Using the knuckle of a divertor cassette as an example, the overall accuracy of the developed visual system meets ITER requirements, and

[1] Five of the publications are listed as contributors to this thesis, one publication, [1], as a citation.

the conducted experiments with the manipulator demonstrated the efficiency of the method.

Smartbooms2 is a project in the mining industry that requires a heavy manipulator with a hydraulic hammer to autonomously break rocks in a cluttered outdoor environment. Based on the output data of the three-dimensional (3D) sensors, several solutions are proposed. Examining a popular time-of-flight (TOF) sensor, this thesis explores state-of-the-art unsupervised machine learning methods and proposes a novel clustering method. Using an industrial stereo camera, this thesis proposes a novel 3D rock detection and localization pipeline. The results and system accuracy are detailed in published research papers.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

CAD	Computer-Aided Design
CLS	Cassette Locking System
CNN	Convolutional Neural Network
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DOF	Degrees of Freedom
DTP2	Divertor Test Platform 2
GMM	Gaussian Mixture Model
ICP	Iterative Closest Point
ITER	International Thermonuclear Experimental Reactor
LIDAR	Light Detection and Ranging
R-CNN	Region-Based Convolutional Neural Networks
RANSAC	Random Sample Consensus
RGB	Red, Green, Blue
RH	Remote Handling
RHCS	Remote Handling Control System
RMS	Root Mean Square
SIFT	Scale Invariant Feature Transform
SSD	Single Shot Multibox Detector
SURF	Speeded-Up Robust Features
SVD	Singular Value Decomposition
TCP	Tool Center Point

TOF	Time-of-Flight
UDP	User Datagram Protocol
VPS	Visual Perception System
VR	Virtual Reality
VTT	Technical Research Centre of Finland
WARD	Ward’s Minimum Variance Method
YOLO	You Only Look Once

LIST OF PUBLICATIONS

P-I **L. Niu**, S. Smirnov, J. Mattila, A. Gotchev and E. Ruiz. Robust pose estimation with a stereoscopic camera in harsh environments. IS&T International Symposium on Electronic Imaging. 2018, Vol. 2018.9, 126–1.

P-II **L. Niu**, M. M. Aref and J. Mattila. Clustering Analysis for Secondary Breaking Using a Low-Cost Time-of-Flight Camera. 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP). IEEE. 2018, 318–324.

P-III **L. Niu**, K. Chen, K. Jia and J. Mattila. Efficient 3D Visual Perception for Robotic Rock Breaking. 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE). IEEE. 2019, 1124–1130.

P-V **L. Niu**, L. Aha, J. Mattila, A. Gotchev and E. Ruiz. A stereoscopic eye-in-hand vision system for remote handling in ITER. Fusion Engineering and Design, 2019, Vol. 146, 1790–1795.

UNPUBLISHED MANUSCRIPT

P-IV **L. Niu**, S. Lampinen, L. Hulttinen, J. Niemi, and J. Mattila. "Autonomous Robotic Rock Breaking Using a Real-time 3D Perception System", 2020.

1 INTRODUCTION

Robots are widely used in industrial work cells, as they can provide superior quality, speed and accuracy in highly repetitive tasks. In these relatively fixed and structured industrial settings, robots can be pre-programmed for consistent task performance and used in 24/7-style operations. Another important benefit in heavy-duty applications and in hazardous environments is that the use of remotely controlled robots can improve safety. Moreover, vision-generated guidance information can allow robots to select and vary their motion targets, thus enabling more flexible automation systems. Vision-based robotic systems incorporate techniques from optics, image processing, computer vision, and machine learning. Unlike pure computer vision research, vision-based robots must incorporate many aspects of robotics (such as robot kinematics, reference-frame calibration, hand-eye calibration, and robotic control algorithms) into an integrated control system that enables stable physical interaction with the manipulated objects.

In this thesis, two challenging real-world scenarios that could benefit from vision-based robotic system development are considered. The two scenarios, namely ITER and Smartbooms2, are described in more detail in the following sections.

1.1 ITER

The International Thermonuclear Experimental Reactor (ITER) is the world's largest fusion experiment[1]. The goal of ITER, currently under construction in Cadarache, France, is to demonstrate the scientific and technological feasibility of fusion energy. Reactor lifecycle management that allows for continuous operation is one of the main development challenges, due to the extreme conditions inside the reactor vacuum vessel during its operation (high temperatures, magnetic fields and radia-

[1] https://www.iter.org

tion). Due to reactor material erosion and impurities resulting from nuclear fusion, such as helium ash, the reactor is subject to scheduled maintenance (Figure 1.1), but since the vacuum-vessel radiation levels make human access impossible, teleoperated robots need to be employed. One such remote maintenance operation is the scheduled replacement of the lower part of the reactor, called the divertor, which consist of 54 modular elements called cassettes, each weighing approximately 10 tonnes. Research on ITER maintenance by use of heavy-duty robots has been carried out in divertor test platform 2 (DTP2) in Finland over the past two decades. Tampere University has acted as a major in this ITER remote handling (RH) research organisation, together with the Technical Research Centre of Finland (VTT). RH primarily utilizes conventional teleoperated robotic manipulators in the man-in-the loop style, relying on skilled operators. Such work requires a high level of concentration and precision and does not allow any room for mistakes. Hence, its cognitive burden on the teleoperator can be very high, which affects operator alertness negatively and can consequently increase accident-proneness due to human errors. Development of a user-assisting 3D visual perception system (VPS) was furthered to enhance safety and performance in RH operations. However, the ITER reactor is subject to numerous challenges, such as very confined space with limited camera field of view and illumination, possible drifting in object positions due to a high magnetic field and extreme temperatures, a high level of radiation and resulting increase in camera noise over time, non-Lambertian reflectance of reactor elements on polished steel surfaces, and millimetre-level clearances in operations. These harsh conditions in the ITER are major challenges to the development of a VPS, even if the environment is structured. Therefore, the first research objective is to find solutions to this problem.

1.2 Smartbooms2

The second research challenge studied is autonomous secondary rock breaking, which was one of the research objectives in a Business Finland project called Smartbooms2, a joint collaboration between Finnish companies Rambooms, Technion and Novatron and Tampere University. In the mining and construction industries, secondary rock breaker manipulators, as shown in Figure 1.2, are used extensively. These human-operated manipulators are equipped with impact hammers, and their task

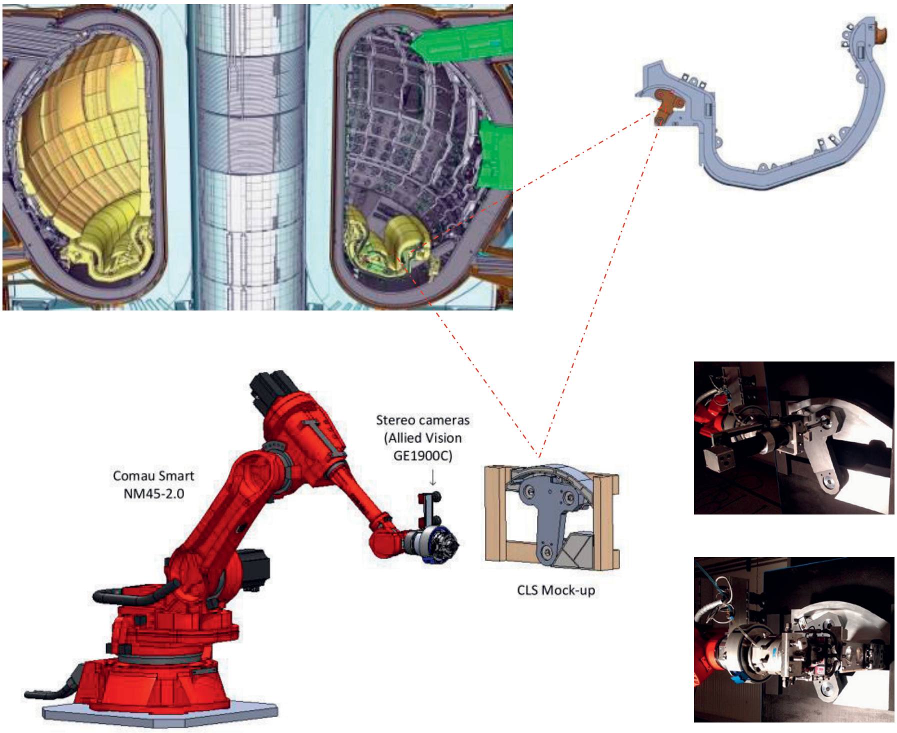


Figure 1.1 The knuckle of diverter cassettes in the ITER reactor.

is to detect oversized rocks on a mineral crusher grader plate and to break them. Such work is event-based, and thus a fast response is only needed when oversized rocks are encountered. What is more, the work environments in underground deep mining can be dangerous. This field robotics scenario in an outdoor environment is fairly structured in terms of known scene dimensions and fixed camera eye-to-hand position, however, the detected object shapes are arbitrary, and they can form overlapping piles of objects. The second research objective of this thesis is to provide solutions to automatize this event-based rock breaking process.

1.3 Research Problems (RPs)

The challenges of vision-based robotic manipulator systems are often centered around the problem of acquiring an accurate 6 DOF pose estimate (Cartesian 3 DOF posi-

Figure 1.2 Rock breaking at mining sites [2],[3].

tion and 3 DOF orientation) of objects of interest. This is essential in both man-in-the-loop type teleoperation tasks and in autonomous robotic manipulation tasks. In the former, the operator benefits if the target-object pose with respect to the controlled robot tool center point (TCP) is displayed for his/her guidance. Whereas in the latter case, the robot is automatically controlled by visual features extracted from an image of the target object.

Moreover, for overall robotic system performance, the object pose estimate has to be converted into a robot coordinate system, which involves eye-in-hand or eye-to-hand calibrations depending on camera location. Also, for real-world, vision-based robot control scenarios, the vision-based system should provide the pose estimate at a level of 10 Hz or higher.

The two RPs addressed in this thesis are as follows:

RP.I: Can the robustness of target-object pose estimation be improved in challenging real-world, heavy-duty robotic scenarios?

RP.II: Can fast detection and localization for objects be obtained without prior known geometry in a scenario with piles of overlapping objects?

1.4 Requirements and Research Scope

Apart from the scientific challenges, the requirements and research scope for robotic virtual perception are associated with the tasks that heavy-duty manipulators perform.

In ITER, the replacement of divertor cassettes requires thousands of tool operations on the cassette locking system (CLS). Such maintainance tasks are performed through RH, where the operator remotely operates a heavy-duty robotic manipulator with the aid of several tools: a jack tool, a pin tool and a wrench tool. The clearance between the tool and the CLS for each operation can be as low as 3 mm. The acquired images from radiation-tolerant cameras are grayscale and low resolution and might contain unexpected sensing noises from the harsh environment.

For safe and accurate RH operations, robust and precise pose estimation of the CLS parts is required. The major goal of the project is to design a generic 3D perception vision system that meets ITER RH requirements.

To achieve maximum accuracy, the design of the vision system has to take the following factors into account:

- Selection of camera for 3D perception.
- Camera calibration methods.
- Robust depth from stereo method.
- Selection of extrinsic camera configuration for higher accuracy.
- Design of a robust 6 DOF pose estimation method.

In view of these factors, a stereo pair of close-range cameras has been deemed essential, and a plane-sweeping method has been employed for the precise 3D reconstruction of target objects in dense point cloud. For maximum accuracy and an occlusion-free field of view, eye-in-hand camera configuration has been adopted. In the most prominent part of the study, a novel edge-point ICP method has been proposed for robust pose estimation of the knuckle P-I. The methods and process for intrinsic and eye-in-hand (extrinsic) calibration are detailed in P-V and my paper [1].

A high-level architecture of the remote handling control system (RHCS) with a stereoscopic vision system is presented in Figure 1.3, where the RHCS is built based on a restructuring of the old infrastructure [4]. In the experimental setup, the Co-

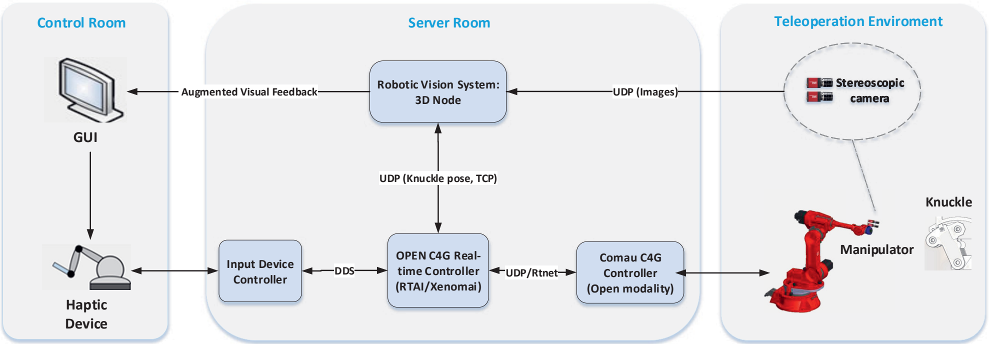


Figure 1.3 Top-level architecture of the remote handling system with stereoscopic visual perception.

mau NM-45 is attached with a pair of stereo cameras to its end-effector. The RHCS is composed of a real-time determinstic OpenC4G controller acting as an equipment controller for data communication with the vision system and other systems, a real-time input device controller for a haptic device, a real-time Comau C4G controller operating in open modality mode, and a tool exchanger. The RH manipulator is teleoperated through a 6 DOF haptic device that drives the stereo camera closer to the target object knuckle, and the pose estimation of the kunckle is performed in a replicated ITER environment [P-V].

The RH environment enables further development of advanced control schemes for semi-automatic manipulation tasks. For example, the robust estimated pose of a target can guide the manipulator to move a tool automatically to the aligned pose, such that the operator can finalize the tool insertion process by driving the tool along the z-axis of the manipulator's tool coordinate.

In Smartbooms2, the research scope is the deployment of a 3D VPS for autonomous rock breaking. Autonomous rock breaking includes three steps: perception, decision making and robotic controls. The first step is scene understanding, and the goal is to detect and localize each individual rock in a cluttered scene. In the second step, decisions on the order of rock breaking can be classified according to the size the rock, the height of the rock surface or the manipulator TCP position. Among these criteria, the manipulator TCP position with respect to each rock is adopted for experimental trajectory planning. The last step is to follow the calculated trajectory, keep the impact hammer in a given pose, and maintain pressure against the rock.

A high-level architecture of the system is depicted in Figure 1.4. The research focus is on the design of an accurate, real-time 3D VPS.

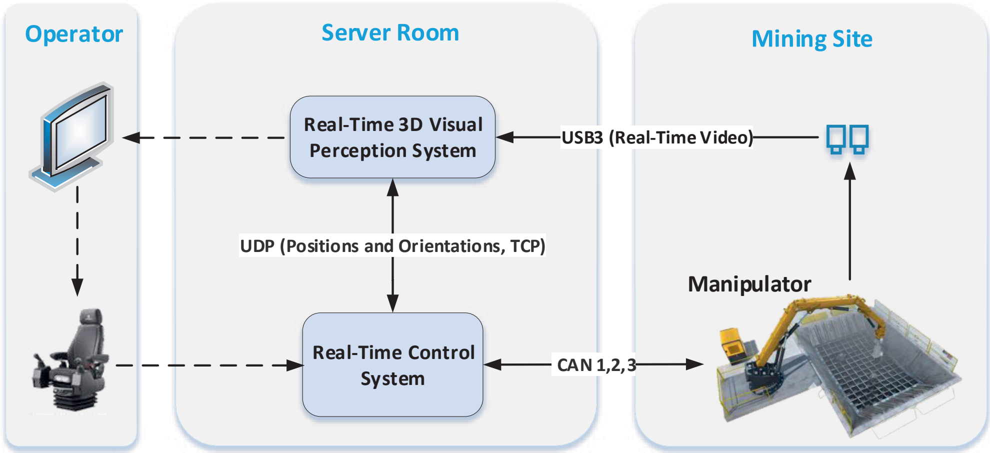


Figure 1.4 Top-level architecture of the autonomous rock breaking system.

It is obvious that rocks collected from mining sites can not be characterized by a particular feature. They possess a variety of colors, unique surface textures and arbitrary geometries (shapes and sizes). The design of a robotic perception system should take account of the following challenges:

- Hardware selection and component setup, e.g., vision sensor for 3D perception, graphics processing unit (GPU), setup plan, etc.
- Lenses are fragile in close-range hazardous rock breaking operations.
- Ability to tackle unpredictable ambient light under outdoor dynamic illumination conditions.
- Design of a fast, robust and accurate 3D object detection pipeline.
- Detailing a method that can accurately detect all rocks in a scene, including ones that have been occluded by overlapping rocks or the manipulator arm.
- Determine the appropriate position of each rock to break.
- Ability to infer from the surface above the breaking position the appropriate angle for breaking.

This research has been carried out in two phases. In the first phase, an IFM TOF camera[2] as the 3D sensor was chosen for its popularity across the industry [P-II].

[2] https://www.ifm.com/products/ae/ds/O3M150.htm

And in the second phase, a ZED stereo camera[3] was adopted to continue our study in rock detection [P-III]. Based on the experience gained, a generic 3D visual perception pipeline for autonomous rock breaking is proposed in P-IV.

To summarize, in ITER and Smartbooms2, the target objects can be either known objects or unknown objects. "Known objects" refers to objects with known features, i.e., shape, size, color, surface texture, etc. However, in real-world scenes, there are more objects that are not known in advance; they may appear in an arbitrary order and come with unknown features. For example, rocks are objects with arbitrary characteristics. The research problems for these two categories of objects are characterized in Table 1.1.

Table 1.1 Research Scope for Remote Handling and Autonomous Secondary Breaking

	Remote Handling	*Autonomous Secondary Breaking*
Control Scheme	Man in the Loop	Closed-Loop
Camera Setup	Eye-in-Hand	Eye-to-Hand
Object Information	Known, Single	Unknown, Arbitrary

1.5 Thesis Contributions

As the main contributions of this thesis, novel methods and their related deployment in 3D VPSs are presented to facilitate robotic tasks of heavy-duty manipulators in challenging scenarios. The publication contributions are summarised as follows:

P-I The paper proposes an algorithmic improvement of the edge-point ICP method for fine alignment of the sensed point cloud with the reference point cloud in harsh conditions. This novel method significantly enhanced the robustness of point cloud registration in challenging ITER environments. As a consequence, an accurate 6 DOF pose estimate of a target object can be achieved. Given the divertor CLS as the target object, the repeatability test demonstrated its consistent performance in terms of the number of outliers and precision of pose estimation.

[3] `https://www.stereolabs.com/zed/`

P-II A 3D VPS is the key enabler of autonomous robotic secondary breaking. This paper aim was to discover the most appropriate clustering methods for rock detection using the IFM TOF camera. The study began by exploiting the existing start-of-the-art clustering methods for this task and found out that Ward's minimum variance method (WARD) and density-based spatial clustering of applications with noise (DBSCAN) perform better than the rest. Nevertheless, these two methods still have issues in rock detection. To this end, a novel Euclidean clustering algorithm was proposed based on the spatial characteristics of the TOF camera. The conducted experiments revealed that the proposed method has better robustness and overall performance compared to DBSCAN. The method, in contrast with WARD, does not require manual adjustments of parameters while preserving performance. The paper also highlights the constraints of the research due to the limitations of the TOF camera.

P-III This paper continued the previous work in P-II. In order to acquire the rich features of a scene, an industry-ready ZED stereo camera was adopted for its capability to provide high resolution images and dense point clouds. The study focused on deep learning approaches for their advancements in scene understanding. An accurate and fast 3D rock detection method was proposed based on the infrastructure of You Only Look Once version 3 (YOLOv3). The paper also presents methods for accurate 3D reconstruction, 3D position and 3D surface normal estimation of detected rocks. The overall performance of the 3D object detection mechanism was validated by offline videos. For example, conducting the experiment with a video where 12 rocks were overlapping to each other in a pile. The proposed 3D object detection method exhibits versatility in the detection of unstructured objects within a structured environment.

P-IV The paper presents a novel autonomous robotic secondary breaking system, which is an extension of the work in P-III. The 3D VPS, as the key enabler of autonomous operation, was further developed in the direction of commercial settings. The 3D visual perception pipeline was refined to resolve challenges in secondary breaking experiments. The paper addressed the following details: The accuracy and robustness of the rock detection model were improved through training with a larger image data set, with the aid of data augmentation; The ZED camera's intrinsic and extrinsic calibrations were performed, and the precision of the system was validated; The schemes for determining positions and orientations for rock breaking were revised, which significantly improved the success rate of breaking; A real-time 3D viewer was implemented to visualize the detected rock in 3D and validate the correctness and effectiveness of the positions for breaking online; Implementation of data analysis, processing, validation, rendering and communication modules. The experiments were conducted in a real-world setup with a commercial heavy-duty manipulator, that yielded an average rate of 96.41% for rock detection, 11.76 Hz for detection speed, and autonomous rock breaking attempts of 3.3 per minute. These results suggest the advancement of VPS for the productive robotized operation and its readiness to be employed in the mining industry.

P-V The paper leverages the idea of the novel edge-point ICP method in P-I for development of the eye-in-hand stereoscopic vision system with the RHCS. Together with [1], the paper presents the design of the software and hardware architecture of the vision system, the implementation of different operation modes of the vision system, the structure of the software and hardware architecture of the RHCS, the CLS tools, the calibration of system components (manipulator, intrinsic and extrinsic camera parameters, and cassette operation tools), and the communications between each heterogeneous subsystem. Overall, the proposed system can tackle the challenging requirements in ITER application, such as a constraint on image acquisition with low-resolution and grayscale radiation tolerant camera, high level of image noises due to the radiation, non-Lambertian reflectance of reactor elements on shiny metallic surfaces, and deficient illumination of the scene due to constraints on available light sources. Successfully conducting RH experiments in a replicated ITER environment with only a three-millimeters clearance shows that the developed system met the application requirements.

1.6 The Author's Contribution

This section briefly explains the role of the author in each of the listed publications.

P-I The author conceived the idea of the eye-in-hand stereo vision system, contributed to the development of the vision system, integrated the remote handling control system with the vision system, and wrote the paper. M.Sc. Sergey Smirnov helped with depth-map creation, 3D reconstruction development and edited the paper. Professor Jouni Mattila edited the paper. Professor Atanas Gotchev reviewed the paper. Dr. Emilio Ruiz provided an evaluation of the system.

P-II The author conceived the idea, developed the clustering algorithm, and wrote the paper. Dr. Mohammad M. Aref helped edit the paper. Professor. Jouni Mattila reviewed the paper and made corrections.

P-III The paper was written during TUT Mobility (TUT on World Tour 2018) from Oct 2018 to Feb 2019. The author developed the 3D object detection system and wrote the paper. Professor Ke Chen contributed to the discussion of the results and helped edit the paper. Professor Kui Jia contributed to the discussion of the results and provided research facilities. Professor Jouni Mattila reviewed the paper and made corrections.

P-IV The author and M.Sc. Santeri Lampinen contributed equally to the paper. The author conceived the idea of the deep learning-based visual perception system, implemented 3D object detection mechanism, calibrated vision system, and designed the method for detecting suitable breaking position and orientation to enable autonomous robotic control. M.Sc. Santeri Lampinen designed the manipulator control system and managed the implementation. M.Sc. Lionel Hulttinen calibrated the manipulator's forward kinematics model and wrote the corresponding part of the paper. Mr. Jouni Niemi provided industrial insight and views regarding rock breaking, including system evaluation. Professor Jouni Mattila reviewed the paper and made corrections.

P-V The author calibrated the system (robot, tools and camera extrinsic), developed RHCS, deployed the vision system, integrated the stereoscopic vision system to the RHCS, and wrote the paper. M.Sc. Liisa Aha edited the paper. Professor Jouni Mattila reviewed the paper and made suggestions. Prof. Atanas Gotchev reviewed the paper. Dr. Emilio Ruiz evaluated the performance of the whole system.

1.7 Outline

This compendium thesis is comprised of five chapters. The arrangement of chapters and publications is illustrated in Figure 1.5.

Chapter 1 introduces the research problems, scope of the research, and contributions.

Chapter 2 presents state-of-the-art methods in the robotic visual perception field: 3D reconstruction, random sample consensus (RANSAC), and extrinsic camera configurations. Subsequently, the literature review of robotic visual perception is presented in two categories, known and unknown objects. The proposed methods are assessed with state-of-the-art methods in terms of application-specific evaluation metrics.

Chapter 3 consists of summaries of each of the five publications. This chapter explains the connections between the thesis research problems and the publications.

Chapter 4 categorizes discussions into four subjects. The first two subjects feature discussions of the research problems. The other two subjects feature discussions of common extrinsic camera configurations and various factors that influence the precision of a vision system.

Chapter 5 presents the research conclusions based on observations and experiment results and answers the research problems. The last part of the chapter addresses future research.

The publications P-I, P-II, PIII, P-IV and P-V are appended at the end of this thesis.

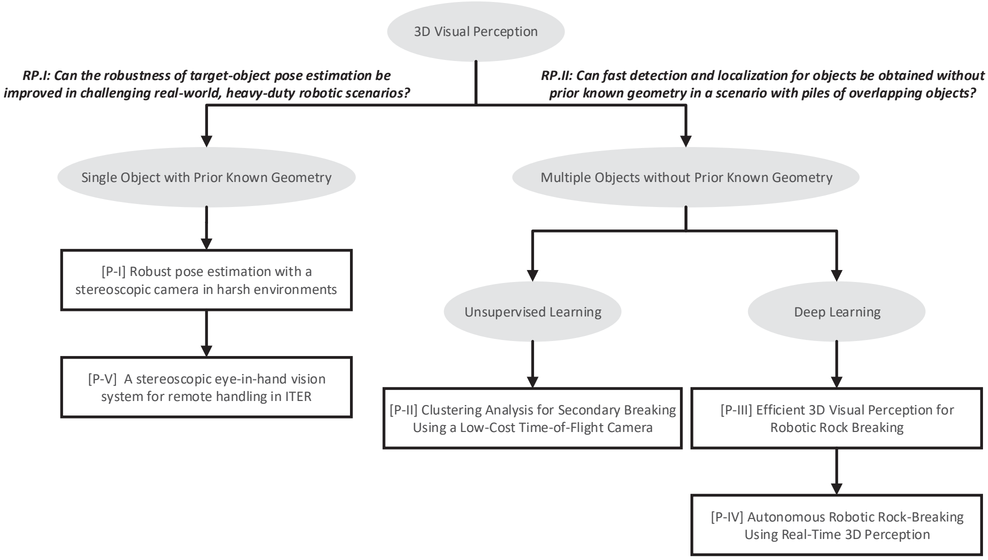

3D Visual Perception
RP.I: Can the robustness of target-object pose estimation be improved in challenging real-world, heavy-duty robotic scenarios?
RP.II: Can fast detection and localization for objects be obtained without prior known geometry in a scenario with piles of overlapping objects?
Single Object with Prior Known Geometry
Multiple Objects without Prior Known Geometry
[P-I] Robust pose estimation with a stereoscopic camera in harsh environments
Unsupervised Learning
Deep Learning
[P-V] A stereoscopic eye-in-hand vision system for remote handling in ITER
[P-II] Clustering Analysis for Secondary Breaking Using a Low-Cost Time-of-Flight Camera
[P-III] Efficient 3D Visual Perception for Robotic Rock Breaking
[P-IV] Autonomous Robotic Rock-Breaking Using Real-Time 3D Perception

Figure 1.5 Thesis structure.

2 STATE-OF-THE-ART ROBOTIC VISUAL PERCEPTION

Robotic vision incorporates techniques from optics, image processing, computer vision, machine learning and deep learning. Unlike pure computer vision research, robotic vision must incorporate a variety of aspects of robotics into its techniques and algorithms, such as reference frame calibration, kinematics, camera to robot extrinsic calibration and the robot's ability to physically affect the environment. This chapter reviews 3D reconstruction from stereo vision, object detection and pose estimation approaches used in robotic perception. Objects in all-purpose robotic applications can be classified of known and consistent geometry or unknown, scalable and varied geometry. Here, some state-of-the-art 3D object detection and pose estimation methods are presented related to the ITER and Smartboom2 projects.

2.1 From 2D-Image-Coordinate-System to 3D-World-Coordinate-System: Scene Restoration

A red, green, and blue (RGB) image produced by a pinhole camera is a projection from a 3D scene onto 2D plane; in this dimension-reduction process, depth information is lost. The reverse process is to infer the 3D geometry and structure of the scene from images, which is known as 3D reconstruction. Knowing the pose of the camera with respect to the robotic manipulator base, the restored 3D scene can help a robot to understand the profiles of objects in a scene and their positions in the robot coordinate system.

2.1.1 3D Reconstruction from Stereo Images

3D reconstruction from a stereo camera image pair consists of the following steps:

- Geometric camera calibration: Camera calibration can be divided into two individual steps, intrinsic and extrinsic calibration [5]. Apart from the correction of lens distortions and finding inherent camera parameters, the goal of intrinsic calibration in stereoscopic vision is to determine the geometric relationship between a point position in a camera coordinate and its projected position in the corresponding left and right image. Extrinsic calibration in stereoscopic vision reflects how a stereo camera is positioned in the world coordinate system. In robotic manipulator applications, the extrinsic parameters are determined by the pose of the left eye of the camera with respect to the base of the manipulator. Section 2.1.3 has the details of application scenarios.

- Image acquisition: The pair of images for a scene or object are simultaneously acquired by the left and right eye of a stereo camera. The result is two disparate camera images of the same scene, like human sight. The acquisition step has to ensure illumination invariance between image pairs.

- Depth from stereo: Estimation of scene geometry from a stereoscopic camera is called stereo matching or the depth from stereo problem. Stereoscopic vision uses the binocular disparity between two camera images for depth estimation. Conventional depth-from-stereo methods are based on stereo-image rectification [6], which in some cases might underperform due to the introduction of artificial camera transformation and excessive image interpolation steps. Moreover, a deviation from a geometrically parallel camera configuration is possible (e.g., the camera's optical axes might be crossed), thus introducing substantial image deformation. In comparison, the plane-sweeping method [7] allows direct processing of the captured imagery via calibrated camera parameters for the generation of a depth map, and it does not require rigorous geometrically parallel camera configuration. For maximized robustness and accuracy of the vision system, P-I adopted this approach.

- 3D restoration: The 3D geometry of a scene is reconstructed according to the intrinsic parameters of a calibrated camera and the depth map of a stereo camera. The accuracy of 3D reconstruction depends on the precision of stereo correspondence and camera calibration methods.

2.1.2 3D Data Preprocessing: RANSAC

The lens of a pinhole camera introduces radial distortion. On the other hand, unexpected noises from harsh environments may introduce outliers to a generated 3D point cloud. These outliers, which could severely influence a performance of a vision system, can be largely excluded by applying RANSAC [8], on account of its robust adaptive solutions for different noises as compared to simple thresholding methods. RANSAC is a learning technique to estimate parameters of a model from the given data, which contain both inliers and outliers; points belonging to that model are considered inliers. The algorithm starts by randomly picking minimum number of points needed to form a sample to initialize the model. Then it gets the consensus set with the points within error bounds, i.e. the distance threshold. It repeats this until a good model is found, which contains the most inliers. RANSAC works for general models, which require a minimal set, the smallest set from which the model can be computed. The algorithm can be terminated either by reaching a big enough consensus set or by repeating it N times and then returning the model with the biggest set.

RANSAC is a frequently used algorithm with real-world sensor data, as it is robust enough for large numbers of outliers in noisy imagery. In the ITER application, RANSAC was applied to remove outliers by distance threshold, thus inliers of CLS component's model were acquired [P-I]. In the Smartbooms2 project, RANSAC was adopted to find the best fitting plane on the surface of each rock ([P-II], P-III] and [P-IV]).

2.1.3 From 3D Camera Coordinate System to 3D Robot World Coordinate System

A 3D scene is reconstructed in a camera coordinate system by using intrinsic camera parameters. A point in a camera coordinate system describes its position with respect to the optical center of the left eye of the stereo camera. For a robotic manipulator, it is essential to know where this point is relative to the robot base coordinate system, which coincides with the world coordinate system.

There are two camera configuration scenarios: eye-in-hand or eye-to-hand. The robot manipulator's end-effector is regarded as a hand, and the camera as an eye.

2.1.3.1 Eye-in-Hand Vision

Eye-in-hand camera configuration refers to when a camera is rigidly mounted on the moving end-effector of a robot, as depicted in Figure 2.1, which is the camera configuration used in the ITER application.

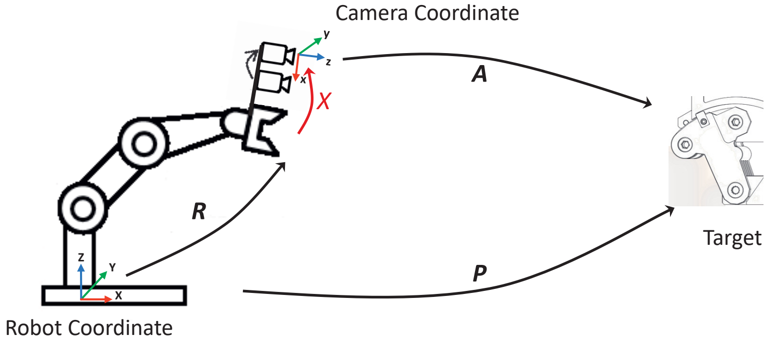


Figure 2.1 A use case of the eye-in-hand vision in the ITER project.

The target object in the world coordinates can be formulated by: $P = RXA$, where P indicates the pose of an object in the robot coordinate system, A is the pose of the object in the camera coordinate system, X is the hand-eye transformation matrix, and R is the current pose of the robot's end-effector in the robot coordinate system. In order to represent the reconstructed 3D scene in robot coordinate system, the pose of the camera with respect to the robot's end-effector X has to be known. And this can be obtained by eye-in-hand calibration [9]: $AX = XB$ where A and B are the robot's end-effector and camera poses between two successive time stamps, respectively.

2.1.3.2 Eye-to-Hand Vision

In eye-to-hand camera configuration, the camera is at a fixed point in the world coordinate system observing both the robot's end-effector and the workspace. Figure2.2 depicts the eye-to-end setup scenario in the Smartbooms2 project.

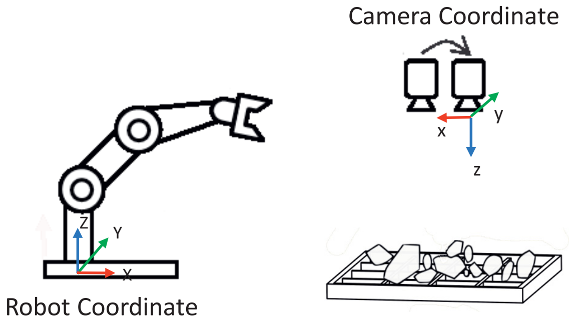


Figure 2.2 A use case of the eye-to-hand vision in the SMARTBOOM2 project.

The position of the 3D reconstructed scene in robot-world coordinate system is determined by the pose of the camera with respect to the robot base. This is also known as the extrinsic camera matrix, represented in the form of transformation matrix H. It consists of rotation matrix R and translation vector t. Rigid transformation is a geometric transformation that preserves the same shape and size in camera and robot world coordinate systems. Given 3D points $C(x_n, y_n, z_n) \in \mathbb{R}^3$ in the camera coordinate system, their corresponding points $W(X_n, Y_n, Z_n) \in \mathbb{R}^3$ in the robot-world coordinate system, and transformation matrix H:

$$C = \begin{pmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \vdots & \vdots & \vdots & \\ x_n & y_n & z_n & 1 \end{pmatrix}, W = \begin{pmatrix} X_1 & Y_1 & Z_1 & 1 \\ X_2 & Y_2 & Z_2 & 1 \\ \vdots & \vdots & \vdots & \\ X_n & Y_n & Z_n & 1 \end{pmatrix}, H = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \tag{2.1}$$

it follows $C^T = HW^T$. Consequently, the 3D points in robot-world coordinate system can be computed by: $W = (H^{-1}C^T)^T$. The approaches for computing transformation matrix H can be categorized as the singular value decomposition (SVD)-based [10, 11] and quaternion based [12, 13, 14]. For the highest level of accuracy and stability, a SVD-based method was adopted [15], which requires more than three pairs of two corresponding points. According to the SVD approach:

$$\Big[U, S, V\Big] = SVD((C - \frac{1}{N}\sum_{n=1}^{N} C^i)(W - \frac{1}{N}\sum_{n=1}^{N} W^i)^T) \tag{2.2}$$

from which R can be obtained: $R = VU^T$, where U and V are orthonormal matrices, and N is the number of pairs of correspondence points from C and W.

Subsequently, the translation vector t can be obtained:

$$t = \frac{1}{N} \sum_{n=1}^{N} W^i - R(\frac{1}{N} \sum_{n=1}^{N} C^i) \tag{2.3}$$

2.2 Perception of Known Objects: Remote Handling at ITER

2.2.1 Model-Based Object Pose Estimation: ICP

Finding an object with known geometric properties in a scene is a typical research question. Capturing objects with the same geometrical appearance at the appropriate level of specificity are commonly relied on predefined 3D computer-aided design (CAD) model of the target.

A classic method of finding such an object utilizes geometric matching of a target object surface with its model surface by performing the iterative closest point (ICP) algorithm [16].

ICP takes two sets of point clouds as input: a model or reference point cloud and the sensed point cloud. Let $M = \{m_i\}$ denotes the model point set, and $m_i = [m_{ix}, m_{iy}, m_{iz}]^T$ in 3D, where i =1, 2, ...N_M and N_M is the number of points in the model shape. Similarly, let $P = \{p_i\}$ denotes the sensed scene shape point set, and $p_i = [p_{ix}, p_{iy}, p_{iz}]^T$ in 3D, where i =1, 2, ...N_p and N_p is the number of points in the sensed scene shape. The output of the algorithm is registration parameters R and t, where R is an operator which applies rotation to its argument (a point) and t is a vector representing translation parameters, $t = [t_x, t_y, t_z]^T$ in 3D.

There are three steps to be done: search for the closest point, search for the best transformation for the correspondence and align the data set. These steps are repeated iteratively.

Firstly, ICP pairs every point of a target set of the scene with the closest point of a model set. For every point p_i in the sensed scene shape P, the algorithm searches for the closest point m_j in the model shape M to the scene point p_i using the Euclidean distance as follows:

$$d(p_i,\ M) = \min_{k=1,...N_M} d(p_i,\ m_k) = \min_{k=1,...N_M} ||p_i - m_k|| \tag{2.4}$$

Once the closest point $m_j \in M$ (model point set) satisfies the equality, then j is the index of the closest point p_i

$$d(p_i, m_j) = d(p_i, M) \implies j = \underset{k\,=\,1,\ldots N_M}{\text{argmin}}\ d(p_i, m_k) \tag{2.5}$$

Secondly, once these corresponding pairs of closest points between two object surfaces are matched, then the transformations R and t for minimizing the error E are computed as follows:

$$E = \sum_{i\,=\,1}^{N_p} ||m_i - (sRp_i + t)||^2 \tag{2.6}$$

where s is the scale factor, and $sRp_i + t$ registers the scene point P to the corresponding model point M.

Lastly, the sensed-scene target object is then rotated and translated by the computed transformation. The iteration process is repeated until the error E falls below a predefined threshold or the number of iterations reaches a chosen constant.

The error decreases monotonically until converging to a local minimum, and if the initial condition is given properly, the algorithm may converge to the global minimum. The issue with the ICP algorithm is in its complexity (i.e., number of points N_p), which grows exponentially with the number of points.

There are many ICP variants [17, 18, 19]. One common improvements has been reducing the influence of outliers on the global error. ICPs can be categorized as SVD [11] based or quaternions [13, 14] based for minimizing the error metric using a closed-form solution.

2.2.2 Edge-Point ICP in ITER's Harsh Environments

A high quality target point cloud is an essential requirement for conventional ICP algorithms. A relatively moderate fraction of outlying points in the input cloud can significantly degrade the performance of an ICP based method, thus preventing its usage in real-world applications [19] [20] . This is particularly important for stereoscopic cameras, as the depth maps and reconstructed 3D point clouds from this kind of passive vision system can deteriorate in ITER's harsh environment due to the following reasons:

- Shiny metallic surfaces: depth estimation becomes unstable due to violation of the Lambertian reflectance model [21].
- High levels of noise: false matches within textureless areas.
- Low-resolution grayscale imagery: another constraint of stereo-matching algorithms.

All these difficulties result in erroneous depth values in depth maps and, consequentially, contaminate sensed point clouds of the scene with outliers [22].

Despite challenging conditions inside the ITER reactor, strong luminance gradients in stereo images are features that can be trusted for their error-free behaviour. In textureless or smooth scenes, strong image gradients usually correspond to object boundaries or significant changes in the surface (e.g., a slope). In contrast to other robust image features, such as scale invariant feature transform (SIFT) [23] or speeded up robust features (SURF) [24], image gradients are much denser and more tolerant to noise in images. Nevertheless, using the object boundaries as matching primitives can also limit the selection of the underlying ICP method.

Subject to the constraints of ITER's environmental conditions, the surface normals generally cannot be estimated at borders and object edges, only point-to-point minimization is possible. More advanced point-to-plane [8] or generalized plane-to-plane [9] minimization approaches cannot be utilized. The recently developed edge-point ICP method [4] is capable of coping with this type of constraint. The method successfully works when the estimated point cloud contains few outliers and a good initialization point is provided.

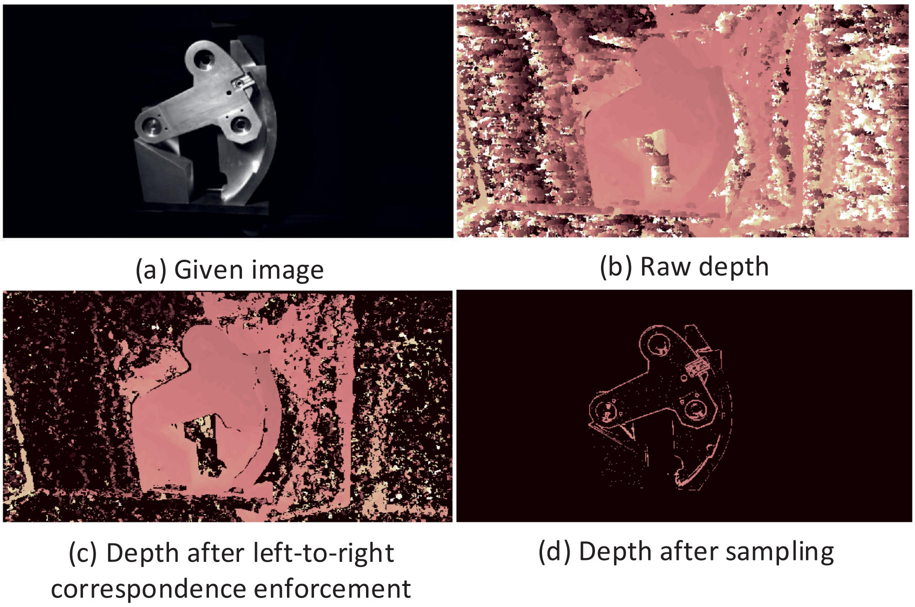
In P-I, for the preparation of high-quality sensed-scene point clouds, the following filtering steps have been considered:

- Left-to-right correspondence enforcement: Two depth maps are used from both left and right cameras in order to compare their values and remove inconsistent ones. This filtering procedure is based on the assumption that the depth of some points in the scene should be the same while looking from both cameras.
- Intensity-based thresholding: This is based on the color value of corresponding depth pixels. As specular reflections in an image are usually overexposed, the color value of a depth pixel that equals or exceeds 254 is considered an outlier. The same is true of too-dark pixels, which are considered irrelevant (too

distant or too slanted).

- Intensity-gradient-based thresholding: The remaining depth values are mostly valid, but can still contain 3D points in the middle of a smooth, flat surface, which are not only useless for pose estimation with the ICP method, but also consume memory and computation power. The useful 3D points from the pre-filtered depth map are pixels having a large magnitude gradient value in the corresponding color (luminance) image. In smooth, textureless scenes, a large luminance gradient usually corresponds to a discontinuity between different objects or surfaces, or alternatively, a sudden illumination change. Similar change-of-surface points can be located on the 3D CAD model using gradients in a surface normal.

An example of the point cloud preparation process is given in Figure 2.3. In order to further improve the robustness to the outliers in the sensed point clouds, the alignment of the sensed point cloud with the reference model point cloud is carried out in two steps, namely, coarse alignment and fine alignment. In coarse alignment, the presence of the approximate planarity structure of the object surface was determined by the RANSAC plane-fitting method, for its robustness in linear model regression and outlier removal. When the CAD model is aligned with its major plane (i.e., model origin and X-Y coordinates belonging to it), the obtained


(a) Given image

(b) Raw depth

(c) Depth after left-to-right correspondence enforcement

(d) Depth after sampling

Figure 2.3 Sampling of the knuckle of the cassette locking system (CLS).

plane parameters can directly be used for the initialization of the rotation matrix. The median centroid of the object is used as an initial value of translation.

The pre-rotated CAD model is rendered on a virtual camera using conventional computer graphics methods. The intrinsic parameters of the virtual camera and its resolution are the same as the left camera of the stereoscopic camera configuration. Then, the depth map of a rendered image follows same outlier removal procedure. As a result, the reconstructed reference model point cloud is matched to the sensed one. In the end, the fine alignment is performed with ICP.

Figure 2.4 illustrates the flowchart of the proposed ICP implementation. Standard edge-point ICP initializes its model point cloud by sampling only once, which is not robust enough for stereoscopic vision. In P-I, a dynamic CAD model resampling mechanism is presented (new blocks within the dashline): the CAD model is rendered and sampled with an estimated initial pose, which is determined by the sensed point cloud of the object.

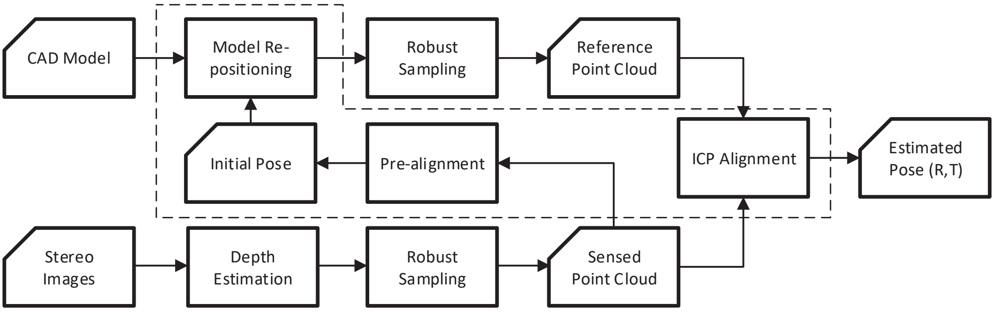


Figure 2.4 Flowchart of proposed iterative closest point (ICP) implementation in [P-I].

As can be seen in Figure 2.5, the blue point cloud has a better initialization pose than the green one, which helps to overcome the issues with local minima.

2.3 Perception of Unknown Objects: Autonomous Rock Breaking

This section presents two approaches for detecting objects without prior geometrical information. The research question comes from the Smartbooms2 project, as shown in Figure 2.6, which requires real-time detection and localization of rocks in a

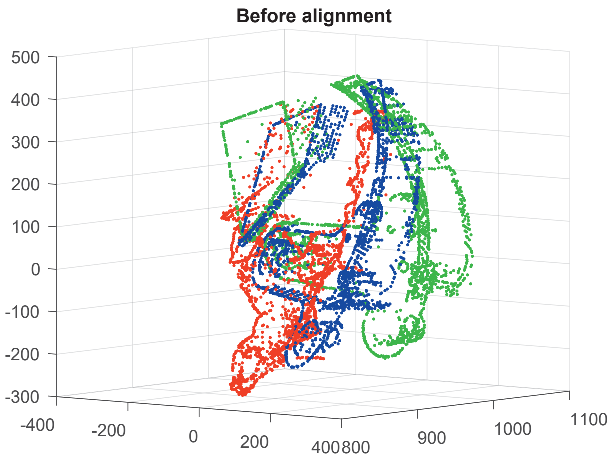


Figure 2.5 Comparison of sampled point cloud: red is sensed, green is from standard ICP, blue is proposed ICP in [P-I].

cluttered scene. Both a TOF camera and a stereo camera were utilized as 3D sensors. The IFM TOF camera provides a sparse 3D point cloud of the scene, while the ZED stereo camera generates a dense point cloud of the scene through 3D reconstruction.

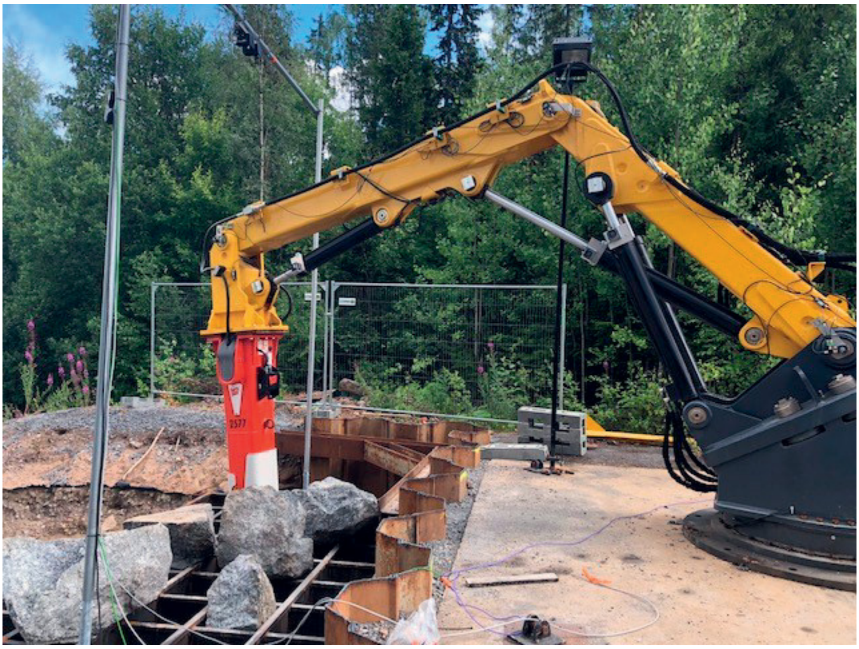


Figure 2.6 The scene of secondary breaking with guidance of a IFM O3M150 time-of-flight (TOF) camera.

2.3.1 Unsupervised Learning: Clustering

The industrial IFM TOF camera is popular in outdoor applications because of features such as IP67 waterproofing, vibration resistance and a wide range of operating temperatures. It provides a point cloud of the scene that describes the geometrical shape of objects and their surrounding environment, as shown in Figure 2.7. Usually each point contains 3D positions (X,Y,Z) and an intensity value, but it does not provide features like color values. In order to detect objects, a clustering method is required to estimate point clusters from the point cloud. Clustering is an unsupervised learning task [25] that deals with structuring unlabeled data. Performing segmentation using clustering allows for the freedom to discover an arbitrary number of objects of any shape in the data.

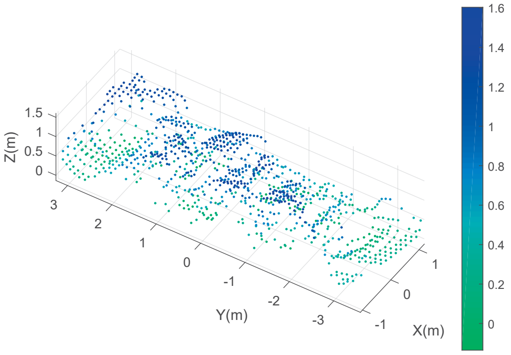


Figure 2.7 Raw sensor data from the TOF camera.

Many existing methods are based on spatial neighborhoods that use the Euclidean distance between two points as the dissimilarity function, i.e. points that are close in the 2D or 3D space form a cluster. By leveraging such properties of point cloud data, the best segmentation results can be achieved. However, many benchmark methods require specification of the number of K clusters being inputted in order to perform clustering (which is a hassle), such as K-means [26] [27], X-means [28], the gaussian mixture model (GMM) [29], spectral clustering [30], and WARD [31]. Other clustering methods such as minimum description length (MDL) based clustering [32], affinity propagation [33] and mean shift [34] either require other hyperparameters or are not suited for data when the clusters have very different sizes. Table 2.1 classifies state-of-the-art clustering algorithms available from the scikit-learn library [35].

Table 2.1 Clustering with or without the optimal number of K clusters for a data set

Inferring K as output	*Specifying K as input*
DBSCAN	K-means
Affinity propagation	Gaussian mixture model
Mean shift	Spectral
	WARD

These start-of-the-art clustering algorithms are evaluated based on the same data set, where the number of rocks varies from 6 to 10. Given a scene of ten rocks as an example, WARD and DBSCAN [36] demonstrate better performance than the rest of the benchmark clustering methods.

The results with DBSCAN imply that density-based clustering can deal with 3D Euclidean-structured data quite well [20]. Despite a few mistakes, DBSCAN is robust enough to infer the number of K clusters. The algorithm treats clusters as dense regions. For each point, DBSCAN searches the number of points within the distance ϵ as a neighborhood, and when the point has at least the pre-defined minimum points *MinPts* in its neighborhood, then the point is considered as core point. That is to say, all of these core points forms dense regions, and points that have fewer than *MinPts* neighborhoods are regarded as noise.

Observing that the 3D point cloud data from the TOF camera are represented as a 3D grid structure, a new clustering method was proposed in P-II. In comparison with DBSCAN, which searches spatial neighbourhoods within a pre-defined sphere, the proposed algorithm searches neighbourhoods within a pre-defined 3D grid (Δx,Δy,Δz), which is in line with the inherent configuration of the TOF camera.

The proposed algorithm is derived from absolute point density values presented specifically to the TOF camera that can be used as global parameters for clustering. The advantages of the proposed algorithm are twofold: Firstly, it predicts the number of clusters, i.e., the number of rocks in the scene. Secondly, no input parameters are required. Similar to DBSCAN's *MinPts*, the proposed method also specifies the minimum number points per cluster i_{min} as initialization for a possible smallest cluster.

The conducted experiments in P-II followed hard clustering criteria, i.e., every point must belong to a proper cluster. In all test scenarios, the proposed method outperforms DBSCAN in the accuracy of the given TOF point cloud data.

2.3.2 Deep Learning: Convolutional Neural Network Based Object Detection

The TOF camera, which is mounted about 5 meters above the grizzly, represents a scene with 1024 (64×16) points. This is a very low-resolution point cloud, which also does not contain color features but only the intensity, and thus it is unable to perform well in cluttered scenes where small rocks overlap each other. Therefore, a higher resolution ZED stereo (4416×1242) camera was utilized. Thus, more research methods became available.

Object detection generally refers to the classification and detection of objects in 2D image or 3D point cloud. The common supervised learning approach is to use a trained convolutional neural network (CNN) to classify and detect a single object, and then to slide it across the image, which is not only slow, but also computationally expensive. In recent years, deep learning architecture has become ubiquitous in object detection [37, 38, 39, 40, 41, 42, 43, 44, 45], all of which is based on CNNs.

The progress in 3D object detection research has been significant, however, current studies have been mainly focusing on objects with known geometries [46, 47, 48] on benchmark datasets, or light detection and ranging (LIDAR) based applications [49, 50, 51, 52, 53, 54, 55, 56, 57] without taking into account overlapping objects. Compared to high resolution RGB images, 3D point clouds are irregular, thus typical CNNs are not well suited to directly process them [58]. LIDAR point clouds are relatively sparse and unstructured, and the plausible 3D shapes presented by point clouds are often unable to represent all the detailed features of objects; they are inadequate to interpret the details of complex scenes, such as when a pile of irregularly shaped small rocks are overlapping each other. Moreover, annotations of 3D point clouds are time consuming. Overall, in comparison with the 3D object detection approaches of 3D point clouds, 2D object detection methods based on images are more sophisticated for industrial deployments.

State-of-the-art, real-time 2D object detection methods can be categorized into two main groups: region based or single shot based. The former includes region-based convolutional neural networks (R-CNN) [59], fast R-CNN [60], and faster R-CNN [61]. The latter includes single shot multibox detector (SSD) [62] and you only look once (YOLO). YOLO detectors [63, 64, 65] have become a widely used alternative to R-CNN variants by achieving superior detection efficacy.

YOLOv3 is the current YOLO model for object detection. It takes RGB images as an input and then predicts an output value as a classification; it relies on annotated real-image data as a ground truth, and then computes the error between the ground truth and model estimated output as a loss function. The average of the entire training set is used to compute the cost function, minimizing it through back propagation steps to compute gradient descent in order to achieve the global optimum in parameter weights. The network is based on darknet-53 as a feature extractor, and the major hyperparameters for tuning are learning rate and batch size.

Because Smartbooms2 requires only the detection of rocks, i.e., only one class of objects, the output vector of YOLOv3 is quite simple, as shown in Figure 2.8, where the output vector y contains five elements: the probability of predicted value p between 0 and 1, and the position (b_x, b_y) and size (b_w, b_h) of the bounding box in the image.

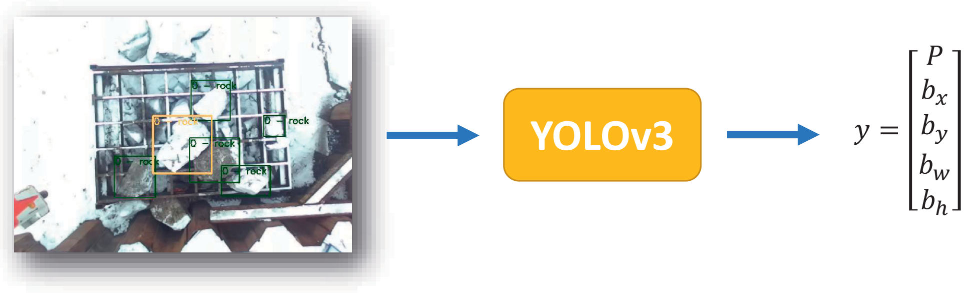

rock
YOLOv3
$y = \begin{bmatrix} P \\ b_x \\ b_y \\ b_w \\ b_h \end{bmatrix}$

Figure 2.8 YOLOv3 output format for rock detection.

Given a test image and the trained CNN, the YOLOv3 workflow is as follows. Firstly, the test image is divided into cell grids. Based on the size of the test image, the size of the grid cells in pixels varies. Secondly, each grid cell is used for predicting a set of bounding boxes. For each bounding box, the network also predicts the confidence that the bounding box encloses a particular object as well as the probability of the object belonging in a particular class. Lastly, a non-maximum suppression is used to eliminate bounding boxes with a low confidence level, as well as redundant bounding boxes enclosing the same object. In case of overlapping, YOLOv3 provides anchor boxes for each grid to allow for detection of multiple objects.

YOLOv3 is one among the fastest and the most accurate object detection algorithms for 2D images. In view of its performance, P-III adopted its existing architecture and then extended it to 3D object detection. As illustrated in Figure 2.9, the detected rocks are represented as 2D regions enclosed by bounding boxes, much like 3D reconstruction from stereo images, and the detected regions on the left image (reference image) of the stereo camera can be reconstructed as 3D regions with the corresponding depth maps. Thus, the detected rocks in point clouds enclosed with 3D bounding boxes are generated.

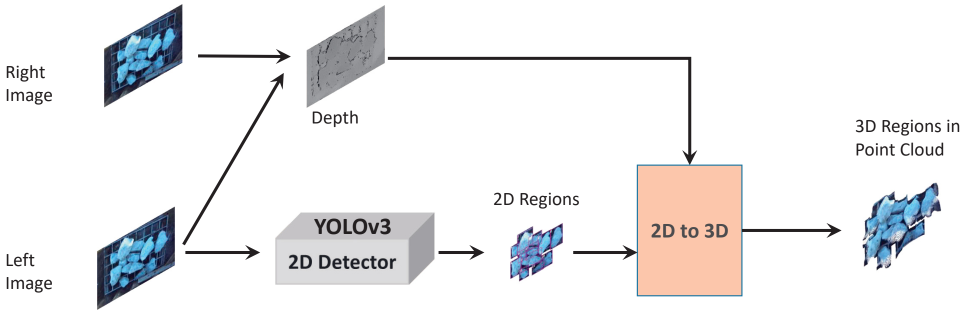
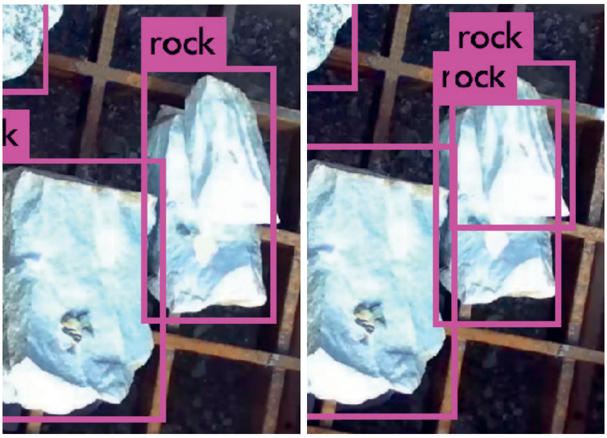


Figure 2.9 The 3D object detection mechanism in [P-III].

The performance of YOLOv3 essentially relies on data; as the works in [66, 67, 68] show that with a large and varied data set, deep learning models work very well, and the dataset continues growing, the deep learning neural networks perform better, with higher accuracy. The Smartbooms2 rock data set initially consisted of 4733 images[1] collected from the field test site (Figure. 2.6), where the amount of rocks in the scene varied from one to 15. Nevertheless, the image datasets contain only high-contrast images taken under normal sunny lighting conditions, thus images taken in other outdoor conditions are missing, such as during days with rain, snow or fog. Though real-world data is the best option for any neural network training, the collection of data in different outdoor conditions can be difficult and time consuming. Data augmentation, the process of generating realistic synthetic data, is one way to bridge the experimental scenario reality gap [69, 70, 71, 72].

For a stereo camera, dynamic outdoor illumination conditions can be challeng-

[1] `https://github.com/epoc88/SecondaryBreakingDataset`. It has now been expanded to 23850 images.



(a) Original model [P-III] (b) Improved model [P-IV]

Figure 2.10 Compared detection results following data augmentation, the scenario depicts a smaller rock on top of a bigger rock under overexposed lighting conditions.

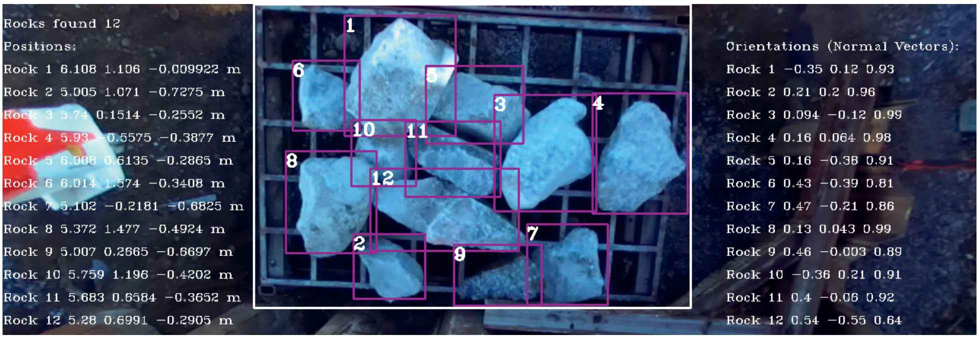
ing, as very bright lighting conditions can make object edges indistinguishable. Thus, it is beneficial to feed more images into the training model when the deep neural network fails, as deep neural networks will over time accumulate and pick up patterns. P-IV presents an experiment for the evaluation of model performance, and the result is depicted in Figure 2.10.

Further performance measures are conducted using the mean average precision (mAP) metric. An average precision of 97.61% was reached at a intersection over union (IOU) threshold of 0.5, with an average detection time of 85 ms per frame.

Rock detection is a classification problem, while the localization of rocks is a regression problem. Each detected rock's position is the center of its enclosed bounding box, which is a relative position (between 0 and 1) to a specific grid cell in the image, thus, it cannot be used as a rock breaking position. Based on two years of experiments, P-IV proposed an effective breaking scheme, a new search mechanism for the breaking positions, as well as novel methods for searching for orientation angles for breaking.

The final autonomous secondary breaking experiments were conducted with between 6 and 12 rocks in the scene. An example result is shown in Figure 2.11, with 12 rocks in the scene.

To validate the above results in real-time, a 3D viewer, shown in Figure. 2.12, was implemented using Point Cloud Library (PCL) in C++. The estimated positions



```
Rocks found 12
Positions:
Rock 1 6.108 1.106 −0.009922 m
Rock 2 5.005 1.071 −0.7275 m
Rock 3 5.74 0.1514 −0.2552 m
Rock 4 5.93 −0.5575 −0.3877 m
Rock 5 6.008 0.6135 −0.2865 m
Rock 6 6.014 1.574 −0.3408 m
Rock 7 5.102 −0.2181 −0.6825 m
Rock 8 5.372 1.477 −0.4924 m
Rock 9 5.007 0.2665 −0.6697 m
Rock 10 5.759 1.196 −0.4202 m
Rock 11 5.683 0.6584 −0.3652 m
Rock 12 5.28 0.6991 −0.2905 m
```

```
Orientations (Normal Vectors):
Rock 1 −0.35 0.12 0.93
Rock 2 0.21 0.2 0.96
Rock 3 0.094 −0.12 0.99
Rock 4 0.16 0.064 0.98
Rock 5 0.16 −0.38 0.91
Rock 6 0.43 −0.39 0.81
Rock 7 0.47 −0.21 0.86
Rock 8 0.13 0.043 0.99
Rock 9 0.46 −0.003 0.89
Rock 10 −0.36 0.21 0.91
Rock 11 0.4 −0.06 0.92
Rock 12 0.54 −0.55 0.64
```

Figure 2.11 An example of the secondary breaking scenario with 12 rocks.

for guiding the manipulator's hammer are marked as red spots on the surface of each rock. The 3D viewer provides a 3D representation of detected objected, and visualize the breaking position in live.

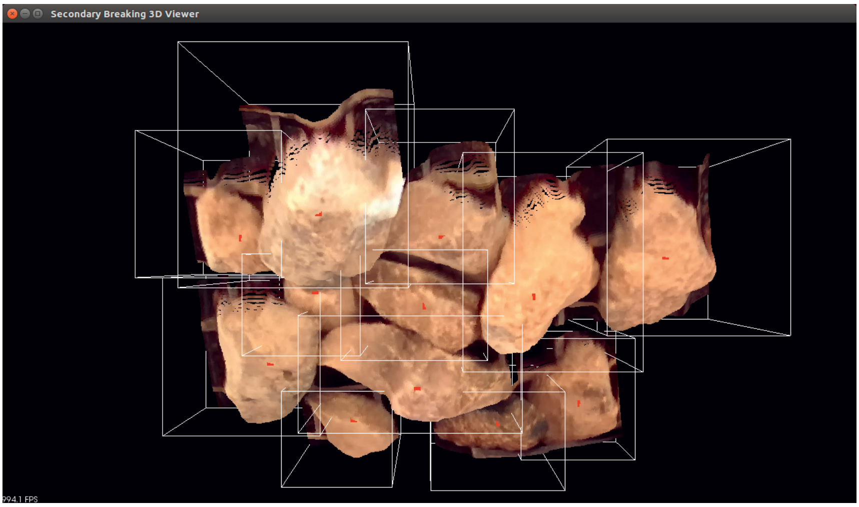


Figure 2.12 A real-time view of detected rocks with breaking positions indicated in red dots.

3 SUMMARY OF PUBLICATIONS

This chapter summarizes each thesis publication to outline the previously mentioned research problems. The hardware and software architectures addressed in P-I and P-V are depicted in [1].

3.1 Robust Pose Estimation with a Stereoscopic Camera in Harsh Environments

Although 6-DOF pose estimation solutions have been well studied, pose estimation inside a radioactive reactor chamber remains challenging. In the ITER fusion reactor, images of the target object in a scene were acquired with a low-resolution radiation-tolerant grayscale camera in which a high level of noise is present. Moreover, the target object appeared to have non-Lambertian reflectance in the case of shiny metallic surfaces, as well as a deformed shape due to erosion. Such extreme conditions create constraints not only for hardware (Figure 2, [1]), but also for generic pose estimation methods. For a rigid object whose prior geometric information is known, the conventional approach is to apply ICP for registration. However, no ICP methods can perform adequately within ITER's harsh environment. P-I proposes a novel edge-point ICP method to robustly align the sensed object with the reference object. In addition, the paper proposes an advanced plane sweeping approach to improve the precision of the stereoscopic vision system.

Experiments were conducted in comparison with the classic edge-point ICP method, as well as the proposed approach. Given two CLS knuckle types as the target objects, the relative accuracy of the vision system was assessed with a repeatability test where both methods were compared in terms of the number of outliers, position stability, and angular stability using images taken from the replicated scene. The pose estimation results verified the efficiency of the proposed edge-point ICP method.

3.2 Clustering Analysis for Secondary Breaking Using a Low-cost Time-of-flight Camera

This paper presents a case study of secondary rock breaking using a low-cost industrial TOF camera. The aim was to find an unsupervised learning approach to make the best use of sparse point cloud data from the scene to achieve rock detection and localization in real time. The paper first highlights an overview of state-of-the-art clustering methods for analyzing TOF camera point cloud data. In light of the issues arising from these existing methods, the paper proposes a novel clustering method based on the spatial characteristics of TOF cameras. The conducted experiments indicate the reliability of the proposed method, which outperformed two of the best state-of-the-art methods for this task, DBSCAN and WARD. The proposed clustering method can accurately detect and localize rocks in a point cloud provided that no rocks overlap. The study also revealed the benefits of utilizing a TOF camera for outdoor applications while highlighting its limitations due to its weak spatial resolution.

3.3 Efficient 3D Visual Perception for Robotic Rock Breaking

This paper presents a further study of secondary rock breaking by means of 3D object detection using deep learning methodology. The aim of the paper was to resolve existing issues in rock detection, such as the detection of small rocks overlapping one another in a pile. The paper introduces a novel 3D visual perception pipeline for rock breaking. It offers solutions for rock detection in an acquired image, reconstructing detected rocks from an image into a 3D point cloud, estimating the position of a detected rock, and leveraging k-dimensional tree and RANSAC algorithms for orientation estimation. The rock detection model was implemented with the state-of-the-art YOLOv3 infrastructure, referencing the darknet-53 convolution neural network as the backbone. The deep learning model was built based on the training of 4733 real-world image data collected from the field. As the result, the model achieved a 97% average precision rate with a detection speed of 10 Hz. The VPS was capable of detecting and localizing irregularly shaped rocks in a clutter scene.

3.4 Autonomous Robotic Rock Breaking Using a Real-time 3D Visual Perception System

In this paper, an end-to-end solution for potential commercial applications of autonomous rock breaking is provided. The work undertaken involved deploying an industry-ready 3D perception system, designing a manipulator control system, onsite camera calibration and accuracy evaluation, manipulator calibration, communication, and auxiliary tasks required for performing autonomous robotic tasks. As the most prominent component of the paper, the 3D virtual perception pipeline was further developed to resolve challenges in actual outdoor experiments. First, the accuracy and robustness of the object detection model were optimized for a dynamic outdoor environment. This was achieved by feeding more images into training data from situations where the current model failed to detect rocks. The new image data set consists of both real-world images and synthetic images that were created via data augmentation. Second, intrinsic and extrinsic camera calibration was conducted indoors and in the field, respectively. The precision of the vision system was assessed by the marker position measurements. Third, new rock breaking mechanism was proposed based on empirical evidence; the new breaking positions were determined by the shape and size of a detected rock rather than its centroid, and the breaking orientations were determined by the rock surface of a circular area within a diameter of 135 mm centered on the breaking position. Fourth, a 3D viewer was implemented for real-time visualization of breaking positions in robot coordinates, which could validate the correctness and effectiveness of breaking positions online. Fifth, the implementation of data analysis, processing, validation, rendering, and communication modules were reported. Rock detection results were analyzed such that only oversized rocks on the grizzly were accepted, while the rest were ignored. Positions and orientations for rock breaking were rendered according to rock height and size. The rendered data facilitated the decision-making process for final breaking operations. Finally, data communication with the manipulator was implemented via UDP communication, where each pair of position and orientation values sent were indexed. Apart from these improvements, the paper also addresses the Rambooms manipulator's DOF constraints and provides a solution for utilizing the 3D orientation information. The efficacy of the 3D VPS is demonstrated in the final autonomous rock breaking experiments.

3.5 A Stereoscopic Eye-in-hand Vision System for Remote Handling in ITER

P-I proposed a novel edge-point ICP method for 6-DOF object pose estimation in cluttered scenes. The relative accuracy of the vision system was tested with two instances of knuckles in camera coordinates. In P-V, a stereoscopic eye-in-hand robotic perception system was implemented and integrated with the RHCS. The aim of the paper was to validate the accuracy and reliability of the stereoscopic eye-in-hand vision system for fulfilling the generic ITER vision system requirements, as well as transferability to other RHCSs. Alongside [1], this work detailed vision system software and hardware architecture design, implementation of different vision system operation modes, RHCS software and hardware architecture design, calibration of system components (manipulator, intrinsic and extrinsic camera parameters, and the cassette operation tools), and communication between the heterogeneous subsystems. The vision system's precision and robustness were verified in a demonstration utilizing a vision-guided pin tool and jack tool operation in a replicated ITER environment, where a clearance of 3 mm was required for tool operations. The experiments illustrate the general applicability of the vision system for other RHCSs and the feasibility of the novel edge-point ICP method for object pose estimation under harsh ITER conditions using a low-resolution radiation-tolerant camera.

4 DISCUSSION

This chapter discusses the relevant research problems, explains research outcomes, and addresses research limitations, divided by topic.

4.1 Visual Perception in Remote Handling

Can the robustness of target-object pose estimation be improved in challenging real-world, heavy-duty robotic scenarios?

P-I suggests using the edge properties of objects for improving the reliability of detection. In the process of 3D scene reconstruction, P-I proposes the plane sweeping method for accurate and robust depth estimation instead of the conventional rectification-based method. To ensure robustness in 3D registration, P-I presents two phase alignments: coarse and fine alignment. The former leverages the RANSAC algorithm to identify a strong initialization point for fine alignment, while the latter applies a novel edge-point ICP method to accurately align the sensed object point cloud with the reference object point cloud. In doing so, an object's pose in a robot world coordinate system can be obtained by transforming its pose in camera coordinates using pre-calibrated eye-in-hand extrinsic camera parameters, as well as the pose of the manipulator TCP. The study also focuses on outlier removal methods. Our findings suggest that distortions in the sensed point cloud can be effectively removed by applying a sequence of filtering methods. These methods include left-to-right correspondence checks, intensity-based thresholding, and intensity gradient-based thresholding. Nevertheless, potential threats to robustness may still exist due to unpredictable ITER environments. Unexpected outliers removal is a major area for further study, since registration can be disrupted in the presence of outliers.

4.2 Visual Perception in Autonomous Secondary Breaking

Can fast detection and localization for objects be obtained without prior known geometry in a scenario with piles of overlapping objects?

The development of object detection methods for secondary breaking was carried out in two stages. In the first stage, a TOF camera was employed to acquire the scene in a 3D point cloud. As a common type of unsupervised learning, clustering techniques can be used to cluster distinct types of objects, as underscored by the solution presented in P-II. The study in P-II also addresses the limitations of the IFM O3M150 TOF camera with a resolution of 64x16. Such a low spatial resolution restricts object detection methods. As a result, only large and detached objects can be detected, small rocks can hardly be recognized, occluded objects cannot be distinguished, and a pile of rocks can only be regarded as one rock. The study also demonstrates how rock detection approaches are heavily influenced by the type of 3D sensor used, its spatial resolution, and the depth accuracy of its working range, as all of these factors determine the available features and how they can be used. In view of this, in the second stage a ZED stereo camera with a sensor resolution of 4416x1242 was adopted, which allows the extraction of rich features from a scene. In return, more research approaches become available.

For a cluttered scene in which rocks overlap, P-III presents a rock detection solution that incorporates YOLOv3, a prominent object detection algorithm. The convolutional layers of the deep neural network are capable of capturing important object features in an image, performing well with small object detection. Given RGB images as an input, the detected objects in the outputs of the deep neural network are enclosed by bounding boxes, which is characterized by parameters of the center, the height and width of the bounding box. As no convolutional deep neural networks determine the shape of the object, P-III presents a solution via 3D reconstruction of the detected object from stereoscopic imagery. The reconstructed 3D point cloud of the object in the camera's frame not only represents the geometry of the objects, but also enables further estimation of the position and orientation of the object.

Generally speaking, the performance of the deep learning model depends of quality and quantity of the data it was given. To improve performance under diverse outdoor weather conditions (including rainfall, snowfall, and fog), new training data were created via data augmentation in P-IV. Therefore, it is beneficial to feed more real and synthetic images into training for improving the model.

Like other supervised learning methods, YOLOv3 requires labeling a large amount of image data to train the model, which is time consuming. Since occlusions can result from obstructions by the manipulator arm. In future, this can be remedied by using two or more mounted cameras for multiple view geometry, and thus requires the registration of multiple camera point clouds to reconstruct a scene.

4.3 Demand for Application Specific Cameras and Setup

Cameras are ideal tools for robotic perception. Depending on the application requirements, a camera can either be moving (for instance, mounted to the robot arm in the ITER scenario) or fixed in a workspace (as in the Smartbooms2 scenario).

In ITER, a close-range camera is required to provide millimeter-level accuracy. An eye-in-hand setup is beneficial in obtaining precise measurements between the end effector and the target, and the pose of a camera can also be adjusted to ensure the best field of view without occlusion. Apart from this, a radiation-tolerant camera with a higher resolution sensor would further improve the precision and robustness of the overall system.

In Smartbooms2, secondary rock breaking takes place on a grizzly in an outdoor environment. In the experiment, the camera was fixed five meters above the grizzly to provide a view of the entire workspace. This eye-to-hand camera configuration led to an occlusion scenario caused by the presence of the moving manipulator arm. However, a multiple-camera setup can be applied to resolve this issue. For another use case scenario depicted in Figure 1.2, eye-in-hand camera configuration is necessary. However, such a setup requires a custom design for tackling vibration and lens protection challenges. Moreover, an outdoor industry-ready stereo camera must fulfill various outdoor requirements, such as being waterproof, temperature tolerant, and vibration resistant.

4.4 System Precision

Precision is an important factor in characterizing the performance of any robotic system. For a vision-guided robotic system, precision relies on image processing algorithms, the inherent accuracy of the camera and robotic manipulator, the calibration methods for the camera and robotic manipulator, and measurement methods. However, in practical experiments, errors between estimated and ground truth values are unavoidable, and system-level errors are accumulated from all sub-level components.

In ITER, remote handling leverages an open-loop control scheme. As a consequence, the overall system accuracy depends on the accuracy of all components in the chain. For example, the target pose of the robot frame is determined by the intrinsic and extrinsic calibration of the eye-in-hand camera, the pose of the manipulator's TCP, and the target pose with respect to the camera. While system calibration errors can be minimized with pose-based visual servoing, this process requires closed-loop control. As such, for safety reasons, this approach is not used in RH. To improve overall system precision, it is essential to refine the selection of all components in the system, as well as improve image processing algorithms, calibration, and measurement methods.

To validate the robustness and accuracy of the eye-in-hand vision system, P-I presented a repeatability test for operating distances of 500–1500 mm in which both positional and orientation errors were compared with those of the classic ICP method. The estimated pose of the target object, however, could not be compared with a ground truth. Instead, the accuracy of the vision system had to be assessed using tool operation experiments where robot and tool calibration errors were present. In P-V, the conducted experiments with tool operations were carried out in a replicated ITER environment in which only 3 mm of clearance was available. The success of these experiments validated the overall precision of the vision system and the effectiveness of the pose estimation method.

In comparison with ITER's millimeter-level accuracy requirement, the Smartbooms2 requirement is 150 mm. Nevertheless, after several rounds of calibration, the actual system precision was found to be greater than this. For the vision system, the maximum Cartesian error was 67.19 mm when operating at a five-meter distance. Kinematic calibration yielded accuracy within 8.37 mm for the system's kinematics, and the control system yielded a maximum Cartesian error of 60 mm in free space trajectory tracking.

5 CONCLUSION

This chapter concludes the studies on visual perception in challenging real-world scenarios. The relation between research methods is illustrated in Figure 1.5. In this thesis, major progress was made in addressing RP.I and RP.II.

In the RP.I scenario, the target objects were ITER reactor components whose 3D CAD models were available in advance. However, these components were subject to small drifts in their 6-DOF poses due to extreme heat and high magnetic fields inside the ITER reactor. Thus, a VR representation of the RH environment may not reflect the actual scene accurately. This thesis outlines the precise and reliable navigation of RH maintenance operations toward the development of a robust and accurate 3D VPS.

Due to the constraints of harsh ITER conditions, research methods in such contexts are limited. The related study had to navigate various challenges to improve the robustness of the target object pose estimation technique. To achieve this goal, several efforts in designing the stereoscopic VPS were presented, such as eye-in-hand configuration, a robust depth estimation method for the accurate 3D reconstruction of target objects using low-resolution grayscale stereo images, a variety of approaches for outlier removal, and a novel edge-point ICP algorithm for robust pose estimation.

Finally, the conduct of demanding RH operations demonstrates that the developed system can cope with the limitations set by a harsh ITER environment, such as image acquisition with low-resolution and grayscale radiation tolerant camera, high level of image noises due to the radiation, non-Lambertian reflectance of reactor elements on shiny metallic surfaces, and deficient illumination of the scene due to constraints on available light sources. As a conclusion, the developed VPS meets generic ITER requirements and can significantly improve the RH operator experiences. It not only assists the human-operator to locate remote objects quickly and accurately, but also ensures RH tasks to be performed efficiently and safely, as well as reducing operator stress.

In the RP.II scenario, object detection methods that rely on predefined CAD models are infeasible, as rocks do not possess a regular shape or specific surface geometry. Moreover, developing an autonomous rock breaking system requires advanced robotic visual perception capable of instantly detecting and localizing overlapped rocks in a cluttered scene under dynamic outdoor conditions. The thesis presents two relevant case studies.

In the first case, a popular low-cost TOF camera was employed to generate a sparse point cloud of the scene. The rocks in the scene were represented as 3D Euclidean grid-structured data that allowed for global parametrization. The study proposes a novel unsupervised learning algorithm, which outperformed the state-of-the-art DBSCAN and WARD methods. It also addresses the research limitations caused by the TOF camera's spatial resolution constraints.

In the second case, a ZED stereo camera was adopted for its high resolution and compact size. The study leveraged recent advancements in CNN-based deep learning models, which can aggregate the features of a full RGB image regardless of complexity so that object detection can be performed on a granular and regional level of the image.

The thesis presents significant work in deploying real-time 3D VPS for autonomous robotic application, which involved data preparation, enhanced training for a deep learning model, proposing and implementing a novel 3D rock detection pipeline, and designing and implementing an innovative rock breaking mechanism.

Overall, the proposed robotic VPS meets the requirements for the mining industry with its average rock detection rate (97.61%), real-time performance (11.76 Hz), and capability of autonomous rock breaking without any human intervention. The results offer a clear indication of the technological readiness of such system.

Visual perception starts at 3D sensors, but real processing is done by a computer. Classical computer vision incorporates geometric methods that employ the mathematics necessary for understanding the geometry of a 3D scene. Meanwhile, recent developments in machine learning and deep learning approaches have greatly advanced the understanding of 3D scenes.

Overall, this thesis commits to combining computer vision, machine learning or deep learning techniques in order to maximize the value of visual perception for robotics and to contribute to cutting-edge technological advancements.

BIBLIOGRAPHY

[1] L. Niu, O. Suominen, M. M. Aref, J. Mattila, E. Ruiz and S. Esque. Eye-in-hand manipulation for remote handling: Experimental setup. *IOP Conference Series: Materials Science and Engineering*. Vol. 320. 2017, 012007.

[2] Sandvik Mining and Construction. *Hydraulic hammer rammer 2577, operator's manual*. Available at `https://www.rammer.com/en/products/hydraulic-hammers/excellence-line/medium-range/2577/` Seen: 10.10.2019. 2016. URL: `https://www.rammer.com/en/products/hydraulic-hammers/excellence-line/medium-range/2577/`.

[3] Sandvik Mining and Construction. *Rammer hammer catalog*. Available at `https://www.marakon.fi/images/pdf/Rammer_HAMMERS.pdf` Seen: 12.10.2019. 2012. URL: `https://www.marakon.fi/images/pdf/Rammer_HAMMERS.pdf`.

[4] P. Alho and J. Mattila. Real-time service-oriented architectures: A data-centric implementation for distributed and heterogeneous robotic system. *International Embedded Systems Symposium*. Springer. 2013, 262–271.

[5] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[6] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision* 47.1-3 (2002), 7–42.

[7] S. Smirnov, M. Georgiev and A. Gotchev. Comparison of cost aggregation techniques for free-viewpoint image interpolation based on plane sweeping. *Ninth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. 2015.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24.6 (1981), 381–395.

[9] R. Y. Tsai and R. K. Lenz. Real time versatile robotics hand/eye calibration using 3D machine vision. *Proceedings. 1988 IEEE International Conference on Robotics and Automation*. IEEE. 1988, 554–561.

[10] N. Ho. Finding optimal rotation and translation between corresponding 3D points. *URL http://nghiaho. com* (2013).

[11] K. S. Arun, T. S. Huang and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on pattern analysis and machine intelligence* 5 (1987), 698–700.

[12] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *Josa a* 4.4 (1987), 629–642.

[13] B. K. Horn, H. M. Hilden and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A* 5.7 (1988), 1127–1135.

[14] M. W. Walker, L. Shao and R. A. Volz. Estimating 3-D location parameters using dual number quaternions. *CVGIP: image understanding* 54.3 (1991), 358–367.

[15] A. Lorusso, D. W. Eggert and R. B. Fisher. *A comparison of four algorithms for estimating 3-D rigid transformations*. University of Edinburgh, Department of Artificial Intelligence, 1995.

[16] P. J. Besl and N. D. McKay. A Method for registration of 3-D shapes. *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. International Society for Optics and Photonics. 1992, 586–606.

[17] D. Chetverikov, D. Stepanov and P. Krsek. Robust Euclidean alignment of 3D point sets: the trimmed iterative closest point algorithm. *Image and vision computing* 23.3 (2005), 299–309.

[18] M. Tomono. Robust 3D SLAM with a stereo camera based on an edge-point ICP algorithm. *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, 4306–4311.

[19] R. Marani, V. Reno, M. Nitti, T. D'Orazio and E. Stella. A modified iterative closest point algorithm for 3D point cloud registration. *Computer-Aided Civil and Infrastructure Engineering* 31.7 (2016), 515–534.

[20] S. Du, Y. Xu, T. Wan, H. Hu, S. Zhang, G. Xu and X. Zhang. Robust iterative closest point algorithm based on global reference point for rotation invariant registration. *PloS One* 12.11 (2017), e0188039.

[21] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.2 (2003), 218–233.

[22] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha and M. Beetz. Towards 3D point cloud based object maps for household environments. *Robotics and Autonomous Systems* 56.11 (2008), 927–941.

[23] D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, 1150–1157.

[24] H. Bay, T. Tuytelaars and L. Van Gool. Surf: Speeded up robust features. *European conference on computer vision*. Springer. 2006, 404–417.

[25] G. E. Hinton, T. J. Sejnowski, T. A. Poggio et al. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[26] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28.2 (1982), 129–137.

[27] R. L. Thorndike. Who belongs in the family. *Psychometrika*. Citeseer. 1953.

[28] D. Pelleg, A. W. Moore et al. X-means: Extending k-means with efficient estimation of the number of clusters. *ICML*. Vol. 1. 2000, 727–734.

[29] D. H. H. Santosh, P. Venkatesh, P. Poornesh, L. N. Rao and N. A. Kumar. Tracking multiple moving objects using gaussian mixture model. *International Journal of Soft Computing and Engineering (IJSCE)* 3.2 (2013), 114–119.

[30] T. Ma, Z. Wu, L. Feng, P. Luo and X. Long. Point cloud segmentation through spectral clustering. *The 2nd International Conference on Information Science and Engineering*. IEEE. 2010, 1–4.

[31] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58.301 (1963), 236–244.

[32] H. Bischof, A. Leonardis and A. Selb. MDL principle for robust vector quantisation. *Pattern Analysis & Applications* 2.1 (1999), 59–72.

[33] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science* 315.5814 (2007), 972–976.

[34] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), 603–619.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[36] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*. Vol. 96. 34. 1996, 226–231.

[37] Z. Cai, Q. Fan, R. S. Feris and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. *European Conference on Computer Vision*. Springer. 2016, 354–370.

[38] J. Lahoud and B. Ghanem. 2d-driven 3d object detection in rgb-d images. *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 4622–4630.

[39] C. R. Qi, W. Liu, C. Wu, H. Su and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 918–927.

[40] B.-s. Kim, S. Xu and S. Savarese. Accurate localization of 3D objects from RGB-D data using segmentation hypotheses. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, 3182–3189.

[41] D. Lin, S. Fidler and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. *Proceedings of the IEEE International Conference on Computer Vision*. 2013, 1417–1424.

[42] W. Liu, R. Ji and S. Li. Towards 3D object detection with bimodal deep Boltzmann machines over RGBD imagery. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, 3013–3021.

[43] Q. Luo, H. Ma, L. Tang, Y. Wang and R. Xiong. 3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection. *Neurocomputing* 378 (2020), 364–374.

[44] K. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, 770–778.

[45] S. Gupta, P. Arbeláez, R. Girshick and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, 4731–4740.

[46] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. *Asian Conference on Computer Vision*. Springer. 2012, 548–562.

[47] B. Li, W. Ouyang, L. Sheng, X. Zeng and X. Wang. GS3D: An efficient 3D object detection framework for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[48] L. Liu, J. Lu, C. Xu, Q. Tian and J. Zhou. Deep fitting degree scoring network for monocular 3D object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[49] Q. He, Z. Wang, H. Zeng, Y. Zeng, S. Liu and B. Zeng. SVGA-Net: Sparse voxel-graph attention network for 3D object detection from point clouds. *ArXiv Preprint ArXiv:2006.04043* (2020).

[50] D. Z. Wang and I. Posner. Voting for voting in online point cloud object detection. *Robotics: Science and Systems*. Vol. 1. 3. 2015, 10–15607.

[51] X. Zhao, Z. Liu, R. Hu and K. Huang. 3D object detection using scale invariant and feature reweighting networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, 9267–9274.

[52] E. Al Hakim. *3d yolo: End-to-end 3d object detection using point clouds*. 2018.

[53] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 4490–4499.

[54] X. Chen, H. Ma, J. Wan, B. Li and T. Xia. Multi-view 3d object detection network for autonomous driving. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 1907–1915.

[55] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong and I. Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, 1355–1361.

[56] B. Yang, W. Luo and R. Urtasun. PIXOR: Real-Time 3D Object Detection From Point Clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[57] M. Liang, B. Yang, Y. Chen, R. Hu and R. Urtasun. Multi-task multi-sensor fusion for 3D object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[58] C. R. Qi, O. Litany, K. He and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. *Proceedings of the IEEE International Conference on Computer Vision*. 2019, 9277–9286.

[59] R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, 580–587.

[60] R. Girshick. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*. 2015, 1440–1448.

[61] S. Ren, K. He, R. Girshick and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*. 2015, 91–99.

[62] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg. SSD: Single shot multibox detector. *European Conference on Computer Vision*. Springer. 2016, 21–37.

[63] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*. 2016, 779–788.

[64] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, 7263–7271.

[65] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *ArXiv Preprint ArXiv:1804.02767* (2018).

[66] T. Tran, T. Pham, G. Carneiro, L. Palmer and I. Reid. A bayesian data augmentation approach for learning deep models. *Advances in Neural Information Processing Systems*. 2017, 2797–2806.

[67] N. Rusk. Deep learning. *Nature Methods* 13.1 (2015), 35.

[68] G. Marcus. Deep learning: A critical appraisal. *ArXiv Preprint ArXiv:1801.00631* (2018).

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

[70] C. Sakaridis, D. Dai and L. Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126.9 (2018), 973–992.

[71] X. Peng, B. Sun, K. Ali and K. Saenko. Learning deep object detectors from 3d models. *Proceedings of the IEEE International Conference on Computer Vision*. 2015, 1278–1286.

[72] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, 699–715.

PUBLICATIONS

PUBLICATION

I

Robust pose estimation with a stereoscopic camera in harsh environments

L. Niu, S. Smirnov, J. Mattila, A. Gotchev and E. Ruiz

IS&T International Symposium on Electronic Imaging 2018.9 (2018), 126–1
DOI: `10.2352/ISSN.2470-1173.2018.09.IRIACV-126`

https://doi.org/10.2352/ISSN.2470-1173.2018.09.IRIACV-126

Robust Pose Estimation with a Stereoscopic Camera in Harsh Environments

Longchuan Niu*, **Sergey Smirnov****, **Jouni Mattila***, **Atanas Gotchev****, **Emilio Ruiz*****
*** Laboratory of Automation and Hydraulics Engineering, Tampere University of Technology, Tampere, Finland**
**** Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland**
***** Fusion for Energy, Barcelona, Spain**

Abstract

Remote teleoperation of robotic manipulators requires a robust machine vision system in order to perform accurate movements in the navigated environment. Even though a 3D CAD model is available, the dimensions and poses of its components are subject to change due to extreme conditions. Integration of a stereoscopic camera into the control chain enables more precise object detection, pose-estimation, and tracking. However, the conventional stereoscopic pose-estimation methods still lack robustness and accuracy in the presence of harsh environmental conditions, such as high levels of radiation, deficient illumination, shiny metallic surfaces, etc. In this paper we investigate the ability of a specifically tuned iterative closest point (ICP) algorithm to operate in the aforementioned environments and suggest algorithmic improvements. We demonstrate that the proposed algorithm outperforms current state-of-the-art methods in both robustness and accuracy. The experiments are performed with a real robotic manipulator prototype and a stereoscopic machine vision system.

Introduction

Computer Aided Teleoperation (CAT) usually implies several different aspects or tools within the robotic operation chain. The main goal of the teleoperation in our application is to perform maintenance and tool manipulations with several kinds of objects inside a radioactive fusion reactor, where human presence is prohibited.[1]

In order to perform operations during the reactor maintenance break, a robotic manipulator must insert different tools inside several mounting holes for the different pre-defined reactor components. Even though the 3D CAD models of the components to be manipulated are known with high accuracy in advance, these elements are subject to small drifts in their poses, which have six degrees of freedom, and material deformation due to extreme heat and magnetic loads during machine operation. For precise and reliable teleoperation, the environment dimensions and poses have to be estimated accurately and converted into the robot's world coordinates [1]. Once that relation, that is, the rigid-body transformation is found, operations such as tool pickup, insertion, turning, retraction, and putting down can be made semi-automatic.

The problem of pose estimation, however, remains challenging, due to the harsh environment within the chamber. *Radiation tolerant* cameras are the only sensors capable of working in the chamber, and no stationary equipment is allowed. Apart from the low resolution and grayscale output of these cameras, other limitations connected to the environment are also present, including a high level of image noise due to the radiation; deficient illumination of the scene due to constraints on available light sources; non-Lambertian reflectance of shiny metallic surfaces and objects, etc. All these are difficulties that make any vision-based object detection and pose-estimation system problematic.

A previous study on pose-estimation CAT systems based on the 3D template matching algorithms showed significant limitations of the monocular approach [2]. In our application [3], we use a stereoscopic camera mounted to the last joint of a robot manipulator as a sensing tool to perform vision tasks, object detection, and pose-estimation. The same camera system can also be used by the operator, for instance when inspecting objects or the robot itself.

A stereoscopic camera system can reconstruct the geometry of a 3D scene based on stereo correspondences. Subsequently, it generates a depth map in the form of a grayscale image describing the geometry. We utilize this property in order to recover a *3D point cloud* representation of a scene, then try various *iterative closest point* (ICP) alignment approaches [4, 5] in order to detect and finally recover the pose of a target object.

Problems and Limitations

Current ICP methods are limited by the use scenario. Depth maps and point clouds generated by a stereoscopic camera system are significantly degraded due to various factors of the operating environment, and thus only a small portion of points can be trusted. For instance, the depth of shiny surfaces usually cannot be well estimated due to violation of the Lambertian reflectance model. High levels of noise can also result in false matches within textureless areas, and low-resolution grayscale imagery significantly limits the discriminative power of the stereo-matching algorithms. All these difficulties result in systematically erroneous depth values (outliers), which significantly disorient conventional general-purpose ICP methods.

Strong luminance gradients are the only features in the stereo images that can be trusted for their error-free behaviour. In the textureless and smooth scenes, strong image gradients usually correspond to object boundaries or significant changes in the surface (e.g., slope). In contrast to other robust image features, such as scale-invariant features (SIFT) [6] or speeded-up robust features (SURF) [7], image gradients are much denser and tolerate image noise.

Nevertheless, using the object boundaries as matching primitives can also limit the selection of the underlying ICP method.

[1] The nuclear-fusion reactor, constructed within the ITER project (http://www.iter.org).

As the surface normals generally cannot be estimated at borders and object edges, only *point-to-point* minimization is possible. More advanced *point-to-plane* [8] or generalized *plane-to-plane* [9] minimization approaches cannot be utilized.

The recently proposed *edge-point ICP* method [4] is capable of operating within this type of constraints. The method successfully works when the estimated point cloud contains few outliers and when a good initialization point is provided. From the algorithmic point of view, outliers are not only wrongly estimated depth values, but also points that have no corresponding points in the target (model) point cloud, or vice-versa.

Another substantial property of depth-from-stereo methods is the generation of content-dependent occlusion artifacts in their output. Occlusion hole-prediction methods exist, but they all rely on high-quality depth of the neighboring zones and use some guessing mechanisms, which is not allowed in precise alignment tasks. During the preparation of reference point clouds, based on the supplied CAD models, such artifacts are usually not taken into account, as it is not possible to predict from which viewpoint the object will be captured. Thus, large numbers of reference points may become outliers, with no corresponding point in the estimated cloud. Depending on the number of mismatched points, performance of the ICP alignment can be seriously degraded.

Contributions

In this paper we propose an efficient method to increase the robustness and the accuracy of the ICP alignment in which target point clouds are estimated using stereoscopic capture in the harsh industrial environments. We use the *approximate planarity* assumption in order to recover good initialization points for the ICP algorithm and illustrate its suitability for successful convergence. In contrast to conventional methods, we also use dynamically sampled reference point clouds, especially targeted to each particular stereo-observation. We model artifacts appearing in the depth-from-stereo methods in order to minimize the number of outliers in the reference clouds and thus increase final alignment accuracy.

Prior Art

Depth-from-Stereo

Estimation of the scene geometry from a binocular camera setup is usually called *stereo-matching* or the *depth-from-stereo* problem. Even though this field is already well developed, and many advanced techniques are available, in our problem we not only required estimating the depth but also correctly manipulating the depth values, projecting them back to the 3D space with real-world coordinates. Therefore, conventional stereo-matching methods, based on stereo-image rectification [10], might underperform due to the introduction of artificial camera transforms and excessive image interpolation steps. Moreover, a deviation from geometrically parallel camera configuration is possible (e.g., the camera optical axes might be crossed), thus introducing substantial image deformation in rectification-based methods.

Instead, *plane-sweeping depth estimation* methods, using calibrated camera parameters, allow direct processing of the captured imagery [11]. Figure 1 illustrates the depth-estimation method, based on the plane-sweeping principle. In this method, the entire observable scene is divided into a number of fronto-parallel planes (hypothesizes), where stereo correspondences might be found. Such hypothesizes can be selected for example by selecting the possible depth range (i.e., minimum and maximum possible depth values) and number of layers, which controls the trade-off between fidelity and computational complexity of the method.

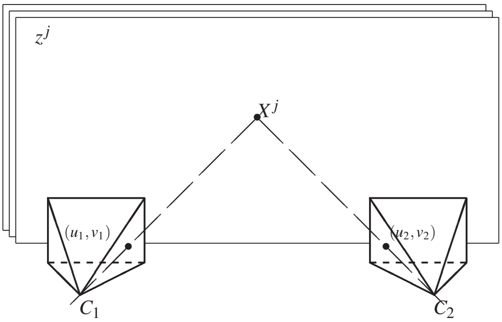


Figure 1. *Illustration of the plane-sweeping principle of the depth-from-stereo estimation methods*

For every hypothetical depth z_j, one can project a pixel (u_1,v_1) from a reference camera to a 3D space, using pre-calibrated camera matrix C_1:

$$\mathbf{X}^j = C_1^{-1}\dot{\mathbf{x}}_1, \tag{1}$$

where $\dot{\mathbf{x}}_1$ is the homogeneous projective coordinate of a current pixel $\dot{\mathbf{x}}_1 = (u_1 \cdot z_j, v_1 \cdot z_j, z_j, 1)^T$, $\mathbf{X}^j$ is the resulting point coordinate in a 3D space; and $j = 1,..,N$ where N is the selected number of layers.

Every obtained 3D point $\mathbf{X}^j$ can be further projected onto the sensor plate of a second camera using a similar equation:

$$\dot{\mathbf{x}}_2 = C_2\mathbf{X}^j \tag{2}$$

where $\dot{\mathbf{x}}_2$ is a projective pixel position in a second camera image plane, and the actual pixel coordinates can be recovered as:

$$u_2 = \frac{\dot{\mathbf{x}}_2.x}{\dot{\mathbf{x}}_2.z} \quad , \quad v_2 = \frac{\dot{\mathbf{x}}_2.y}{\dot{\mathbf{x}}_2.z} \tag{3}$$

Similarly to conventional rectification-based methods [10], one can construct a 3D cost volume, in which pixel dissimilarities are calculated between the original pixel in the reference camera and the corresponding pixel in the second one:

$$C(u,v,j) = \|I_1(u_1,v_1) - I_2(u_2,v_2)\|, \tag{4}$$

where I_1 and I_2 denote the first and second images, respectively, and because the (u_2,v_2) coordinates are not necessarily integers, the corresponding sampling should be performed for instance with bilinear interpolation.

After appropriate cost aggregation [11], the depth map can be recovered by using the so-called *winner-takes-all* approach:

$$Z_1(u,v) = z_{\hat{j}}, \hat{j} = arg\min_j \tilde{C}(u,v,j), \tag{5}$$

where $\tilde{C}(\cdot)$ denotes the aggregated cost volume.

The coordinates of the point cloud in the reference camera can now be reconstructed using the same equation as in (1), replacing $z_{\hat{j}}$ with the estimated value.

ICP Methods

Since the first invention of the ICP method [12], many updates have been proposed [5, 4, 13]. One of the directions for improvements has been reducing the influence of outliers on the global error. Thus, many widely accepted techniques remove too many point correspondences while calculating global error [5]. A number of linearized methods were suggested using SVD [14], quaternions [15], and dual quaternions [16] for minimizing the error metric with a closed-form solution. High quality of the sensed (input) point cloud is an essential requirement for conventional ICP algorithms. A relatively moderate fraction of outlying points in the input cloud can significantly degrade performance of the method, thus preventing its usage for real-world applications. This is an important aspect for point clouds estimated via stereoscopic camera in harsh environments. As the passive vision systems (including depth-from-stereo methods) usually fail in the presence of textureless or shiny (i.e., non-Lambertian) surfaces, their depth maps become corrupted with a high number of false estimates. Consequently, input point clouds could be contaminated with outliers, thus preventing use of the technique for pose estimation tasks.

Edge-point ICP [4] uses an additional type of filtering step, where points not connected to a strong image gradient are removed from the point cloud. Even though this operation can significantly reduce the number of available points in the cloud, their discriminative power significantly improves, thus resulting in better performance, especially in cases when textureless areas dominate the scenes.

HandEye Calibration and World Coordinates

The object pose in terms of camera coordinates has to be transformed into robot world coordinates, for which hand-eye calibration [17] is needed. Figure 2 indicates the relationship between the robot end-effector, the camera, and the object in the world coordinates with the formula:

$$P = R \cdot X \cdot A \tag{6}$$

where P is the required pose of an object, A is the estimated alignment in the camera coordinate space, X is the eye-hand transformation matrix, and R is the current position of the robot hand/wrist.

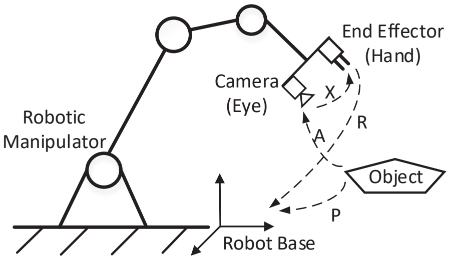


Figure 2. Hand-eye calibration and world coordinates

Sampling of CAD Models

Sampling of CAD models is usually done once during algorithm development and all estimated points in the point cloud are matched against this reference cloud.

An example of sampling of CAD models is provided in Figure 3, which shows a sensed point cloud before alignment.

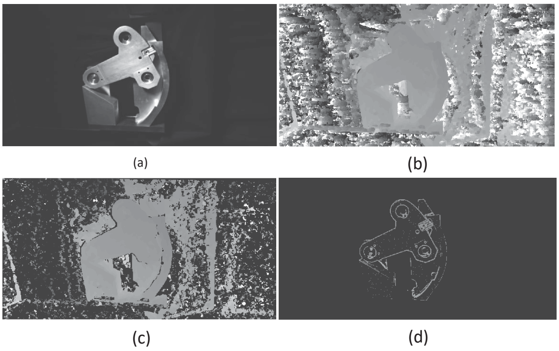


Figure 3. Sampling of CAD model: (a) given image, (b) raw depth, (c) depth after L2R, and (d) depth after sampling

Proposed Method

Typical industrial environments, which are also considered in our application, usually contain many planar surfaces. Such surfaces are easier to manufacture and they are more convenient when constructing large-scale structures. Target objects can also be considered as having at least one major planar surface, facing the stereoscopic sensor. Even though a strict planarity constraint may not be fully satisfied due to obstacles and other features on the object surface, often we can still rely on the *approximate planarity of the surfaces*. In our method, we propose imposing such constraints in order to estimate a good initialization point for the alignment algorithm and to avoid point mismatches due to occlusion artifacts.

The point cloud estimated from a scene can be analyzed for the presence of plane structures. This can be done, for instance, using the random sample consensus (RANSAC) [18] plane-fitting method. General plane-fitting methods in 3D point clouds usually utilize a generalized plane equation:

$$ax + by + cz + b = \mathbf{a}^T \hat{\mathbf{x}} = 0, \tag{7}$$

where $\mathbf{a} = [a, b, c, d]^T$ is the vector of plane parameters to estimate, and $\hat{\mathbf{x}} = [x, y, z, 1]^T$ is the homogenous point coordinate from the cloud.

A conventional way to perform the analysis is to select three random points from the cloud, fit the plane parameters and estimate the number of other points that belong to the same plane with some kind of tolerance. The process is repeated multiple times, and the plane equation containing the largest number of inliers is considered the largest plane found in the scene.

As the point cloud estimated with the stereo-camera setup usually does not capture highly slanted or parallel-to-the-optical axis planes, we can utilize a relaxed plane equation:

$$z = ax + by + c = \mathbf{a}_s^T \hat{\mathbf{x}}_s. \tag{8}$$

Following a similar RANSAC methodology, the matrix of three selected points X and the vector of corresponding depth values $\mathbf{z}$ can be utilized to recover the plane parameters using the

Moore-Penrose pseudo-inverse:

$$X = \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{pmatrix}, \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \tag{9}$$

$$\mathbf{a}_s = \mathbf{z} \cdot X^T (XX^T)^{-1} \tag{10}$$

Here, $n = 3$ for the initial plane estimation and can be arbitrary during the plane refinement stage, when plane parameters are estimated using all the found inliers. Inliers can be selected using pre-defined threshold value θ, as points whose distance to plane is lower than a threshold $|ax_i + by_i + c - z_i| < \theta$.

The parameter θ can also control the expected proximity of an object surface to a plane model. For objects with dominating planarity, θ can be reduced to account only for possible depth estimation errors, while for objects containing many bumps or cavities, larger values of θ can be beneficial.

When the CAD model is aligned with its major plane (i.e., model origin and X-Y coordinates belong to it), the obtained plane parameters can directly be used to estimate good initialization of the rotation matrix. Two of the Euler angles can be estimated as:

$$\beta_x = tan^{-1} b, \tag{11}$$

$$\beta_y = -tan^{-1} a, \tag{12}$$

where β_x and β_y are Euler angles around X and Y axes, respectively.

Rotation around the Z axis cannot be estimated by such a coarse method; however, the generic assumption of vertical camera orientation can still be used to provide meaningful initialization. As a guess for an initial translation, we use the median-centroid of a point cloud. This assumption may introduce certain limitations of the method, particularly when a significant part of the surrounding scene is also visible to the stereo camera setup.

Advanced CAD Model Sampling

Apart from the transform matrix, we also propose a method to reduce the number of mismatches in the point cloud estimated by using the depth-from-stereo method. As the rotational component in the true underlying transformation can be arbitrarily large, projective distortions appearing in the sensed images may be significant. We use dynamic CAD model re-sampling as a mechanism to reduce possible outliers in the model point cloud, hence improving the accuracy of the final alignment.

In conventional ICP methods, the model point cloud is usually statically defined and re-used every time a new observation is made. In practical cases, however, excessive numbers of mismatched points prevents this use.

We use heuristics in order to remove possible outliers from the reference cloud. For instance, a left-to-right correspondence check rendering is done with the transform found in the initial alignment step. We render images for both the reference and secondary camera (with the same configuration as in the stereoscopic setup). This allows us to apply the same left-to-right correspondence check as in the estimated depth. We use rendered images of a CAD model to find strong edges in the scene and prepare a point cloud according to the same process as for the source point cloud. Applying these heuristics, the reference cloud contains the same amount of occlusion and similar results with regard to the edge properties as the source cloud.

For efficient processing, we propose the following scheme. Figure 4 shows the procedure per frame.

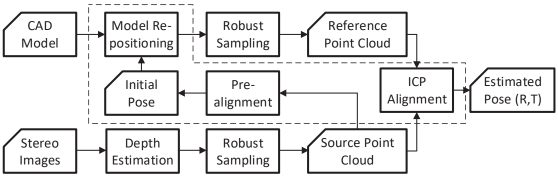


Figure 4. *Flowchart of proposed ICP implementation*

Standard edge-point ICP initializes its model point cloud by sampling only once, which is not robust in the case of a stereoscopic camera. Thus, in our proposed ICP (new blocks within the dashline), we render our CAD model such that it shows approximately what the camera is seeing. The model point cloud is estimated every time using the pre-alignment, and we sample the CAD model relative to our initial estimated alignment.

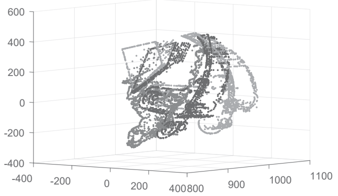


Figure 5. *Comparison of sampled point cloud; Red, sensed; Green, standard ICP; Blue, proposed ICP*

As we can see in Figure 5, the blue point cloud has a better initialization point than the green one, which helps to overcome the issues with local minima.

Experiments

In order to validate the proposed method, we run two sets of experiments, based on photographs of two target objects, namely a CLS mockup and a knuckle, illustrated in Figure 6. The CLS mockup is made of steel, from laser-cut sheet material welded together with high precision, while knuckle was mainly 3D printed with a fused deposition modeling (FDM) printer, sanded, and then painted with shiny metallic paint; the manufacturing accuracy is thus lower for the knuckle. Surrounding elements in the knuckle were made with precise steel-cutting approaches. Overall, both target objects well represent the expected reflectivity and textureless properties of the application, as well as corresponding to an underlying CAD model with tolerances up to 0.2 mm.

In order to obtain a comprehensive set of experimental data, we gathered a significant number of stereo-images using different camera offsets, orientations and different illumination conditions. Overall, 31 stereo-pairs per target object were acquired using the calibrated stereo camera setup. Camera positions were selected such that for the closest (to the object) camera position, the target object barely fit in the camera view, while for the position further away from the target objects, it occupies just a small fraction of the image, representing a wide range of distances. Figure 6 shows two of the acquired images.

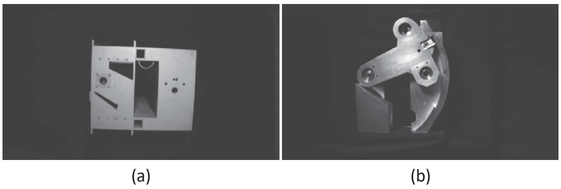


Figure 6. *Sample images of (a) “CLS-mockup” and (b) “knuckle” objects used in the experiments.*

The main goal of our experiment was to estimate the robustness, reliability, and accuracy of the proposed method as well as to compare it with competing approaches. In order to estimate robustness, for every acquired stereo-pair we independently ran the alignment algorithm 30 times and measured the number of false pose estimates, that is, when the aligned object completely disagrees with the acquired data.

This can be done in semi-automatic mode, in which the software asks the operator to confirm whether current alignment was successful. Figure 7 shows two alignment results, where the CAD model was projected to the camera space and rendered according to the estimated object pose. Two images, the acquired and the rendered one, are combined together in different color channels and presented as a single RGB image, which we refer to as the “augmented” image. Such representation can easily be evaluated by the operator for correctness of alignment and thus be selected as correct or not.

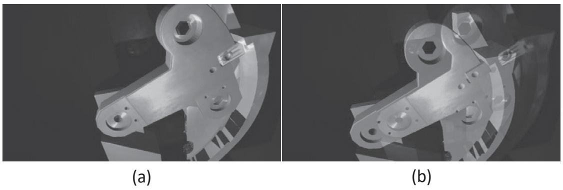
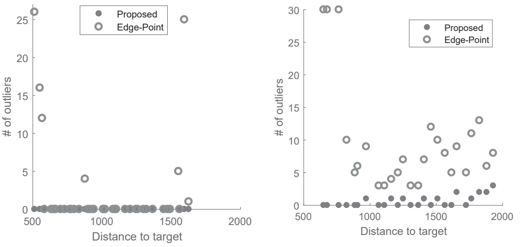


Figure 7. *Example of ICP alignment with augmented images: (a) successful, (b) unsuccessful*

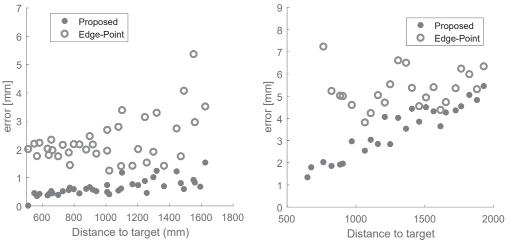
The 31 observations of every stereo-pair provided a number of pose estimates, including the relative rotation and the translation between the camera and the CLS mockup or knuckle. We used semi-manually estimated positions as the threshold for selection of correct estimates, or inliers. All inliers from these observations are averaged together in order to obtain the centroid of the estimated points. Now, by measuring the Euclidean distance between the centroid and every other estimate, one can obtain the average displacement (deviation) for this particular stereo-pair. While taking an estimated Z (depth) value as a reference variable, one can plot a figure in which the horizontal axis represents depth (Z-distance between camera and the object) and the vertical axis shows the respective deviation value. Figure 8 and Figure 9 show these graphs for a few different experiments.

As we can see, when the target object is too close to the camera, it can no longer observe all the distinctive edges. This indicates a general lower limit of the pose-estimation system where too-close observations are not reliable. In addition, the images show that both the CLS mockup and the knuckle achieve good accuracy and stability within the middle range. This can be explained as fairly consistent behavior within the expected operational range. With the increase in distance, the repeatability error grows but also becomes unstable, which could suggest the existence of an upper limit for the system. This limit, however, was not reached during these sets of experiments.



(a). *“CLS-Mockup”* ***(b).*** *“Knuckle”*

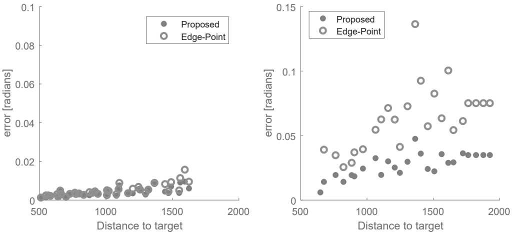
Figure 8. *Number of outliers for CLS mockup and knuckle datasets.*



(a). *“CLS-Mockup”* ***(b).*** *“Knuckle”*

Figure 9. *Position stability (repeatability) for CLS mockup and knuckle datasets.*

A similar procedure can be done for the rotational part of the found transforms. We extract the rotation matrix from each estimated transform and convert them to a vector of Euler angles. Then, the mean Euler angle value for one image is chosen as the correct rotation, and the error in the rotations is expressed as the difference between the mean Euler vector and the rest of the vectors. In order to obtain a single variable out of all the observations, we convert the angular error vectors to a list of combined errors, taking the L2 norm of each vector. Then, the mean value of all combined errors is taken to represent the integral error metric for one particular image. The process is repeated for every image in the dataset in order to obtain a closed curve.



(a). *“CLS-Mockup”* ***(b).*** *“Knuckle”*

Figure 10. *Angular stability (repeatability) for CLS mockup and knuckle datasets.*

Figure 10 exhibits similar performance of the method as in Figure 9. The optimal range of the system is reached in the range of 65 to 100 cm, and the the overall integral angular error is on the order of 0.02 to 0.05 rad.

Conclusion

The measurement of repeatability error demonstrates fairly consistent behavior, even though the target object was imaged from different perspectives. Overall, our proposed method has shown more robustness and accuracy than the standard edge-point ICP method in terms of the number of outliers and precision of pose estimation. The results verify the effectiveness of the proposed method.

Acknowledgment

The work leading to this publication has been funded in part by Fusion for Energy and TEKES under Grant F4E-GRT-0689. This publication reflects the views only of the authors, and Fusion for Energy or TEKES cannot be held responsible for any use which may be made of the information contained herein.

References

[1] Pritam Prakash Shete, Abhishek Jaju, Surojit Kumar Bose, and Prabir Pal, "Stereo vision guided telerobotics system for autonomous pick and place operations," in *Proceedings of the 2015 Conference on Advances In Robotics*. ACM, 2015, p. 41.

[2] Z Ziaei, A Hahto, J Mattila, M Siuko, and L Semeraro, "Real-time markerless augmented reality for remote handling system in bad viewing conditions," *Fusion Engineering and Design*, vol. 86, no. 9, pp. 2033–2038, 2011.

[3] L. Niu, O. Suominen, M.M. Aref, J. Mattila, E. Ruiz, and S. Esque, "Eye-in-hand manipulation for remote handling: Experimental setup," in *International Conference on Robotics and Mechatronics*. IOP Conference Series, Hong Kong, 2017.

[4] Masahiro Tomono, "Robust 3d slam with a stereo camera based on an edge-point icp algorithm," in *IEEE International Conference on Robotics and Automation*, Kobe, Japan, May 12-17 2009.

[5] Dmitry Chetverikov, Dmitry Stepanov, and Pavel Krsek, "Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm," *Image and Vision Computing*, vol. 23, no. 3, pp. 299–309, 2005.

[6] David G Low, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[8] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat, "Comparing icp variants on real-world data sets," *Autonomous Robots*, vol. 34, no. 3, pp. 133–148, 2013.

[9] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun, "Generalized-icp.," in *Robotics: science and systems*, 2009, vol. 2, p. 435.

[10] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[11] S. Smirnov, A. Gotchev, and M. Georgiev, "Comparison of cost aggregation techniques for free-viewpoint image interpolation based on plane sweeping," in *Ninth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM*, 2015.

[12] Paul J Besl, Neil D McKay, et al., "A method for registration of 3-d shapes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[13] Roberto Marani, Vito Reno, Massimiliano Nitti, Tiziana D'Orazio, and Ettore Stella, "A modified iterative closest point algorithm for 3d point cloud registration," *Computer-Aided Civil and Infrastructure Engineering*, vol. 31, no. 7, pp. 515–534, 2016.

[14] K Somani Arun, Thomas S Huang, and Steven D Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 5, pp. 698–700, 1987.

[15] Berthold KP Horn, Hugh M Hilden, and Shahriar Negahdaripour, "Closed-form solution of absolute orientation using orthonormal matrices," *JOSA A*, vol. 5, no. 7, pp. 1127–1135, 1988.

[16] Michael W Walker, Lejun Shao, and Richard A Volz, "Estimating 3-d location parameters using dual number quaternions," *CVGIP: image understanding*, vol. 54, no. 3, pp. 358–367, 1991.

[17] Roger Y Tsai and Reimar K Lenz, "Real time versatile robotics hand/eye calibration using 3d machine vision," in *Robotics and Automation, 1988. Proceedings., 1988 IEEE International Conference on*. IEEE, 1988, pp. 554–561.

[18] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

Author Biography

Longchuan Niu is a PhD student at TUT, Tampere, Finland, in the field of robotics and computer vision. He received his M.Sc. with distinction at TUT in 2000. After that he worked at Nokia R&D Finland as a senior software engineer until joining TUT in 2016.

Sergey Smirnov received his B.Sc. degree from Yaroslavl Demidov State University, Russia (2003), and M.Sc. degree from TUT, Finland (2010). His research interests include depth image-based rendering (DIBR), image analysis, and 3D reconstruction and visualization.

Jouni Mattila received his M.Sc. and Dr. Tech. degrees in 1995 and 2000, respectively, both from TUT, Tampere, Finland. He is currently a Professor in Machine Automation in the Laboratory of Automation and Hydraulics, TUT. His research interests include machine automation, developing nonlinear model-based control systems for robotic mobile manipulators and off-highway machinery, etc. He is currently a Technical Editor of the IEEE/ASME Transactions. on Mechatronics.

Atanas Gotchev is a professor of the 3D Media Group at the Laboratory of Signal Processing at TUT and serves as Director of the national-wide research facility Centre for Immersive Visual Technologies (CIViT). He has broad competence of 3D imaging gathered as Chair of Research Exchange Committee of FP6 3DTV Network of Excellence, Scientific Coordinator of the FP7 project Mobile3DTV, and Project Manager of the FP7 Marie Curie IAPP Action PROLIGHT. His research expertise is in the areas of sampling and reconstruction of multi-dimensional signals; multi-sensor 3D scene sensing and reconstruction, and signal processing for ultra-realistic displays. He has co-authored about 170 scientific publications and has eight invention disclosures.

Emilio Ruiz Morales received his M.Sc. degree in 1990 from ULB, Brussels. He is currently a Senior Engineer and Technical Responsible Officer of several R&D projects at the Remote Handling Project Team/ITER delivery of the Fusion For Energy agency. His background expertise and research work are in control systems and surgical, nuclear and remote handling robotics.

PUBLICATION

II

Clustering Analysis for Secondary Breaking Using a Low-Cost Time-of-Flight Camera

L. Niu, M. M. Aref and J. Mattila

2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP) 2018, 318–324

DOI: `10.1109/ICICIP.2018.8606682`

Clustering Analysis for Secondary Breaking Using a Low-Cost Time-of-Flight Camera

Longchuan Niu, Mohammad M. Aref, Jouni Mattila
Laboratory of Automation and Hydraulic Engineering
Tampere University of Technology
Tampere, Finland
longchuan.niu@tut.fi, m.aref@ieee.org, and jouni.mattila@tut.fi

***Abstract*—The integration of robust perception in a heavy-duty manipulation control system is an enabler for autonomous mining. This paper aims to analyze performance and robustness of clustering methods for object recognition during the secondary breaking stage of mining. Secondary breaking refers to breaking over-sized rocks into smaller pieces for the purpose of grinding and extraction of valuable ores and minerals. Therefore, recognition of rock pieces is the detection of unstructured targets within a structured environment. The clustering methods are experimentally evaluated by several sets of scenes of point clouds as outputs of a Time-of-Flight camera (ToF). The challenges of rock detection from sparse 3D point cloud data are addressed. In outdoor conditions, ToFs generally provide coarse but robust output in short sample times. Therefore, some clustering methods can be prone to numerical and statistical errors. This paper highlights the weaknesses and strengths of three methods for the secondary breaking application. We propose an algorithmic method for exploiting the existing clustering and segmentation methods efficiently in the detection loop to determine a suitable contact point and approaching angle for a hydraulic jack hammer. The results verify effectiveness of the proposed approach for scattered outputs of low-cost ToFs.**

***Index Terms*—range sensing, time-of-flight camera, automatic extraction, 3D point clouds, clustering**

I. Introduction

In the mining and construction industry, valuable minerals are often extracted from the earth by blasting. Secondary breaking is needed to ensure effective processing when the blasted rock pieces are too large for the feeder or crusher. A manipulator carrying a hydraulic rock hammer is used to break these over-sized rocks into smaller pieces as shown in Figure 1. These breaker manipulators are operated by human operators, but such operational tasks are repetitive, mentally and physically demanding. To achieve improvements in mining process autonomy, production rates and performance, will require that artificial intelligence methods be integrated in the perception and control elements of operational tasks.

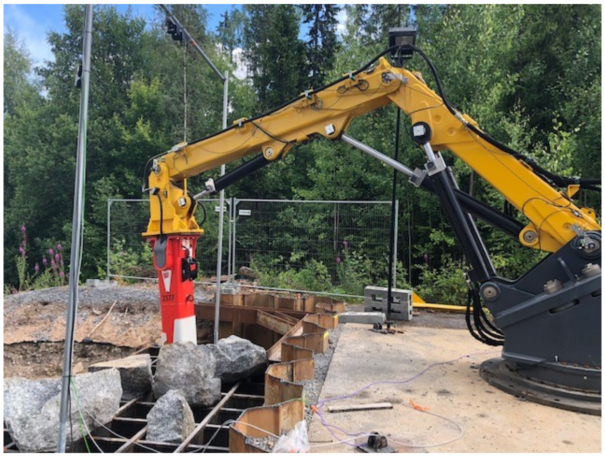


Figure 1: Secondary breaking with guidance of the ToF camera

For this application, different sensor types are available based on different measurement technologies; these include stereo vision cameras, projective light cameras, and time-of-flight cameras. All of these sensors are capable of generating arrays of 3D positions occupied by objects as a point cloud. For the sake of robustness and reliability in outdoor conditions, we consider ToF cameras as the sources of point clouds. The aim of this research is to analyze the point clouds to obtain necessary information of stone pieces for the hammer manipulator.

There is comprehensive research on the recognition of rocks and unstructured minerals. A considerable amount of relevant research has been done on discontinuity studies of outcrop rock mass in mining, such as the detection of outcrop rock mass by block modeling [1], by fuzzy k-means clustering [2], rock mass joints recognition using Density-Based Scan Algorithm with Noise (DBSCAN) [3], or rockfall detection using DBSCAN clustering [4], planar surface detection [5], [6]. All the addressed methods are tolerant to limited failures in detection and imprecise segmentation in preprocessing of point clouds.

Although these methods are capable of clustering in certain conditions, they do not address autonomous secondary breaking applications when the quantity of objects (clusters) is unknown. In comparison with outcrop rock mass segmentation methods, secondary breaking requires an exact prediction of the number of rocks on the metal grid and accurate localization of each of them. Incorrect segmentation may cause the manipulator's hammer to hit the rock and possibly damage the metal grid underneath. Therefore, failures can significantly affect the lifetime of the robot and production rate.

The robotic perception process starts by sensing of rocks using the ToF camera. The generated cluttered scene in the form of 3D point clouds has to be processed for performing normalization as well as the filtering out of the background objects, outliers, and noise. The remaining cleaned data consists of sets of surface point clouds of rocks in unknown shapes and configurations measured from a single viewpoint.

The next step is to further segment the data into meaningful subsets representing pieces of rocks by clusters of data using clustering methods. Performing segmentation using clustering makes it possible to discover an arbitrary number of objects of any shape in the data. It allows segmenting objects in the point cloud without need of templates, textures and geometries.

Although the recognition of objects includes the detection of their quantity, it is considered as an input for many clustering methods. This means that the number of rocks in the field of view need to be calculated. This is the main limitation that prevents use of many state-of-the-art methods for clustering and motivates us for proposing the method in section III. The proposed method overcomes this limitation and recognizes the rocks and number of rocks, namely the K parameter. To obtain a baseline for comparison, we use the proposed method and extracted K, for the purpose of comparing them with the outcomes of similar methods.

In a review of well-known clustering methods addressed in [7], such as the K-means, Gaussian mixture models, Ward, and Spectral, we recognize that the K parameter is a crucial piece of information assumed to be known in the methods discussed in Section II. Therefore, as explained in Section III, our first step toward clustering needs to be the estimation of the number of clusters, K.

Among the clustering methods which do not require prior information about the number of rocks in the scene, DBSCAN has successful applications in dealing with spatial point cloud data [3] and [4]. However, for proper functionality, DBSCAN also requires adjustment of two parameters depending on the scene and is therefore not suitable for autonomous predictions. Moreover, our experiments in Section IV demonstrated that its accuracy for localization of rocks is not consistently acceptable. Other clustering methods such as Affinity Propagation and Mean Shift been unsuccessful in coping with our data. Therefore, we demonstrate experimental results of two methods, DBSCAN and Ward, as well as our proposed method.

For the purpose of experiments, we exploit sets of point clouds gathered by ToF. The reason why we use a ToF camera is due to its close-to-real-time capabilities and promising spatial resolution. Compared to other conventional point scanner cameras such as stereoscopic cameras [8] and RGB-D cameras [9], ToF cameras have a number of advantages, including simplicity, speed, affordability, and efficiency. ToF cameras are also able to measure the distances within a scene in a single shot.

On the other hand, unlike an RGB-D camera, which provides us with data rich in features like colors and texture, a time-of-flight (ToF) camera only provides depth data. The sparse 3D point cloud data from a ToF camera contains position and intensity in grayscale of surface points gathered from the camera's viewpoint. The limited number of features can be challenging for object recognition. Moreover, sometimes highly specular objects in the scene results in the ToF camera failing to capture objects.

This paper is organized as follows. Section II introduces cluster analysis and clustering methods. Section III describes the proposed method and the algorithm for clustering. Section IV describes the experimental framework and point cloud data used to evaluate the performance of benchmark clustering methods in comparison with the proposed method. The results are demonstrated in plots and tables for comparison. Finally, in the conclusion section we summarize our findings for clustering of point cloud data from secondary breaking experiments.

II. Clustering

Clustering is the process of finding similarities among individual points so that they can be segmented. Many methods exist for clustering arbitrary data. As we do not have a prior knowledge about the number of clusters (i.e. K) in our data set from the ToF camera, many methods are unable to be used.

One study [10] proposes that the Bayesian Information Criterion (BIC) be maximized, while another approach [11] is to start with a large value for k and keep removing centroids (reducing k) until it no longer reduces the description length. [12] starts with one cluster, then continues to split clusters until the points assigned to each cluster have a Gaussian distribution. Unfortunately, there is no explicit answer as to which method to use for a spatial clustering problem.

Here are some selected clustering methods from the Scikit-Learn library [7]. Of course, the same rule is applicable to other clustering techniques within the same category.

TABLE I: Determining the number of clusters in a data set

Independent of K parameter	***Required Prior K***
DBSCAN	K-means
Affinity propagation	Gaussian mixture model
Mean shift	Spectral
The Proposed Method	Ward hierarchical

In the following, we address major clustering algorithms in detail. However, for the purpose of experimental evaluation, we only select DBSCAN and Ward together with our proposed method because of their better performance and convenient requirements.

A. Centroid-based clustering, K-means

Choosing the most straightforward spatial clustering method such as K means will in turn require that the user choose the correct number of expected clusters (i.e. K value). Though the Elbow method [13] could be used to determinethe K value, a consistent prediction is not ensured. Another issue with K-means is that it is very sensitive to the initial position and random results appear from the same input data.

B. Distribution-based clustering, Gaussian mixture model

The Gaussian mixture model (GMM) uses the expectation-maximization algorithm on the prior K value. The resulting clusters can easily be defined as objects likely belonging to the same distribution. This algorithm is not suitable for our data, as it converges to local optimum and multiple runs produce different results.

C. Spectral clustering

In order to segment a point cloud through spectral clustering, the point cloud has to be represented as a graph. This is carried out by connecting each point with its neighbors and assigning the edge a weight that describes the similarity. The segementation problem is resolved by NP hard [14]. This method is highly dependent on the similarity matrix and prior K value.

D. Affinity propagation

Affinity propagation is exemplar-based clustering which iteratively searches the set of data points until it best describes the input data found based on the similarities between them [15]. The method does not require the K value before running the algorithm; instead, there are two parameters: *preferences* determines the number of clusters, the higher its value the more clusters it generates, and *damping factor* decides the speed of the algorithm's convergence, preventing oscillation.

E. Mean Shift

Mean shift clustering [16] is built on the concept of kernel density estimation. The method works by shifting a kernel on each point toward a higher density in the data set until they converge. It has only one parameter *bandwidth*, which determines the number of clusters K; however, K may not be a monotonic function of bandwidth. In such a case it will likely fail to find all clusters.

F. Linkage, Ward

Ward linkage [17] involves an agglomerative hierarchical clustering algorithm. It is known for being a minimum variance method based on the sum of squares of errors (SSE) of each cluster; i.e., the sum of squares of deviations from the cluster centroid. Giving K as a parameter, it will attempt to merge K clusters by analyzing all possible pairs of joined clusters and identifying which joint produces the smallest increase in SSE. In spatial agglomeration clustering, the distance measure between two clusters K and L is usually defined as a squared euclidiean distance.

$$\Delta(K,L) = \sum_{j\epsilon K\cup L} ||\vec{x}_j - \vec{m}_{K\cup L}||^2 - \sum_{j\epsilon K} ||\vec{x}_j - \vec{m}_K||^2 - \sum_{j\epsilon L} ||\vec{x}_j - \vec{m}_L||^2 = \frac{n_K n_L}{n_K + n_L} ||\vec{m}_K - \vec{m}_L||^2$$

where $\vec{m}_K$ and $\vec{m}_L$ are mean vectors within cluster K and L, n_K and n_L are the number of points in cluster K and L respectively. Δ is the merging cost of combining the clusters K and L.

G. Density-based clustering, DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a typical density-based clustering algorithm [18]. The algorithm does not require prior K as the parameter; instead, it has two paramters: ϵ is the distance for searching neighboring points in 3D space and minimum points per cluster $MinPts$. All neighboring points in euclidean space within ϵ distance will be connected to form a density-connected cluster. Any unallocated points with a distance further then the predefined threshold ϵ with its nearest neighbor, or points within a cluster whose size is less than $MinPts$ will be regarded as noises. This is a useful spatial clustering method for our data.

While the theoretical foundation of the benchmark methods is excellent, many of them are not suitable for spatial data. Our point cloud data consists of uneven-sized clusters, whose geometry is non-flat. Furthermore, no prior information about the number of clusters and wideness of their coverage. The following cluster methods are therefore unsuitable because of their significant dependency on prior knowledge: K-means, Spectral , the Gaussian mixture model, Affinity propagation and Mean shift.

III. The proposed method

We tackle the clustering problem using euclidean clustering. This is a simple data clustering approach in a euclidean sense in which points that are closer to each other are clustered together by making use of a 3D subdivision of the space.

A. Automatic extraction of rocks

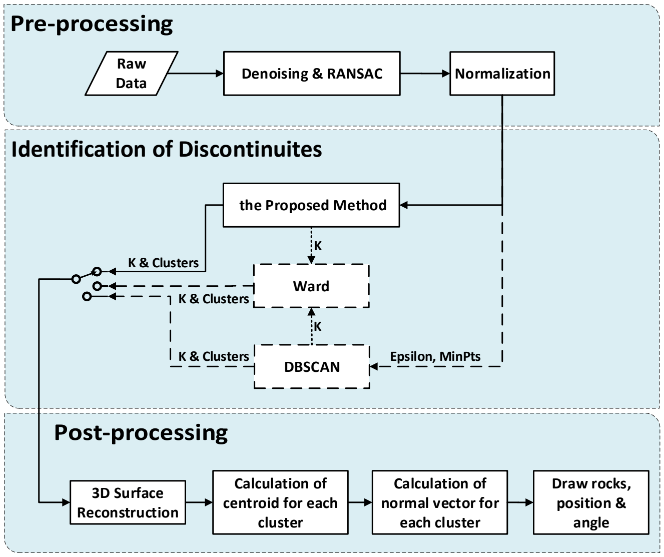

Pre-processing
Raw Data
Denoising & RANSAC
Normalization
Identification of Discontinuites
the Proposed Method
K
K & Clusters
Ward
K & Clusters
K
K & Clusters
DBSCAN
Epsilon, MinPts
Post-processing
3D Surface Reconstruction
Calculation of centroid for each cluster
Calculation of normal vector for each cluster
Draw rocks, position & angle

Figure 2: Flowchart detailing the automatic extraction of rocks

Figure 2 presents three main steps of automatic extraction of rocks; i.e., pre-processing, identification of discontinuites and post-processing, First, the raw data must be preprocessed with a denoising filter, which is effective when dealing with a point cloud which has significant noise and outliers. A random sample consensus (RANSAC) plane filtering algorithm is then applied to remove the ground and the metal grid out of the scene. This approach relies solely on object shapes to perform segmentation. In this situation, orphan points among raw point clouds are also considered as outliers [19].

Automated identification of discontinuites sets using clustering techniques, especially the estimation of the number of clusters (i.e., K value) within the scene, is a bottleneck in the whole segmentation process. This is resolved by the proposed clustering method, in which we compare its performance with DBSCAN, and Ward, which requires the input parameter K from either the proposed method or DBSCAN. The data normalization in pre-processing is performed according to clustering methods.

In the post-processing section and in addition to visualization of cluster and 3D surface geometric reconstruction for all rocks, we also calculate the centroid position of each cluster and normal vector of them. These values will be used by a manipulator.

B. The Proposed Unsupervised Clustering Method

As a matter of fact, the density of the point cloud is homogeneous as collected from the ToF camera. Moreoever, the sampling interval and beam divergence is fixed and the angle of observation is 0 degrees, as shown in Figure 1.

Segmenting such data can be done by performing clustering based on spatial neighborhood; meaning points that are closer in the 3D space form a cluster. We intend to leverage such properties of our data, as we notice that the surface point clouds of rocks are linearly sampled at a constant interval in camera coordinate with product specific configurations along each axis. This means we can utilized these configurations for clustering in order to achieve the best segmentation results. The proposed algorithm is derived through absolute point density values presented specific to the ToF camera. Different cameras may require calibration before correction factors are established.

In addition to predicting the number of clusters, another clear advantage of this algorithm is that no input parameters are required, only the needed configurations to define noises. Similar to DBSCAN's parameter *MinPts*, the recommendation from [18], i_{min} is set to 4 in the algorithm initialization. This number suits our application as well, especially given the size requirements of the application and the robot's metal grid as its tabletop.

IV. Experiments and Results

The proposed approach is evaluated by comparison against the benchmark clustering methods DBSCAN and Ward (from the Scikit-Learn 0.19.2 clustering library [7]), both of which utilize the module sklearn.cluster written in Python. DBSCAN and Ward methods are selected because they outperform the other methods addressed in Section II in their clustering performance.

Algorithm 1: Proposed method

```
Input:
   Set of 3D point cloud data, $P_{ToF}$ copied into $P_{data}$ and
$(x_i, y_i, z_i) \in P_{data}, i = 1, 2, \ldots, i_{max}$
Output:
   Set of labeled clusters of data
$(x_i, y_i, z_i, c_k) \in \{P_{labeled}\}, k = 1, 2, \ldots, k_{max}$
   Number Of detected clusters $k_{max}$
Initialization :
   (Constant) clustering thresholds: $\Delta x_{max}$, $\Delta y_{max}$,
$\Delta z_{max}$
   (Constant) Minimum number of points per cluster:
$i_{min}$
   Zero number of detected clusters : $k_{max} \leftarrow 0$
   No prior clusters: $P_{label} \leftarrow \emptyset$
   Set of temporarily assigned data points: $\{Q\} \leftarrow \emptyset$
   Current cluster number: $l \leftarrow 0$
while $P \neq \emptyset$ do
      $l \leftarrow l + 1$
      Append $(x_1, y_1, z_1)$ from $P^1_{data}$ to $Q$
      $P_{data} = P_{data} - \{(x_1, y_1, z_1)\}$
      $update \leftarrow 1$
    while $update \neq 0$ do
            $update \leftarrow 0$
        for $j \leftarrow 1$ to $Length(Q)$ do
            $(x_j, y_j, z_j) \leftarrow Q^j$
            for $i \leftarrow Length(P_{data})$ downto 1 do
                $(x_i, y_i, z_i) \leftarrow P^i_{data}$
                if $(|x_i - x_j| < \Delta x_{max})$ and
                  $(|y_i - y_j| < \Delta y_{max})$ and
                  $(|z_i - z_j| < \Delta z_{max})$ then
                    Append $(x_i, y_i, z_i)$ from $P^i_{data}$ to $Q$
                    $P_{data} = P_{data} - \{(x_i, y_i, z_i)\}$
                    $update \leftarrow 1$
                end
            end
        end
    end
    if $Length(Q) \geq i_{min}$ then
        $k_{max} \leftarrow k_{max} + 1$
        for $j \leftarrow 1$ to $Length(Q)$ do
            $P_{labeled} \leftarrow P_{labeled} \cup \{l, Q^j\}$
        end
    end
    Empty temporarily made set: $Q \leftarrow \emptyset$
end
```

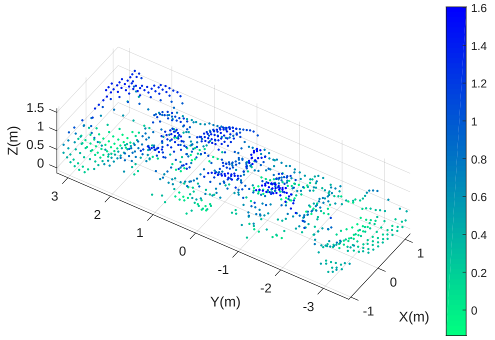


Figure 3: Raw sensor data in the camera coordinate, point cloud output of the time-of-flight camera. Depth (z) values of the points are represented by their color.

A. *Experimental Setup and Data Sets*

The data was collected at an experimental rock secondary breaking site, where a single ToF camera was installed at a fixed postion above the metal grid with rocks (see Figure 1). The IFM O3M150 ToF camera features on 50Hz, 64x16 resolution, generating up to 1024 point clouds. The raw data (shown in Figure 3) contains ten rocks on the metal grid in the scene. The blue points in the figure indicate the rocks and the metal grid frame while the green points indicate the ground. The data will be pre-processed by RANSAC and a denoising filter to remove outliers. The result of clean data after normalization is shown in Figure 4, where only ten clusters of points are left which illustrate 3D surface point clouds of ten rocks. As the experimental conditions were known, such as geometry between the ToF camera and the metal grid, size of each grid (60x60 cm) and number of rocks, a comparison can be done between benchmark methods and our proposed method.

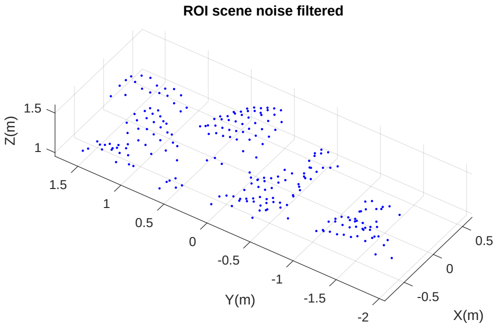


Figure 4: The point cloud after filtering of outliers.

B. *Comparison between benchmark methods and the proposed method*

Given a scene of ten rocks as a use case, the clustering results for Ward is shown in Figure 5 and DBSCAN is shown in Figure 6, respectively. For better visualization, all 3D plots from Figure 5 to Figure 7 are viewed in 2D X-Y coordinate, from -Z direction , which reflects the same view from the ToF camera.

Our results are evaluated according to hard clustering critieria, i.e., each data point must belong to a cluster completely.

Ward hierarchical clustering uses the parameter of *Maxclust* and *K* to construct a maximum of K clusters using the distance criterion. The result in Figure 7 shows that all clusters were found despite misclassified adjacent points between clusters 1 and 2, as well as 4 and 5. The algorithm behaves consistently with different data. We therefore used it to make comparisons with the proposed method in Table II.

The DBSCAN algorithm does not require an initial K value. Instead, adjustment of two additional parameters *epsilon* and *minPts* is needed, and one may not know these values in advance. As a result, the criteria for soft clustering appears satisfactory, while for hard clustering it is a bit problematic. As shown in Figure 6, some points that belong to clusters 1,5 and 6 are treated as outliers.

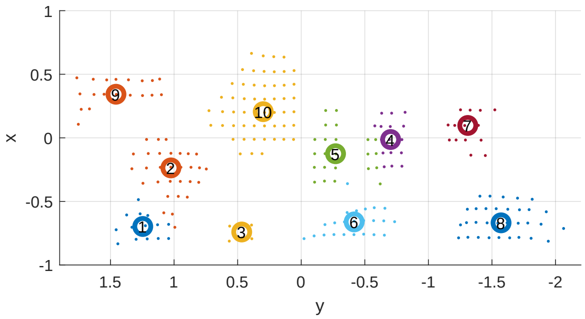


Figure 5: Clustering by Ward hierarchical clustering

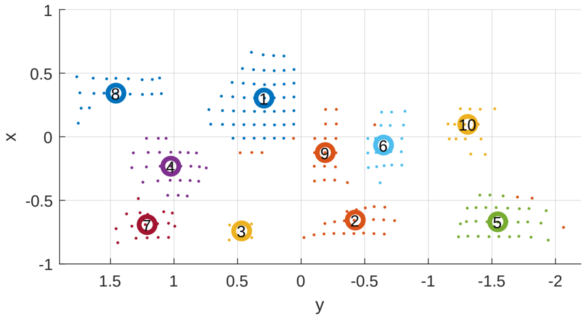


Figure 6: Clustering by DBSCAN clustering

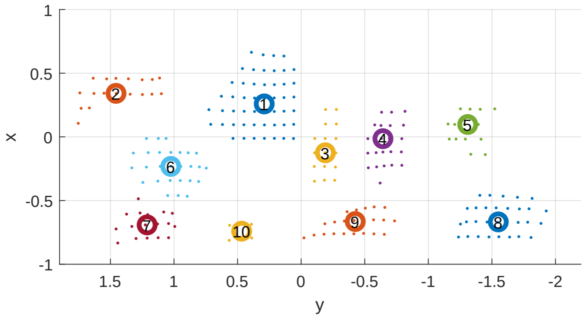


Figure 7: Clustering by the proposed method

We validate the proposed method through experiments with a variety of clusters of different sizes. The outcome of these experiments demonstrates the potential of the proposed method. As shown in Figure 7, all points are properly labelled without error.

C. *Comparison to ground truth*

The ground truth is that the numbers of clusters are known in advance and each cluster to which every point belongs to is also known. Concerning the accuracy of measurements, as improving sparse data derived from highly dense data requires camera hardware changes, we therefore must rely on camera data as the base for comparison of different methods.

TABLE II: Maximum Distance Error, DBSCAN, WARD vs Proposed

Clusters	*DBSCAN*	*Ward*	*Proposed*
4	0	0	0
5	0.0802	0	0.0367
6	0.2520	0.0442	0.0718
7	0.1104	0.0593	0.0373
8	0.1399	0.0153	0.0171
9	0.0538	0.0557	0.0211
10	0.0985	0.0779	0

Table II illustrates maximum euclidean distance errors from each method for each scenario involving a given number of rocks. The result is calculated from the ideal centroid postion of each rock in euclidean distance in a 2D X-Y coordinate space. Note that the diameter of each rock in X-Y coordinate space should be practically larger than 0.6m, i.e., the size of the grid. Otherwise, rocks may roll down from the grid.

The accuracy of the proposed method depends on noise. In certain cases (such as that shown in Figure 7), there is no noise and thus the error can be zero.

The error from the Ward method in Figure 5 derives mainly from mislabeled adjacent points. Therefore, the impact on accuracy is moderate. This occurs because for a small fraction of points the inaccuracy does not affect the overall distribution of distance severely. As we can see from Table II the maximum deviation of euclidean distance in X-Y coordinate space is approxmate 8 cm.

For DBSCAN, larger errors are observed because points belonging to the same cluster are scattered over a larger region. Compared to DBSCAN, the proposed method's labeled points are more centrally distributed; therefore, the errors are much smaller.

Subject to camera hardware limitation, when the distance between rocks is less than camera resolution, i.e., 0.12m, discontinity is unlikely to be identified, as two neighbhouring clusters are likely to be treated as one.

D. Post-processing

After clustering, we performed 3D surface reconstruction of rocks, and as shown in Figure 8, we calculated and visualized the hammer's contact point for each rock by giving its position and angle. These data were then sent to the manipulator.

V. Conclusion

By comparative study on major clustering methods, this paper proposes a systematic way for automatic rock recognition and target perception for autonomous mining jackhammer manipulators for secondary breaking. To fulfill the requirements of such an application, we represented segmentation of surface point clouds of rock masses as well as population estimation of rock pieces and their normal-to-surface vectors. Note that in previous studies, a lack of prior knowledge with regard to the quantity of rocks prevented the use of major contemporary clustering approaches.

The proposed method, together with several state-of-the-art methods, has been examined through experiments with different arrangements of objects. It is shown that the proposed method is capable of detection in a robust and accurate manner for the mining application. The method consists of several point-cloud processing steps, such as the proposed algorithm for estimation of cluster numbers, calculation of centroid position and normal vector of each cluster, pre-processing using RANSAC and denoising filters to remove outliers.

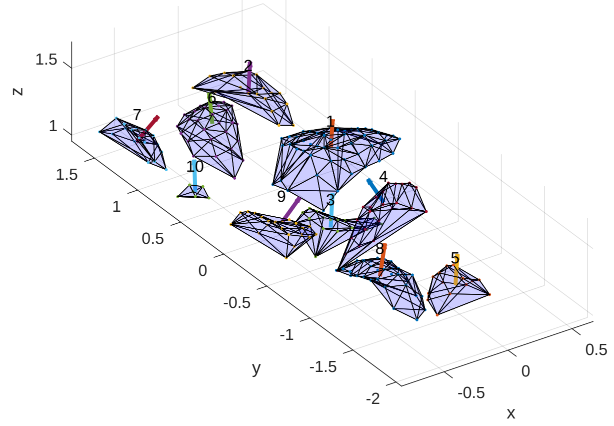


Figure 8: 3D surface reconstruction of rocks based on clustering results

The experimental studies represent a significant difference in the depth, as in any vision-based system. Therefore, this difference is required to be implemented in the clustering algorithm, where the noise and error variations have different behaviors along each principle axes. As the application involves only stationary ToF data, the coordinate has fixed axes. We recommend the use of normal distances in each principal coordinate axis of the camera frame. This, in contrast with point-to-point Euclidean distance, allows us to adjust sensitivity of the algorithms based on the systematic errors and noises of the vision system differently at each direction. In other words, it is not considered equal if some points have the same distances in depth compared to the other directions because the distances in depth are more likely to be affected by measurement noise.

According to experimental results, the proposed method has better robustness and overall performance compared to DBSCAN. Our method, in contrast with WARD, does not require manual adjustments based on the rock arrangements while preserving and sometimes improving performance of WARD. In conclusion, the proposed method improves performance and flexibility of the system while accounting for robustness.

Acknowledgment

The work leading to this publication has been funded in part by Fusion for Energy and TEKES under Grant F4E-GRT-0689. This publication reflects the views only of the authors, and Fusion for Energy or TEKES cannot be held responsible for any use which may be made of the information contained herein.

References

[1] N. Chen, J. Kemeny, Q. Jiang, and Z. Pan, "Automatic extraction of blocks from 3d point clouds of fractured rock," *Computers & Geosciences*, vol. 109, pp. 149–161, 2017.

[2] M. Vöge, M. J. Lato, and M. S. Diederichs, "Automated rockmass discontinuity mapping from 3-dimensional surface data," *Engineering Geology*, vol. 164, pp. 155–162, 2013.

[3] A. J. Riquelme, A. Abellán, R. Tomás, and M. Jaboyedoff, "A new approach for semi-automatic rock mass joints recognition from 3d point clouds," *Computers & Geosciences*, vol. 68, pp. 38–52, 2014.

[4] M. Tonini and A. Abellan, "Rockfall detection from terrestrial lidar point clouds: A clustering approach using r," *Journal of Spatial Information Science*, vol. 2014, no. 8, pp. 95–110, 2014.

[5] M. J. Lato and M. Vöge, "Automated mapping of rock discontinuities in 3d lidar and photogrammetry models," *International Journal of Rock Mechanics and Mining Sciences*, no. 54, pp. 150–158, 2012.

[6] G. Gigli and N. Casagli, "Semi-automatic extraction of rock mass structural data from high resolution lidar point clouds," *International Journal of Rock Mechanics and Mining Sciences*, vol. 48, no. 2, pp. 187–198, 2011.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] L. Niu, S. Smirnov, J. Mattila, A. Gotchev, and E. Ruiz, "Robust pose estimation with a stereoscopic camera in harsh environments," *Electronic Imaging*, vol. 2018, no. 9, pp. 1–6, 2018.

[9] S. Mattoccia and M. Poggi, "A passive rgbd sensor for accurate and real-time depth sensing self-contained into an fpga," in *Proceedings of the 9th International Conference on Distributed Smart Cameras*. ACM, 2015, pp. 146–151.

[10] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *Icml*, vol. 1, 2000, pp. 727–734.

[11] H. Bischof, A. Leonardis, and A. Selb, "Mdl principle for robust vector quantisation," *Pattern Analysis & Applications*, vol. 2, no. 1, pp. 59–72, 1999.

[12] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Advances in neural information processing systems*, 2004, pp. 281–288.

[13] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[14] T. Ma, Z. Wu, L. Feng, P. Luo, and X. Long, "Point cloud segmentation through spectral clustering," in *Information Science and Engineering (ICISE), 2010 2nd International Conference on*. IEEE, 2010, pp. 1–4.

[15] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[17] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[18] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[19] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927–941, 2008.

PUBLICATION

III

Efficient 3D Visual Perception for Robotic Rock Breaking

L. Niu, K. Chen, K. Jia and J. Mattila

2019 IEEE 15th International Conference on Automation Science and Engineering (CASE) 2019, 1124–1130

DOI: 10.1109/COASE.2019.8842859

Efficient 3D Visual Perception for Robotic Rock Breaking

Longchuan Niu[1], Ke Chen[2], Kui Jia[2], and Jouni Mattila[1]

Abstract— **In recent years, underground mining automation (e.g., the heavy-duty robots carrying rock breaker tools for secondary breaking) has drawn substantial interest. This breaking process is needed only when over-sized rocks threaten to jam the mine material flow. In the worst case, a pile of overlapped rocks can get stuck on top of a crusher's grate plate. For a human operator, it is relatively easy to make the decisions about the rock locations in the pile and the order of rocks to be crushed. In an autonomous operation, a robust and fast visual perception system is needed for executing robot motion commands. In this paper, we propose a pipeline for fast detection and pose estimation of individual rocks in cluttered scenes. We employ the state-of-art YOLOv3 as a 2D detector to perform 3D reconstruction from point cloud for detected rocks in 2D regions using our proposed novel method, and finally estimating the rock centroid positions and the normal-to-surface vectors based on the predicted point cloud. The detected centroids in the scene are ordered according to the depth of rock surface to the camera, which provides the breaking sequence of the rocks. During the system evaluation in the real rock breaking experiments, we have collected a new dataset with 4780 images having from 1 to 12 rocks on a grate plate. The proposed pipeline achieves 97.47% precision on overall detection with a real-time speed around 15Hz.**

I. INTRODUCTION

Underground mining continues to progress to deeper levels for tackling the mineral supply crisis in the 21st century [1]. Human worker safety in mines deeper than a kilometer, along with time-consuming human shift worker logistics, is a massive mine operational cost challenge. This has increased demand for the level of autonomous robotics in mining. In deep mines, the extracted material is fed to crushers equipped with grate plates for stopping over-sized rocks (i.e., ore) from falling into the crusher jaws. The grate plate (e.g., a mesh size of 0.5 m x 0.5 m) prevents crusher jamming, but only if over-sized rocks remaining on the plate are immediately broken down into smaller pieces to ensure continuous mine mineral flow. Such rock breaking has been conventionally done by a human operator-driven heavy-duty hydraulic four-link anthropomorphic arm equipped with a hydraulic hammer tool, as shown in Fig. 1.

Recently, robotic rock breaking [2] has attracted wider attention owing to the controllable breaking procedure. Sensory rock perception plays an important role in robotic rock breaking as it provides the automatic over-sized rock detection and the motion target coordinates for the robotic rock

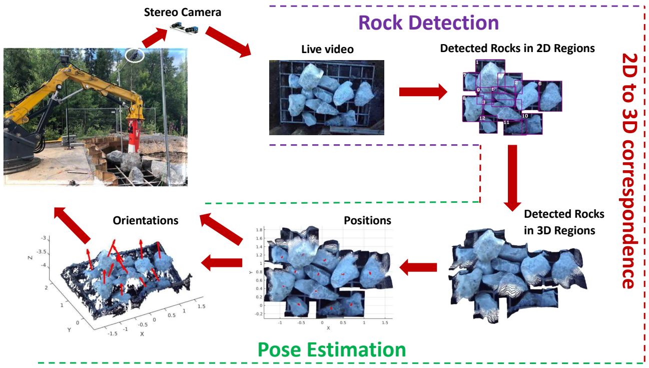


Fig. 1 *3D perception of rocks on a grate plate*

breaker arm. Some rock breaking systems with increased automation levels have been developed, such as the telerobotic rock breaker [3], vision-based mining automation controls [4], and 3D perception for mining robotics [5]. Some studies on rock breaking systems using force sensors [6] and stereo vision [7] have adopted algorithms for computing normals of rock surfaces. Nevertheless, none of the existing methods are capable of understanding the whole rock breaking scene in a complex environment.

For the automatic analysis of a scene, visual 3D perception requires fast and reliable initial detection with accurate object recognition and localization. However, this problem remains challenging due to piled rock scenes having arbitrary shapes, sizes, textures, and colours, as shown in Fig. 1. Pose estimation for objects with prior knowledge of shape was studied using 3D template matching technique in our earlier work [8]. For objects with unpredictable shapes, we have adopted a clustering algorithm for direct point cloud segmentation [9]. This method is used on secondary breaking in an unsupervised learning manner, but it suffers by missing texture-free visual cues for segmenting two rocks close to each other. A lack of contextual information in pure point analyses encourages us to conduct foreground highlighting in the RGB images to improve 3D rock detection. Moreover, such a setting has its significance in the practice of collision avoidance in robot on-line motions.

In this paper, we address the 3D visual perception of rocks via a pipeline visualized in Fig. 2, which consists of three stages: 2D rock detection, 2D-to-3D correspondence of regions, centroid position, and normal-to-surface vector estimation on object point cloud-based surfaces. At the first stage, the rocks displayed in the left image of Fig. 2 are detected as 2D regions (bounding boxes) by the state-of-

[1] Automation Technology and Mechanical Engineering, Faculty of Engineering and Natural Sciences, Tampere University, FIN-33720, Tampere, Finland {longchuan.niu, jouni.mattila}@tuni.fi

[2] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, P.R. China

the-art detector [10] (see Sec. IV-B). For the study, a stereo camera system is used to reconstruct the geometry of a 3D scene based on stereo correspondences. Subsequently, the depth map is generated in the form of a gray-scale image describing its geometry. We utilize this property to recover a *3D point cloud* representation of a textured point cloud of rocks in 2D regions (produced by the 2D detector) to its corresponding point clouds in 3D regions. These are all performed at the 2D-to-3D correspondence stage (see Sec. IV-C), where scene background can be removed and we focus on analyzing rocks in the foreground. At the last stage, based on the predicted point sets for each rock, the centroid of the surface is discovered and its corresponding normal-to-surface vector is estimated by searching for the best fit plane using the nearest points provided by random sample consensus (RANSAC) [11].

Contributions

The novel contributions of this paper are fourthfold. Firstly, we developed an efficient 3D visual perception pipeline for the detection of visible rocks and individual rock 6D pose estimations in cluttered scenes. We achieved an average precision of 97.47% at a real-time speed around 15Hz. Secondly, instead of a conventional stereo-image rectification method, we proposed a plane-sweeping depth estimation method for establishing the 2D to 3D correspondence. Thirdly, on non-Euclidean structured points, we designed a method for estimating the normal-to-surface vectors on detected rock surfaces. Finally, we collected and annotated 4780 different images for the rock detection in a real scale rock breaking robot set-up with the rocks weighing several hundreds kilos each. This dataset is the according to the authors' best knowledge of the first dataset of blasted overlapped rocks.

Experiment results on the new dataset verified the efficacy of the proposed method, which works even if a part of the object is occluded or truncated due to the presence of the robot arm or rocks in a pile. The dataset used for the training has been made available with this paper[1].

This paper is organized as follows: Section II introduces related research on object detection; Section III describes the research problem; Section IV details the methodologies used for the study; Section V explains the experiments that were carried out; and Section VI concludes the paper.

II. Related Work

Object detection is widely studied, and a number of methods based on deep learning has been proposed [12]–[17]. Most existing methods operate using 2D Euclidean convolution on images, which can be categorized into two main groups. The first group is object proposals and image classification, such as region-based convolutional neural networks (RCNN) [18], fast RCNN [19], and faster RCNN [15]. These methods begin by generating thousands of region proposals within the images, and then apply a convolutional classifier to filter the proposals by classification score thresholds. This two-stage setting increases networked training difficulties due to independent training on each individual component in the pipeline. The second group is single shot-based detection, such as SSD [14] and YOLO. Recently, the YOLO detector [10], [12], [13] has become a viable alternative to RCNN variants by achieving superior detection efficacy. Not many 2D-driven 3D object detection studies [20], [21] have been based on both RGB-D images and point clouds. Specifically, utilizing a mature 2D object detector's output to generate 3D object proposals, this reduces the search in entire 3D dense point cloud.

Currently, the majority of 3D object detection methods [22]–[24] operate light detection and ranging (LiDAR) generated point clouds for outdoor applications. Compared to RGB images, LiDAR point clouds are unordered and too sparse to distinguish the severe inter-occlusion between the rocks, which makes the direct application of these methods challenging in a rock-breaking scenario. In light of this, our method maps 2D pixels within predicted bounding boxes into rock point cloud surfaces, which generate a visible rock surface as 3D proposals.

Further state of the art segmentation using an instance segmentation method, such as Mask R-CNN [25], could be performed for each rock within the bounding box. In practice, this method can further boost segmentation performance with the price of higher computational costs, which can be less suitable in real-time applications, such as rock breaking. Our proposed method works effectively in the robotic rock breaking scenario, which is verified in Sec. V.

III. Problem Statement

As mentioned, automatic rock breaking requires fast and reliable detection and localization of every rock in a given scene. Oversized rocks on the grate plate can range from one rock or few rocks scattered around to many rocks in a complex pile overlapping each other. In our real-world robotic rock breaking set-up, we utilize a top-mounted stereo camera to provide video and images for automatic rock recognition and analysis. Given live video or still stereo images as input, the goal is to achieve real-time and sophisticated rock detection in a reference camera (left camera) coordinate, since individual 6D poses have to be shown to the operator and sent to the robot controller.

IV. Methodology

For obtaining required rock poses for the controller, the rock centroid positions $[x, y, z]$ and orientations (i.e. normal-to-surface vectors at their centroids), we conduct three phases in our visual perception system. The first phase is detection, where we employ a 2D object detector [10] for rock detection (see Sec. IV-B). The second phase is 2D to 3D correspondence, where 3D rock surfaces in a point cloud are generated via projection from 2D regions (see Sec. IV-C). In the final pose estimation phase, estimation methods for the centroid position and the normal-to-surface vectors are applied. Fig. 2 illustrates the whole pipeline of the proposed system.

[1] `https://doi.org/10.5281/zenodo.3246919`

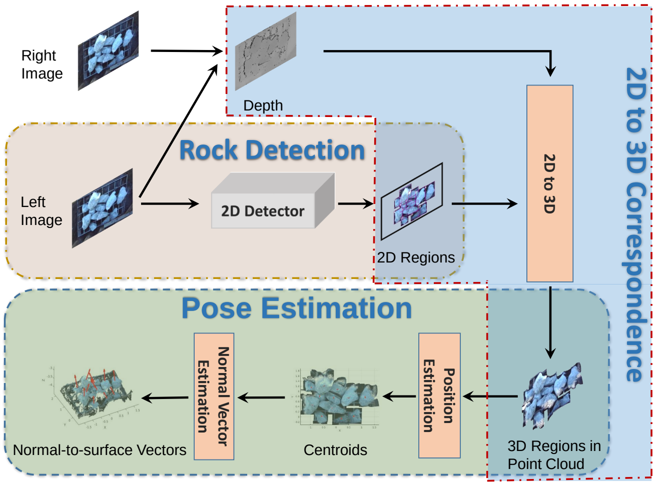


Fig. 2 *Pipeline of the proposed visual perception system*

A. A NEW DATASET FOR ROCK BREAKING

The procedure of data collection and annotation for the new dataset generation in the rock breaking application was organized as follows. The videos were recorded with various amounts of rocks on a grate plate under different outdoor illumination conditions by using a top-mounted stereo camera. However, due to the complex image gathering process in an outdoor environment, the position of camera was not entirely fixed. Therefore, slight camera movements during the video recordings can occur, which leads to background subtraction process failure. In view of this, object detection is considered the best possible approach to cope with the diverse background. In the gathered dataset, 4780 videos were recorded using a pre-calibrated stereo camera compressed in a lossless format in 720p at 15fps. They were further processed offline to extract a selected frame from each video into left and right images, which were used to generate depth maps and point clouds with color information (in ply files).

The Yolo Mark tool[2] was used for left image annotation. To alleviate manual labelling, an automatic labelling tool was implemented. This required manual labelling of 1,000 images among 4,780 previously extracted images, which were then used to train a coarse 2D detector to label the remaining 3,780 images. After automatic labelling, the labelled images were still checked one by one. The quality of automatic labelling is known to be highly dependent on the quality of previously labelled data as well as the coverage of the data set. Therefore, a random data selection mechanism was implemented for this purpose.

B. OBJECT DETECTION

As aforementioned, YOLO [10] was adopted for rock detection in 2D, making it an essential step for further processing. This kind of 2D detector formulates object detection into a regression problem, which addresses localization and recognition in a unified framework via simultaneous prediction of bounding box confidence and class probabilities. To this end, the whole image is divided into regular grids before the network predicts the object's centroids from the given set of candidates for various bounding boxes and object classes. Owing to its efficient detection, we are utilizing the latest network structure [10]. More specifically, the detection network (a variant of darknet-53) consists of 106 convolutional layers, where the prediction is performed at three different scales by predicting 10 times the numbers of boxes, producing more accurate results when detecting small objects.

C. 2D-3D CORRESPONDENCE

The estimation of scene geometry from a stereo camera setup is usually called a *depth-from-stereo* problem, the goal is to estimate the depth of each pixel in a RGB image. Conventional rectification-based stereo-matching methods [26] require excessive image interpolation steps and rigorous geometrically parallel camera configuration. Instead, we adopt the *plane-sweeping depth estimation* method, which allows direct processing of captured imagery [27] via calibrated camera parameters. This enables the setup of multiple cameras for the acquisition of point clouds from different angles in the future. Fig. 3 illustrates the plane-sweeping principle of the depth estimation method using a stereo camera.

The method assumes that the entire scene can be divided into a number of front-to-parallel planes where stereo correspondences could be found. The depth hypotheses can be selected according to the possible depth range ($z_{min} \leq z \leq z_{max}$) and a finite number of layers, to achieve a balance between fidelity and computational complexity.

Another advance of this method is its suitablity for parallel computing, and therefore, a dense 3D reconstruction of a complex scene can be realized in real time through GPU acceleration.

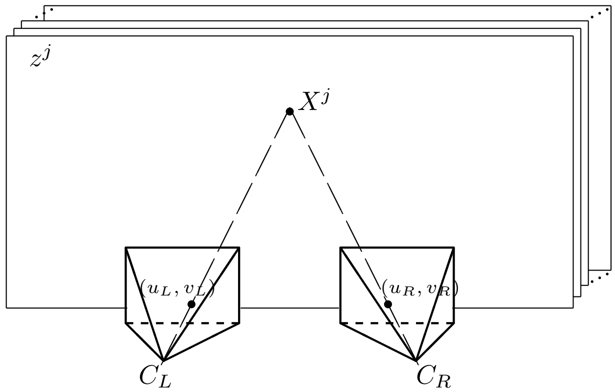


Fig. 3 *An illustration of the plane-sweeping principle of the depth-from-stereo estimation methods for a stereo camera*

For each hypothetical parallel plane with depth z^j, a pixel (u_L, v_L) from a left (reference) camera can be projected to a 3D space, using pre-calibrated camera matrix C_L:

[2] `https://github.com/AlexeyAB/Yolo_mark`

$$\mathbf{X}^j = C_L^{-1}(u_L \cdot z^j, v_L \cdot z^j, z^j, 1)^T = C_L^{-1}\dot{\mathbf{x}}_L, \qquad (1)$$

where $\dot{\mathbf{x}}_L = (u_L \cdot z^j, v_L \cdot z^j, z^j, 1)^T$ is the homogeneous projective coordinate of a current pixel, $\mathbf{X}^j$ is the resulting 3D point coordinate, and $j = 1,..,M$ where M is the selected number of layers.

Then, each projected 3D point $\mathbf{X}^j$ can be further projected onto the image plane of a second camera with a similar equation:

$$\dot{\mathbf{x}}_R = C_R\mathbf{X}^j = (u_R \cdot z^j, v_R \cdot z^j, z^j, 1)^T, \qquad (2)$$

where $\dot{\mathbf{x}}_R$ is a projective pixel in a second camera image plane, and the actual pixel coordinates can be recovered as:

$$u_R = \frac{\dot{\mathbf{x}}_R.x}{\dot{\mathbf{x}}_R.z}, \quad v_R = \frac{\dot{\mathbf{x}}_R.y}{\dot{\mathbf{x}}_R.z}. \qquad (3)$$

We can construct a 3D cost volume in which pixel dissimilarities are calculated between the original pixel in the reference camera and the corresponding pixel in the second one:

$$C(u,v,j) = \|I_L(u_L, v_L) - I_R(u_R, v_R)\|, \qquad (4)$$

where I_L and I_R denote the left (reference) and right camera images, respectively.

Through appropriate cost aggregation [27], the depth map can be recovered as such:

$$Z_L(u,v) = z^{\hat{j}}, \hat{j} = arg\min_j \tilde{C}(u,v,j), \qquad (5)$$

where $\tilde{C}(\cdot)$ denotes the aggregated cost volume.

The 3D coordinates of the point cloud, in accordance with the original pixel in the reference camera, can now be reconstructed using equation (1), replacing $z^{\hat{j}}$ with the estimated value.

D. POSE ESTIMATION

1) Position: The position of a rock is characterized in camera coordinates, indicating it is the geometric center of the bounding box in $x-y$ plane, as it is projected from image coordinates. This position estimation approach is sufficient, as those oversized rocks are with a dimension of at least 500 mm x 500 mm in $x-y$ plane, which allows some millimeter-level deviation.

2) Orientation (Normal-to-surface vectors): Given the location of the centroid of each rock, we estimate its normal vector for the best fitting plane of a nearby point cloud surface. For this goal, the principle of a RANSAC algorithm [11] searches for the best plane among a 3D point cloud surface.

A general plane equation is given as:

$$ax + by + cz + d = \mathbf{n}^T\hat{\mathbf{x}} = 0, \qquad (6)$$

where $\mathbf{n} = [a,b,c]^T$ is the normal vector of plane parameters to estimate and $\hat{\mathbf{x}} = [x,y,z,1]^T$ is the homogeneous point coordinate of the cloud.

The algorithm starts by randomly selecting three points from the cloud, fitting the plane parameters, and detecting all points of the point cloud that belong to the same plane by a given threshold. The process is repeated multiple times, until the plane equation containing the largest number of inliers is determined, the plane is considered as the best fitting plane.

As the point cloud estimated with the stereo-camera setup usually does not capture highly slanted or parallel-to-the-optical axis planes, inliers can be selected using a predefined threshold value θ, where points whose distance to plane is lower than a threshold meet the following condition:

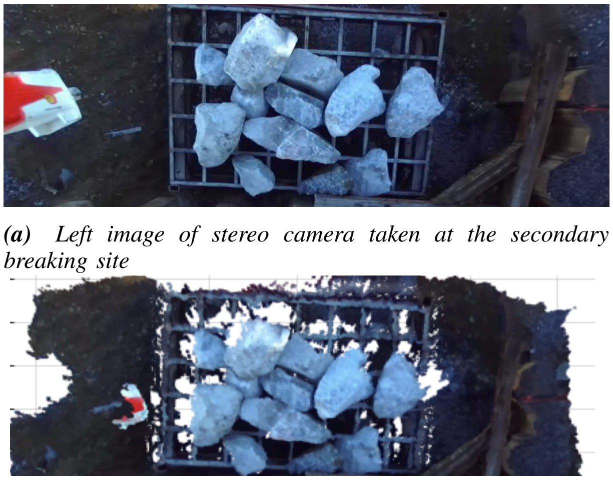
$$(x,y,z) \in Z^3 : 0 \le |ax_i + by_i - z_i + c| \le \theta. \qquad (7)$$

The threshold θ can also control the expected proximity of an object surface to a plane model. For object surfaces containing many bumps or cavities, larger values of θ can be beneficial.

V. EXPERIMENTS

A. Settings

The whole data for rock detection was split into training, validation, and testing sets for fair comparison. Specifically, 70% of the images (in 1280 x 720 resolution) were selected for training, 20% for validation, and the remaining 10% for testing. During parameter tuning, we used training data to fit network parameters by evaluating the performance on the validation set.



(a) *Left image of stereo camera taken at the secondary breaking site*

(b) *An example of point cloud generated from left and depth image*

Fig. 4 *Input images for visual perception*

B. Implementation Details

We set our visual perception system on Ubuntu with the following environment settings:

- OpenCV 3.4.0
- PCL 1.7.1
- CUDA 10.0
- CuDNN 7.4.2
- NVIDIA GeForce GTX 1060 6GB

We implemented all schemes in C++ with OpenCV library and Point Cloud Library (PCL).

From each video frame, we extracted a left image (an example is shown in Fig. 4a) together with a right image, computing its depth map (by means of the proposed plane sweeping method) to generate a point cloud (an example is shown in Fig. 4b). In parallel, the left images with labelled bounding boxes were provided to train the rock detector.

C. Evaluation of Rock Detection

We adopted the off-the-shelf YOLO detector using a variant of darknet-53 [28] in view of its solid detection performance as well as its efficiency during inference. We trained the darknet using our data by setting a learning rate of 0.001, which converges at an average loss of 0.12. We achieved good detection results during testing. Fig. 5 depicts the precision-recall curve, where IoU threshold is 0.75, true positive (TP) is 7144, false positive (FP) is only 125, and false negative (FN) is 141.

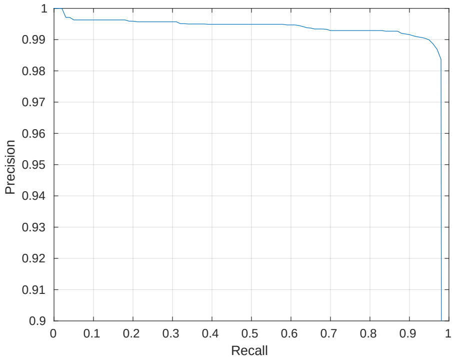
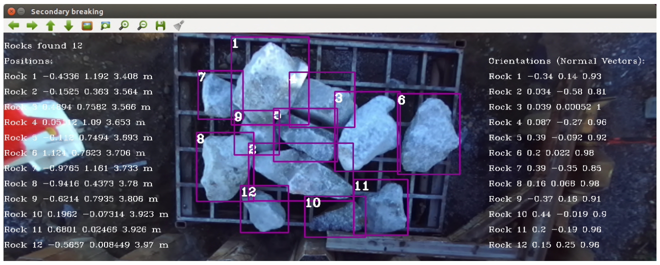


Fig. 5 Precision-recall curve of our proposed method

Moreover, an average precision of 97.47% was reached at ~70 ms per image. We validated the stability of the model using images at different scales and rotations to retain result robustness. Fig. 6 illustrates the detection result from an offline video, and the detected objects provided by YOLOv3 are highlighted with 2D bounding box.



```
Rocks found 12
Positions:
Rock 1 −0.4336 1.192 3.408 m
Rock 2 −0.1525 0.363 3.564 m
Rock 3 0.4894 0.7582 3.566 m
Rock 4 0.05112 1.09 3.653 m
Rock 5 −0.112 0.7494 3.693 m
Rock 6 1.124 0.7623 3.706 m
Rock 7 −0.9765 1.161 3.733 m
Rock 8 −0.9416 0.4373 3.78 m
Rock 9 −0.6214 0.7935 3.806 m
Rock 10 0.1962 −0.07314 3.923 m
Rock 11 0.6801 0.02466 3.926 m
Rock 12 −0.5657 0.008449 3.97 m
```

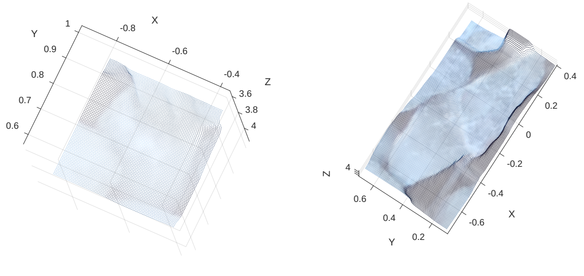
```
Orientations (Normal Vectors):
Rock 1 −0.34 0.14 0.93
Rock 2 0.034 −0.58 0.81
Rock 3 0.039 0.00052 1
Rock 4 0.087 −0.27 0.96
Rock 5 0.39 −0.092 0.92
Rock 6 0.2 0.022 0.98
Rock 7 0.39 −0.35 0.85
Rock 8 0.16 0.068 0.98
Rock 9 −0.37 0.16 0.91
Rock 10 0.44 −0.019 0.9
Rock 11 0.2 −0.19 0.96
Rock 12 0.15 0.25 0.96
```

Fig. 6 Detection and localization of rocks at approx. 15Hz

In addition, this single shot-based rock detector can efficiently localize all rocks at a video frame rate around 15 Hz. In Fig. 6, it can be seen that the rock 2 has a sharp edge in the middle, which is hard to segment properly using unsupervised learning methods [9], while rock 9 is occluded and truncated by rocks 1 and 8, which is harder to recognize using the aforementioned method.

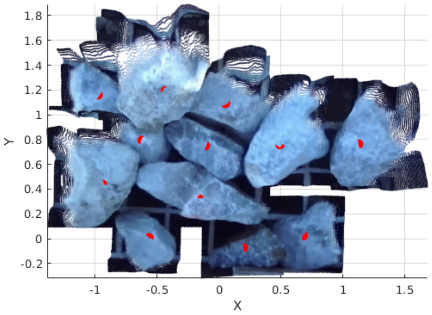
D. Evaluation on Pose Estimation

Here we conduct experiments to evaluate the results of estimating the position and orientation of individual rocks within 3D regions. For each detected 2D region, every pixel within has its 3D corresponding point in 3D point cloud with X,Y,Z and RGBA color. After 2D to 3D correspondence mapping, we obtained their 3D regions in a point cloud. Fig. 7 indicates rocks 2 and 9 in point clouds, through which 6D pose estimation can be performed.

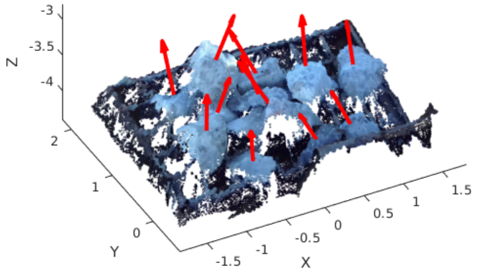


(a) Rock 9 *(b) Rock 2*

Fig. 7 Examples of the point cloud for rock 2 and rock 9 segmented by the projected 3D bounding boxes



(a) Detected centroid positions in 3D



(b) Estimated normal-to-surface vectors in 3D

Fig. 8 Estimation of centroids' positions and normal-to-surface vectors for each 3D region

Fig. 8 illustrates detected 3D regions overall, along with estimated centroids and normal-to-surface vectors for each region. Estimated centroids for each rock are drawn as red spots (as shown in Fig. 8a where they geometrically reside at the center of each rock, even for all occluded rocks).

The estimation of the normal-to-surface vectors was performed using a k-d tree to search for the neighbors (1000 points) around each centroid point. It took the RANSAC method less than four iterations to find the best fitting plane. As no ground-truth normal vectors were available, we visualized the normal-to-surface vectors together with the rocks for quality evaluation. Fig. 8b presents the result of estimating the normal-to-surface vectors shown with red arrows for each rock. As a result, those normal vectors were perpendicular to the estimated main surface plane of each rock. More qualitative results are shown in Fig. 9.

VI. CONCLUSIONS

We have proposed a novel fast method for 3D object detection and target pose estimation for complex scenes containing irregularly shaped and sized blasted rocks that can be in an overlapping pile. Even though object detection using bounding boxes has been widely studied, its extension to 3D in such complicated scenes remains a challenge, especially in a real outdoors environment. On one hand, in real-world outdoor applications, the 3D bounding boxes detector with LiDARs is not an efficient method for solving complex scenes with many sharp changes in the depth and overlying edges that are only visible on the images. On the other hand, 3D detection methods operating solely on dense point clouds can be computationally expensive, rendering the required real-time operation hardly feasible. This paper has presented an efficient online method by taking advantage of fast 2D object detection combined with the 2D to 3D plane-sweeping stereo matching method for 3D object detection. Given secondary rock breaking as an application, the proposed robotic visual perception method can meet the requirements for autonomous breaking required for the mining industry with its reliable object detection, real-time performance, and substantial accuracy on object pose estimation. The experiment results veried the efficiency of the proposed method with 97.47% detection accuracy at 15Hz in real outdoors worksite conditions. Our next research objective is to experimentally verify the success rate of real rock breaking with the machine vision estimated rock surface position as "a sweet spot" for the productive robotized operation.

REFERENCES

[1] P. G. Ranjith, J. Zhao, M. Ju, R. V. De Silva, T. D. Rathnaweera, and A. K. Bandara, "Opportunities and challenges in deep mining: A brief review," *Engineering*, vol. 3, no. 4, pp. 546–551, 2017.

[2] J. J. Green and D. Vogt, "Robot miner for low grade narrow tabular ore bodies: the potential and the challenge," 2009.

[3] E. Duff, C. Caris, A. Bonchis, K. Taylor, C. Gunn, and M. Adcock, "The development of a telerobotic rock breaker," in *Field and Service Robotics*. Springer, 2010, pp. 411–420.

[4] 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.

[5] ——, "3d perception for mining robotics," in *Field and Service Robotics*. Springer, 1998, pp. 46–52.

[6] H. Takahashi and T. Monden, "Automatic breaking system of large rocks by use of force sensors," in *INTERNATIONAL SYMPOSIUM ON ROBOTICS*, vol. 30. Citeseer, 1999, pp. 705–710.

[9] L. Niu, M. M. Aref, and J. Mattila, "Clustering analysis for secondary breaking using a low-cost time-of-flight camera," in *2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2018, pp. 318–324.

[7] A. Iamrurksiri, T. Tsubouchi, and S. Sarata, "Rock recognition using stereo vision for large rock breaking operation," in *Field and Service Robotics*. Springer, 2014, pp. 383–397.

[8] L. Niu, S. Smirnov, J. Mattila, A. Gotchev, and E. Ruiz, "Robust pose estimation with a stereoscopic camera in harsh environments," *Electronic Imaging*, vol. 2018, no. 9, pp. 1–6, 2018.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

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

[13] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

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

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

[16] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European conference on computer vision*. Springer, 2016, pp. 354–370.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[19] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[20] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4622–4630.

[21] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[22] E. Al Hakim, "3d yolo: End-to-end 3d object detection using point clouds," 2018.

[23] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1355–1361.

[24] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection." in *Robotics: Science and Systems*, vol. 1, no. 3, 2015, pp. 10–15 607.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[26] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[27] S. Smirnov, A. Gotchev, and M. Georgiev, "Comparison of cost aggregation techniques for free-viewpoint image interpolation based on plane sweeping," in *Ninth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM*, 2015.

[28] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013–2016.

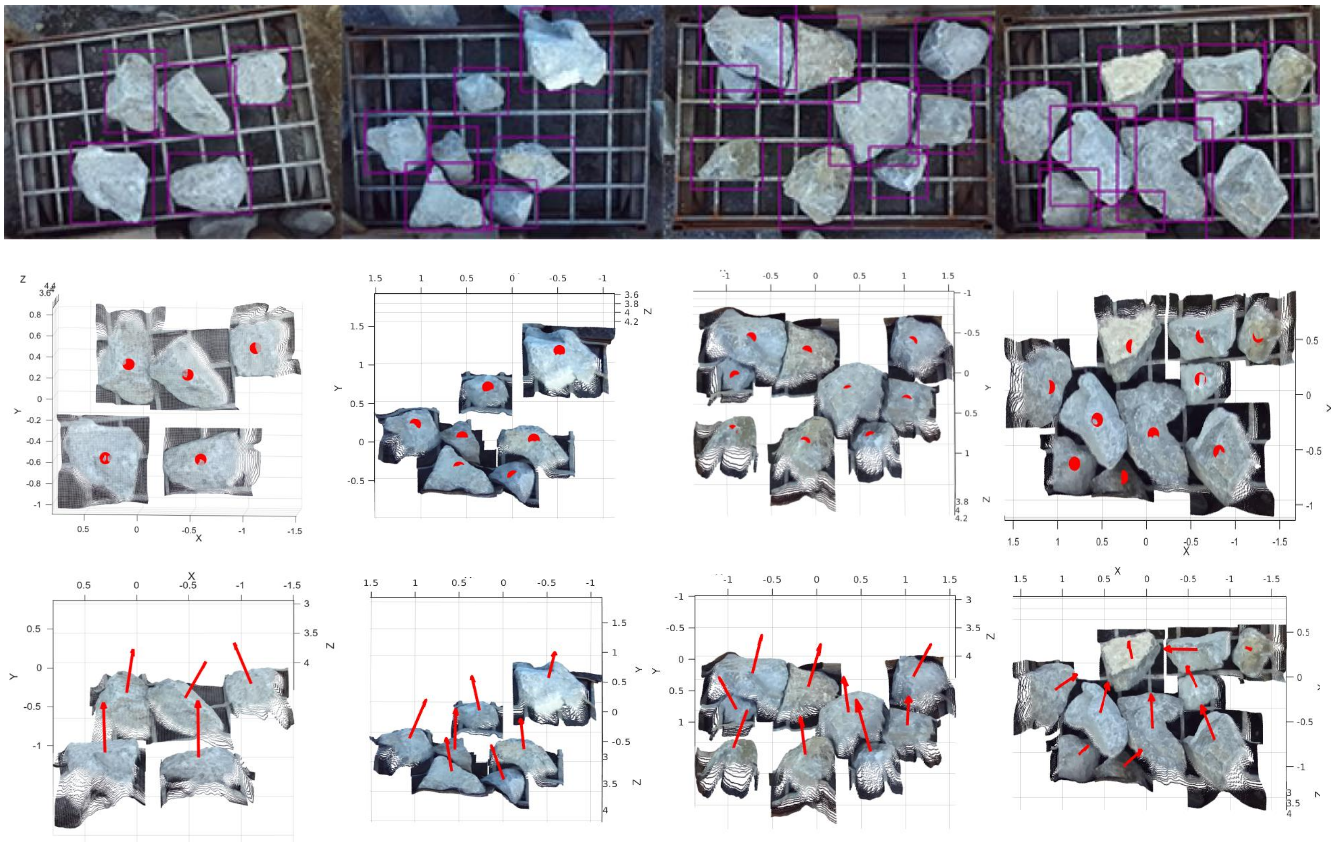


Fig. 9 *More visualization results of detection (top), position (middle), and the normal-to-surface vector estimation (bottom)*

UNPUBLISHED MANUSCRIPT

IV

Autonomous Robotic Rock Breaking Using a Real-time 3D Perception System

L. Niu, S. Lampinen, L. Hulttinen, J. Niemi and J. Mattila

2020

PUBLICATION

V

A stereoscopic eye-in-hand vision system for remote handling in ITER

L. Niu, L. Aha, J. Mattila, A. Gotchev and E. Ruiz

Fusion Engineering and Design 146.(2019), 1790–1795

DOI: `10.1016/j.fusengdes.2019.03.036`

Fusion Engineering and Design 146 (2019) 1790–1795

Contents lists available at ScienceDirect

Fusion Engineering and Design

journal homepage: www.elsevier.com/locate/fusengdes

A stereoscopic eye-in-hand vision system for remote handling in ITER

Longchuan Niu[a,*], Liisa Aha[a], Jouni Mattila[a], Atanas Gotchev[a], Emilio Ruiz[b]

[a] *Tampere University, Tampere, Finland*
[b] *Fusion for Energy, Barcelona, Spain*

ARTICLE INFO

Keywords:
Stereoscopic
Eye-in-Hand
Remote handling
Computer aided teleoperation

ABSTRACT

The International Thermonuclear Experimental Reactor (ITER) maintenance is performed by means of remote handling (RH) systems and with aid of user interfaces such as haptic and joystick devices, virtual reality (VR) systems, and camera views. Many RH operations involving RH equipment, such as robotic manipulator arms, require millimeter accuracy, but camera views are often occluded or of poor quality, and might be unavailable during sensitive steps that require accurate, close-up views. Moreover, the VR system may not reflect the current scene accurately, as physical conditions may have changed under the harsh environment. The purpose of this research was to prototype and evaluate a novel software system, called 3D Node, that locates and detects the position and orientation of a piece of RH equipment or reactor element with respect to a stereo camera pair. The detection information is utilized to adjust the motion trajectories of a robotic manipulator arm. The 3D Node features stereo-camera calibration, target depth mapping, target position and orientation detection, and online target tracking. This paper reports on the 3D Node demonstration on the ITER Divertor RH use case and discusses the system applicability to other ITER RH systems.

1. Introduction

Performing accurate ITER RH maintenance operations inside dark and highly radioactive chambers, where human access is impossible, is extremely demanding. The RH operator can utilize a number of user interfaces for commanding and controlling the RH equipment [1], e.g. a robotic manipulator arm. Other auxiliary interfaces involve live images of the RH equipment and its environment, computer-aided teleoperation (CAT) used in master-slave teleoperation [2], and virtual reality (VR) representing the movements of the RH equipment and its environment [3].

VR displays visual information based on the measured pose of the manipulators and pre-constructed virtual models. Due to the harsh environment, the VR representation may not exactly reflect the actual scene, as physical conditions may have changed, e.g. through material deformation due to extreme heat, or small drifts in the poses of the components to be manipulated. Thus, the pre-defined motion trajectories of the RH equipment have to be adjusted by other interfaces, e.g. a robot perception unit.

The purpose of the study herein is to prototype and demonstrate new means to assist RH operators to successfully perform ITER RH operations. A robot perception unit, namely 3D Node, was designed and developed to introduce new operator assisting features. The new features are based on detection of a target, i.e. a piece of RH equipment or reactor element, and recognition of its position and orientation in a relation to the environment using stereo camera images.

During ITER RH operations, a number of RH operations are identified in which 3D Node information could be helpful. This information could be valuable for updating VR models and implementing augmented reality and synthetic viewing functionalities. However, in this paper we consider 3D Node's usage merely in adjustment of the motion trajectories of RH equipment. A subset of operations related to the Divertor Cassette Locking System (CLS) operations is considered, and the use of 3D Node therein is demonstrated.

2. System architecture

As stated, the RH operator utilizes multiple software systems and user interfaces during an RH operation. As seen in Fig. 1, the novel software system, 3D Node, requires an interface to a manipulator control system and a stereo-camera pair that is attached to, for example, the manipulator arm. The 3D Node receives images from the stereoscopic cameras through GigE Vision protocol and the pose of the manipulator robot's tool center point (TCP) from the manipulator control system. Additionally, it can receive operator input and provide visual feedback to the operator through its graphical user interface (GUI). 3D

[*] Corresponding author.
E-mail address: longchuan.niu@tut.fi (L. Niu).

https://doi.org/10.1016/j.fusengdes.2019.03.036
Received 7 October 2018; Received in revised form 20 February 2019; Accepted 6 March 2019
Available online 22 March 2019

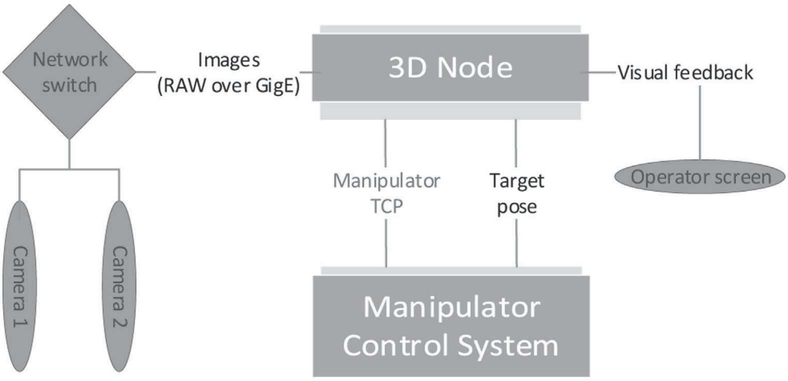


Fig. 1. Top-level architecture.

Node provides the target pose, i.e. position and orientation of the viewed target, to the manipulator control system.

The 3D Node is designed to comply with ITER remote handling control system (RHCS) requirements and is integrated at a later stage of its development into the RHCS as one of its components. Moreover, interfaces to the other systems, such as providing calibration information to VR, can be developed.

3. 3D node software

3.1. Operation modes

3D Node features five modes: calibration mode, video mode, depth mode, detection mode, and tracking mode. The 3D Node GUI has four main parts: the operator control panel and three views. Items displayed in the views depend on the selected mode (two examples in Fig. 2). The purpose and functionality of each mode is elaborated in the following.

Calibration mode is required to calibrate the stereo cameras and the relative position of the cameras with respect to the robot manipulator TCP, i.e. hand-eye calibration, prior to the actual RH operations. In ITER, this would be performed in the Hot Cell facility. In calibration mode, 3D Node captures stereo images of the scene while recording a current manipulator TCP. Images should contain a calibration pattern from diverse locations and angles. 3D Node performs the camera calibration with the aid of stereo images. The hand-eye calibration is also performed based on the stereo images and poses of the manipulator TCP.

Video mode is utilized for inspecting the camera views to confirm that the target object is not occluded by other objects, that lighting is sufficient, and that nothing prevents target detection during the RH operations. In video mode, 3D Node shows the images from both cameras.

Depth mode is used for checking the geometry of the scene or validating the correctness of stereo camera calibration. In depth mode, 3D Node visualizes a depth map of the scene that it has generated.

Detection mode is for detecting the target object and estimating its real pose. In detection mode, 3D Node determines the pose of the target object and aligns a rendered image of the target in the estimated position with the real camera view of the target. If the alignment is correct, the real image and rendered image should correspond to each other as seen in View 1 of Fig. 2b. In addition, 3D Node displays the desired pose of the manipulator TCP in the selected RH operation. The pose values are Cartesian positions in millimeters for X, Y and Z, and in degrees for orientation in Euler angles A, E and R. The pose is updated on RH operator command through the 3D Node control panel (Fig. 2). These values can be utilized for adjusting the motion trajectory of a manipulator arm.

Tracking mode is utilized when the manipulator is moved around to inspect the environment. It differs from video mode, as in tracking mode 3D Node illustrates the rendered image of the target on the camera view.

3.2. Method

3.2.1. Depth from stereo

Estimation of 3D scene geometry from parallel calibrated cameras is known as depth from stereo. In hazardous ITER environments, robust estimation of depth values is crucial. Instead of conventional rectification based methods, we employ the *plane-sweeping depth estimation* method, which uses calibrated camera parameters [4], allowing the captured imagery to be processed directly.

3.2.2. Advanced sampling

As illustrated in [5], an important step for pose estimation is the fine alignment between the sensed target point cloud and the reference point cloud.. This is done by utilizing a state-of-the-art edge point iterative closest point (ICP) algorithm.

Conventional edge-point ICP samples its model point cloud only once. In order to improve the robustness and accuracy of the fine alignment, we used a left-to-right correspondence check and dynamic CAD model resampling as a mechanism to reduce outliers in the model point clouds [4].

An example of a target object is given in Fig. 3a. We render the image to find strong edges, then prepare a point cloud using the left-to-right correspondence check. The sensed point cloud after sampling is shown in Fig. 3b.

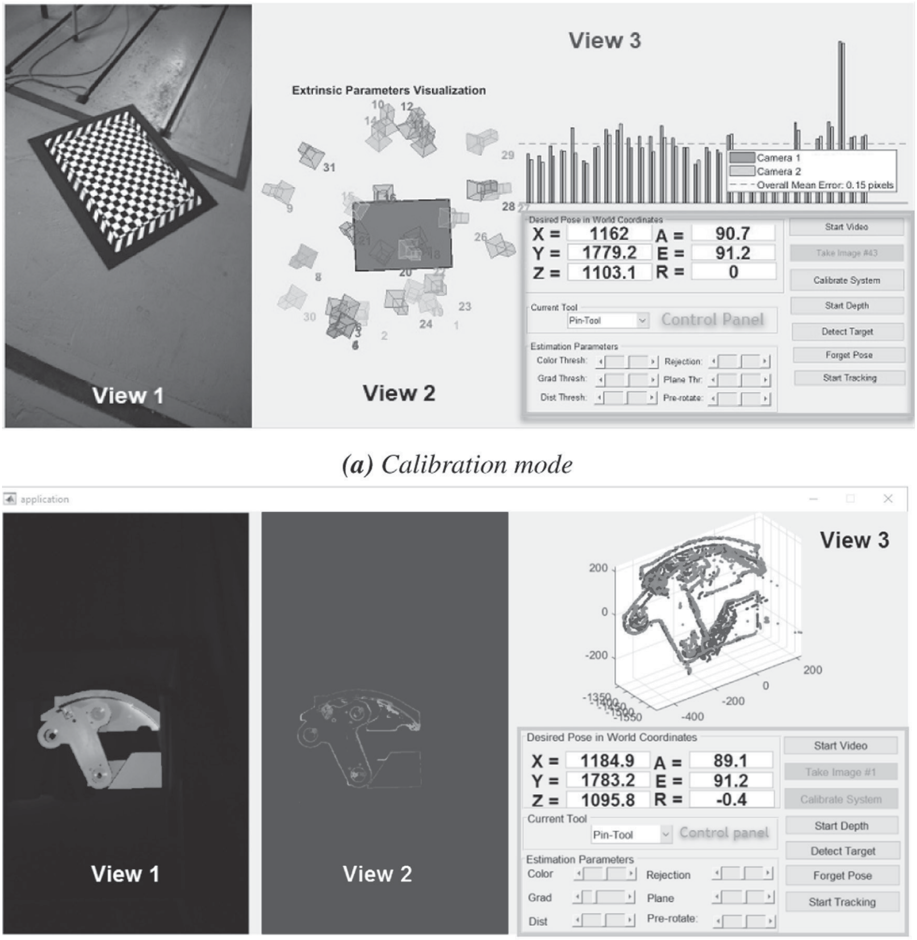
3.2.3. Pose estimation

The 3D Node estimates the target pose based on the stereo camera images and camera poses in the manipulator base frame. The camera pose is calculated by rigid body transformation between the cameras and the manipulator TCP, which is known as hand-eye calibration. The details are presented in [4], [5]. We adopt Tsai's method [6] for the hand-eye calibration.

4. Proof-of-concept demonstration

4.1. Demonstration equipment and setup

As indicated in Fig. 4, in the demonstration setup we used the Comau Smart NM45-2.0 robot as the manipulator with two cameras attached to its wrist. The target object utilized in the demonstration was



Desired Pose in World Coordinates
X = 1162 A = 90.7
Y = 1779.2 E = 91.2
Z = 1103.1 R = 0

Current Tool: Pin-Tool

Estimation Parameters: Color Thresh, Grad Thresh, Dist Thresh, Rejection, Plane Thr, Pre-rotate

(a) Calibration mode

Desired Pose in World Coordinates
X = 1184.9 A = 89.1
Y = 1783.2 E = 91.2
Z = 1095.8 R = -0.4

Current Tool: Pin-Tool

Estimation Parameters: Color, Grad, Dist, Rejection, Plane, Pre-rotate

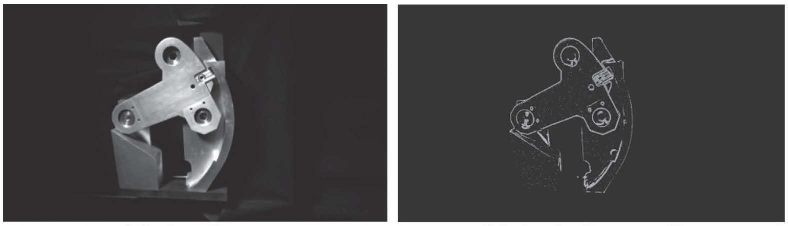
(b) Detection mode

Fig. 2. Sample GUI views: calibration and detection modes.

a test mock-up, which is a 1:1 replica of the Divertor Cassette Locking System (CLS).

The stereo cameras are mounted on an adjustable mounting plate, allowing reconfiguration for particular environments. The accuracy of camera positioning is not an issue as long as the camera field of view is clear. The camera calibration process recovers the underlying stereo camera position every time the camera configuration changes or, for example, when a collision occurs.

The stereo cameras are arranged vertically. This is due to the dimensions between the tool exchanger and the Comau robot wrist. At ITER, the cameras could also be positioned horizontally depending on the manipulator. This is not an issue as the developed 3D Node system



(a) given image *(b) depth after sampling*

Fig. 3. Sampling of CAD model.

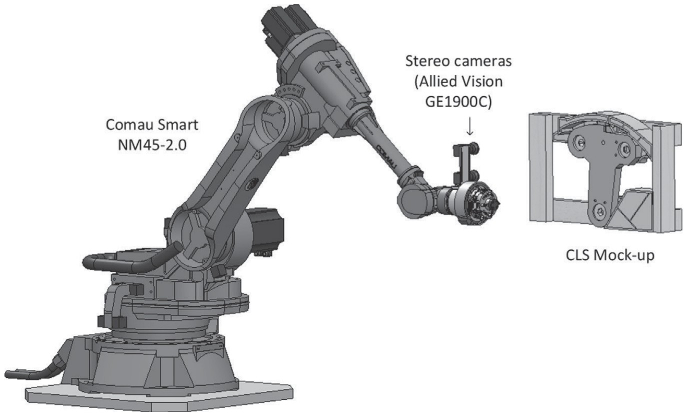


Fig. 4. Comau Smart NM45-2.0 with stereoscopic camera and CLS Mockup.

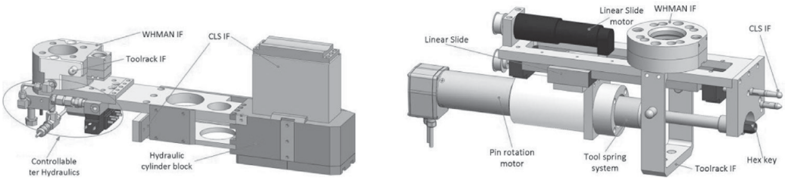
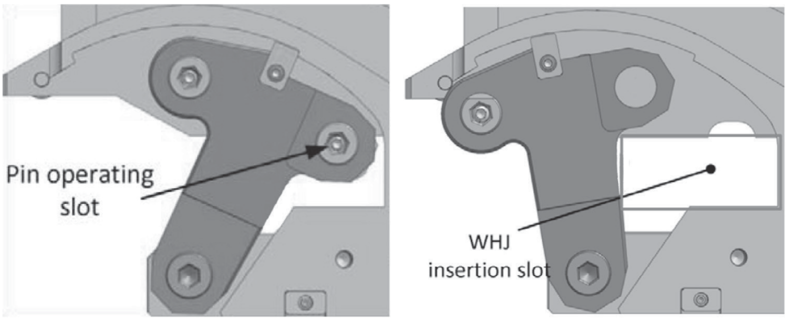


Fig. 5. Cassette locking system tools: jack tool and pin tool.



(a) Location of pin operating slot

(b) Jack tool insertion slot

Fig. 6. Locking mechanism, tool operation location.

works similarly regardless of the camera arrangement.

The current setting is optimized for depth sensing at a range of 400–1500 mm. This is mostly defined by adjusting the camera lenses to deliver optimal sharpness at these distances. For the distance between the cameras, a stereoscopic baseline of 100 mm was chosen as a practical compromise.

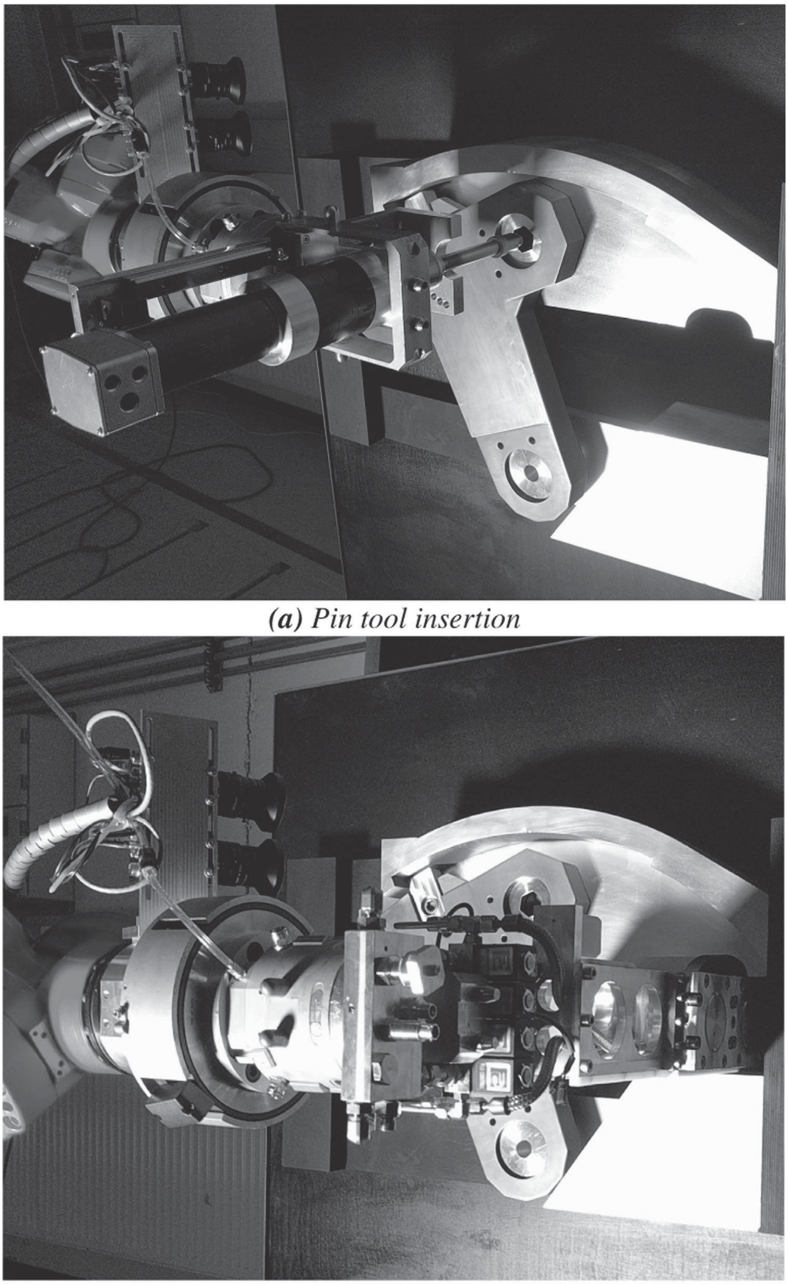
A pair of industrial machine vision digital cameras (Allied Vision GE1900C) was used for the demonstration. The native resolution of the camera is 1920 × 1080, and the sensor size is 1" with an effective pixel size of 7.4 μm. However, such cameras are not usable inside the actual ITER environment due to high levels of radiation; radiation tolerant cameras typically have a lower spatial resolution. We use the "Pixel Binning" feature of the cameras in order to decimate the original resolution as well to automatically convert images to a grayscale format. The resulting effective resolution of 960 × 540 is close to that of the standard radiation tolerant camera.

The Comau Smart NM45-2.0 robot payload capacity is 45 kg. In the 3D Node system demonstration, it operates the Divertor RH equipment tool prototypes, i.e. pin tool and jack tool (Fig. 5). Tool weights are 16 kg for the pin tool and 33.5 kg for the jack tool.

The Comau control system communicates with the 3D Node through User Datagram Protocol (UDP) at 1 Hz. The Comau control system sends the pose values of the manipulator TCP to the 3D Node. Received pose values from the 3D Node will be used to guide the manipulator arm in RH operations. At this moment, the target pose is only displayed within the 3D Node GUI and not sent directly to the Comau control system. Later, the 3D Node communication interfaces will be implemented to comply with ITER RH network communication protocols.

4.2. Demonstration cases

The pin tool and the jack tool are utilized in the CLS operations for unlocking and locking the cassette, and cassette compression, respectively. Fig. 6 indicates the location for these operations. Inserting tools



***(a)** Pin tool insertion*

***(b)** Jack tool insertion*

Fig. 7. Tool operations in the CLS mock-up scene.

into their corresponding slots requires millimeter accuracy in order to guarantee that the operations are performed properly. For example, the horizontal clearance between the jack tool and the slot in the cassette knuckle shown in Fig. 6b is approximately 7 mm. Therefore, we selected these two use cases to validate the functionality of the 3D Node and to give a proof-of-concept demonstration.

The purpose of the demonstration was to determine whether the 3D Node can help the RH operator to execute the RH operations in a physically unknown environment. There are two use cases, one is to insert the pin tool into the operating slot as illustrated in Fig. 6a, and the other is to insert the jack into the insertion slot as shown in Fig. 6b. Just before the demonstration, the target, i.e. the CLS mock-up, was randomly placed, which emulates unpredicted target movement during ITER maintenance operation. The operations were performed in a dark laboratory room with only a single adjustable light source pointed to the target. Prior to any operation, camera and hand-eye calibration were performed.

In both cases the operation sequence is the same when using the 3D Node. At first, the video mode can be utilized for inspecting the scene and the depth mode for validating the correct camera and hand-eye calibration. The detection mode is then utilized to find the actual pose of the target. According to the detected target pose values and calculated pose of the operated tool, the RH operator drives the Comau manipulator and the operated tool towards the calculated pose. As the tool tip reaches the desired pose, the operator can finalize the tool insertion in a peg-in-hole manner by simply driving along the Z-axis (depth) in the manipulator tool frame. The demonstration results are shown in Fig. 7. The successful operation from the insertion of both tools validates that the pose values given by the 3D Node are accurate.

5. Discussion

In order to assess the accuracy of the pose estimation algorithm, we performed a series of experiments [4]. For the observation range between 600 and 1200 mm, the relative accuracy from the repeatability test, i.e. deviation from re-measurement of the same position, is approximately 0.5–1 mm with respect to the position and 0.2-0.4 degrees with respect to the angle, providing that the target object has a planar surface.

Other sources of errors on the accuracy come from camera calibration, hand-eye calibration, and robot calibration. The stereoscopic camera calibration shows excellent stability, and the pixel reprojection error is about 0.15 pixels. Should there be higher resolution radiation tolerant cameras in the future, this would naturally improve the results of the camera calibration. The major proportion of hand-eye calibration error comes from the absolute accuracy of the robot, payload, and path of movement, i.e. possible backlash. Therefore, the selection and calibration of the manipulator are very important to ensure precise end-to-end movement.

6. Conclusions

3D Node is designed to fulfil generic ITER vision system requirements and can be easily integrated to any RHCS. The state-of-the-art pose estimation algorithm is developed to ensure good accuracy and robustness that can be achieved under dark and harsh conditions and with fairly low resolution cameras. The demanding test cases demonstrated its applicability in RH operations. The overall accuracy of the current system is highly dependent on the precision of the manipulator. It can be improved by robot calibration and fine tuning of the hand-eye calibration.

Acknowledgement

The work leading to this publication has been funded in part by Fusion for Energy and TEKES under Grant F4E-GRT-0689. This publication reflects the views only of the authors, and Fusion for Energy or TEKES cannot be held responsible for any use which may be made of the information contained herein.

References

[1] J. Tuominen, A. Muhammad, J. Mattila, L. Aha, H. Saarinen, M. Siuko, D. Hamilton, L. Semeraro, Command and control application framework for interoperable heterogeneous ITER remote handling devices, Fus. Eng. Des. 86 (9-11) (2011) 2067–2070.

[2] M. Viinikainen, J. Tuominen, P. Alho, J. Mattila, Improving the performance of dtp2 bilateral teleoperation control system with haptic augmentation, Fus. Eng. Des. 89 (9-10) (2014) 2278–2282.

[3] S. Esque, J. Mattila, M. Siuko, M. Vilenius, J. Järvenpää, L. Semeraro, M. Irving, C. Damiani, The use of digital mock-ups on the development of the divertor test platform 2, Fus. Eng. Des. 84 (2-6) (2009) 752–756.

[4] L. Niu, S. Smirnov, J. Mattila, A. Gotchev, E. Ruiz, Robust pose estimation with a stereoscopic camera in harsh environments, Electron. Imaging 2018 (9) (2018) 1–6.

[5] L. Niu, O. Suominen, M.M. Aref, J. Mattila, E. Ruiz, S. Esque, Eye-in-hand manipulation for remote handling: Experimental setup, in: IOP Conference Series: Materials Science and Engineering, Vol. 320, IOP Publishing, 2018, p. 012007.

[6] R. Y. Tsai, R. K. Lenz. Real time versatile robotics hand/eye calibration using 3d machine vision, in: Robotics and Automation, 1988. Proceedings., 1988 IEEE International Conference on, IEEE, 1988, pp. 554-561.