Kurt Niel
Peter M. Roth
(Eds.)

PROCEEDINGS

OAGM & ARW Joint Workshop 2016 on „Computer Vision and Robotics“

11th–13th May 2016
University of Applied Sciences Upper Austria
Wels Campus

www.fh-ooe.at/oagm-arw2016

TU Graz
OAGM
OESTERREICHISCHE COMPUTER GESELLSCHAFT
AUSTRIAN COMPUTER SOCIETY
GMAR
sapere aude
fh oö
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Kurt Niel and Peter M. Roth (eds.)

Proceedings of the

OAGM & ARW Joint Workshop 2016

Computer Vision and Robotics

May 11-13, 2016
Wels, Austria

Austrian Association of Pattern Recognition
Austrian Robotic Workshop

Editors

Kurt Niel and Peter M. Roth

Layout

Austrian Association of Pattern Recognition
https://www.aapr.at/

GMAR Austrian Robotic Workshop
http://www.roboticsworkshop.at

Cover

Nicola Spitzer

Sponsors

ISBN (print) 978-3-85125-527-0
ISBN (e-book) 978-3-85125-528-7
DOI 10.3217/978-3-85125-528-7

Preface

The OAGM and ARW Joint Workshop on Computer Vision and Robotics provides a platform bringing together researchers, students, professionals and practitioners from both research directions to discuss new and emerging technologies in the field of machine driven perception and automated manipulation/autonomous movement. Even though there is a long tradition for OAGM workshops (we are celebrating the 40th workshop since 1980) and the ARW workshops (since 2011), which have their roots in the early days of the Austrian RoboCup workshops (2006), this is the first time that both communities are organizing a joint event.

Computer Vision tries to perceive the physical world from image or video data resulting in applications such as scene understanding, object detection and tracking and 3D reconstruction. Thus, the main problems are to find suitable representations and to design and implement efficient (learning) algorithms. In contrast, Robotics aims at dealing with moving arms, graspers, and eventually moving vehicles. There are one or more actuators which have to be controlled accordingly in a planned matter for fulfilling given jobs. Some of them consist of additional sensors, e.g., graspers get some feedback for they can correctly catch and hold object without losing or destroying it; or the mobile device stops in front of an obstacle. These examples clearly demonstrate the relations between both fields. The outer world/the actual scenery is perceived by cameras; a consistent set of knowledge is modeled for the actuator for operating successfully either in a planned or even in an unplanned – standalone – strategy. Thus, there is a considerable interest in describing approaching features and possibilities and how the combination of different technologies could be beneficial.

The aim of the joint workshop is to discuss latest academic and industrial approaches and to demonstrate the recent progress. The call for papers resulted in 28 full paper submissions and additional 9 papers submitted to the industrial/featured talk and poster track, where finally according to the reviews of an international programme committee 34 contributions (26 talks, 8 posters) have been selected for presentation at the workshop. The goal of the workshop is also supported by inviting five internationally established researchers, i.e., Oliver Bimber (JKU Linz), Ales Leonardis (BHAM, UK), Laurent Resquet (TIMA, FR), Andreas Müller (JKU Linz), Andreas Nüchter (JMU, DE), representing both areas.

Kurt Niel (General chair of the workshop)
Peter M. Roth (Chairman OAGM)
Markus Vincze (Chairman ARW)
Wels, May 1, 2016

General Chair

Kurt Niel (FH Upper Austria)

Programme Chairs OAGM

Wilhelm Burger (FH Upper Austria)
Bernhard Moser (SCCH)
Peter M. Roth (TU Graz)

Programme Chairs ARW

Burkhard Stadlmann (FH Upper Austria)
Michael Zauner (FH Upper Austria)
Markus Vincze (TU Wien)

Secretary and Local Arraggement

Nicola Spitzer (FH Upper Austria)
Marion Minnich (FH Upper Austria)

Programme Committee OAGM

Helmut Ahammer (Medical University of Graz)
Nicole Artner (TU Wien)
Horst Bischof (TU Graz)
Christia Eitzinger (Profactor)
Friedrich Fraundorfer (TU Graz)
Harald Ganster (Joanneum Research)
Martin Hirzer (TU Graz)
Florian Kleber (TU Wien)
Matej Kristan (University of Ljubljana)
Walter G. Kropatsch (TU Wien)
Arjan Kuijper (Fraunhofer IGD)
Roland Kwitt (University of Salzburg)
Christoph Lampert (IST Austria)
Vincent Lepetit (TU Graz)
Mathias Lux (Alpen-Adria-Universität Klagenfurt)
Hubert Mara (Heidelberg University)
Branislav Micusik (AIT)
Roland Perko (Joanneum Research)
Justus Piater (University of Innsbruck)
Thomas Pock (TU Graz)
Christian Reinbacher (TU Graz)
Josef Scharinger (JKU Linz)
Andreas Uhl (University of Salzburg)
Martin Urschler (LBI Clinical Forensic Imaging)
Martin Welk (UMIT Hall/Tyrol)
Christopher Zach (Toshiba Research Europe)

Programme Committee ARW

Christian Bettstetter (Alpen-Adria-Universität Klagenfurt)
Alessandro Gasparetto (Università degli Studi di Udine)
Hubert Gattringer (JKU Linz)
Wilfried Kubinger (FH Technikum Wien)
Justus Piater (University of Innsbruck)
Gerald Steinbauer (TU Graz)
Ales Ude (Jozef Stefan Institute)
Hubert Zangl(Alpen-Adria-Universität Klagenfurt)

Awards 2015

The

OAGM Best Paper Award 2015 sponsored by OCG

was awarded to the paper

The Minimum Spanning Tree of Maximum Entropy

by

Samuel de Sousa and Walter Kropatsch.

The

Microsoft Visual Computing Award 2015

was awarded to

Bernd Bickel (IST Austria).

Index of authors

Contents

Keynote Talks

A Holonomic Robot for Rescue Applications

Oliver Bimber

JKU Linz, Institute of Computer Graphics, Austria
oliver.bimber@jku.at

Abstract
This talk summarizes our progress towards a fully transparent, flexible, and scaelable thin-film image sensor. In contrast to conventional image sensors, it does not capture pixels in image space on the sensor surface, but makes integral measurements in Radon space along the sensor's edges. Image reconstruction is achieved by inverse Radon transform. By stacking multiple layers, it enables a variety of information, such as color, dynamic range, spatial resolution, and defocus, to be sampled simultaneously. Multi-focal imaging allows reconstructing an entire focal stack after only one recording. The focal stack can then be applied to estimate depth from defocus. Measuring and classifying directly in Radon space yields robust and high classification rates. Dimensionality reduction results in task-optimized classification sensors that record a minimal number of samples. This enables simple devices with low power consumption and fast read-out times. Combining our sensing approach with lensless coded aperture imaging has the potential to enable entire thin-film camera systems that make the capturing of images, light fields, and depth information possible.

Hierarchical Compositional Representations of Structure for Computer Vision and Robotics

Ales Leonardis

University of Birmingham, School of Computer Science, United Kingdom
a.leonardis@cs.bham.ac.uk

Abstract
Modelling, learning, recognising, and categorising visual entities has been an area of intensive research in the vision and robotics communities for several decades. While successful partial solutions tailored for particular tasks and specific scenarios have appeared in recent years, more general solutions are yet to be developed. Ultimately, the goal is to design and implement proper structures and mechanisms that would enable efficient learning, inference, and, when necessary, augmentation and modifications of the acquired visual knowledge in general scenarios. Recently, it has become increasingly clear that possible solutions should be sought in the framework of hierarchical architectures. Among various design choices related to hierarchies, compositional hierarchies show a great promise in terms of scalability, real-time performance, efficient structured on-line learning, shareability, and knowledge transfer. In this talk I will first present our work on compositional hierarchies related to visual representations of 2D and 3D object shapes for recognition and grasping and then conclude with some ideas towards generalising the proposed approach to other visual entities and modalities.

Event-based Design for Mitigating Energy in Electronic Systems

Laurent Fresquet

Laboratoire TIMA, France
laurent.fresquet@imag.fr

Abstract
Today, our digital society exchanges data flows that are incredibly large and the future promises us a data explosion due to the communications between our technological equipment, robots, etc. Indeed, we are close to widely open the door of the Internet of Things (IoT). This data orgy will waste a lot of energy and will contribute to a non-ecological approach of our digital life. Indeed, the Internet and the new technologies consume about 10% of the electrical power produced in the world. Considering that we are only at the beginning of the IoT, it is urgent to enhance the energetic performances of the electronic circuits and systems. The design paradigm based on synchronizing digital circuit communication with a clock is source of useless activity and of complicated design techniques. The digital circuit design based on local synchronizations, also called asynchronous circuits, is a way to mitigate the power consumption in electronics by only activating the circuitry when an event appears. In addition, another way to reduce energy is to rethink the sampling techniques and digital processing chains. Indeed, by using the Shannon theory, we produce more data than necessary. Indeed, useless data produce more computation, more storage, more communications and also more power consumption. If we go beyond the Shannon theory, we can discover new sampling schemes and new processing techniques able to take advantage of event-based design. Drastically reducing the useless data and activity is maybe the Grail of low-power computing.

Model-Based Control of Industrial Robots – From Theory to Practice

Andreas Müller

JKU Linz, Institute of Robotics, AT, Austria
a.mueller@jku.at

Abstract
Industrial robotics has seen a major overhaul in terms of improved designs, novel kinematics, and actuation concepts. Redundancy, for instance, is becoming an important factor for increasing flexibility and robustness. As such, kinematic redundancy of serial manipulators (mimicking anthropomorphicarms) and actuation redundancy of parallel manipulators are prevailing concepts. Aiming at reducing energy consumption and increasing agility, light-weight robotics is another example of innovation in robotics. While these may not be at the core interest of a majority of robot end users, reducing production and cycle times was and still is an important issue. The solution concept applicable to all these problems is the model-based control. In contrast to classical decentralized control schemes, which are commonly used in industrial robots, model-based control schemes make use of a dynamical model. Standard control systems do not account for such models. This will be vital, however. In this presentation the basic concept of model-based control will be discussed. Particular attention will be given to efficient formulations of the dynamic model accounting for rigid as well as elastic manipulators. Strategies for the geometric calibration and the identification of dynamic parameters will be presented. It will be shown how these concepts can seamlessly integrated in industrial controller hardware.

SLAM goes Industry 4.0 – Mobile Laser Scanning for Flexible Production

Andreas Nüchter

Julius-Maximilans-University Würzburg, Informatics VII: Robotics and Telematics, Germany
nuechter@informatik.uni-wurzburg.de

Abstract
The terrestrial acquisition of 3D point clouds by laser range finders has recently moved to mobile platforms. Mobile laser scanning puts high requirements on the accuracy of the positioning systems and the calibration of the measurement system. We present a novel algorithmic approach to the problem of calibration with the goal of improving the measurement accuracy of mobile laser scanners. We developed a general framework for calibrating mobile sensor platforms that estimates all configuration parameters for any configuration of positioning sensors including odometry. In addition, we present a novel semi-rigid SLAM algorithm that corrects the vehicle position at every point in time along its trajectory, while simultaneously improving the quality and precision of the entire acquired point cloud. Using this algorithm the temporary failure of accurate external positioning systems or the lack thereof can be compensated for. We demonstrate the capabilities of our two newly proposed algorithms on a wide variety of data sets. Applications for the developed suite of algorithms range from 3D mapping for autonomous driving to precise digitization of production lines in the automotive context. We end the talk with a description of an innovative start-up in the area of robotic SLAM.

Industrial Applications/Featured Talks

One-Shot Learning of Scene Categories via Feature Trajectory Transfer

Roland Kwitt[1], Sebastian Hegenbart[1], Marc Niethammer[2]

[1] University of Salzburg, Austria;
[2] University of North Carolina, Chapel Hill, NC, USA

Abstract

The appearance of (outdoor) scenes changes considerably with the strength of certain transient attributes, such as ``rainy'', ``dark'' or ``sunny''. Obviously, this also affects the representation of an image in feature space, e.g., as activations at a certain CNN layer, and consequently impacts scene recognition performance. In this work, we investigate the variability in these transient attributes as a rich source of information for studying how image representations change as a function of attribute strength. In particular, we leverage a recently introduced dataset with fine-grain annotations to estimate feature trajectories for a collection of transient attributes and then show how these trajectories can be transferred to new image representations. This enables us to synthesize new data along the transferred trajectories with respect to the dimensions of the space spanned by the transient attributes. Applicability of this concept is demonstrated on the problem of one-shot scene recognition. We show that data synthesized via feature trajectory transfer considerably boosts recognition performance, (1) with respect to baselines and (2) in combination with state-of-the-art approaches in one-shot learning.

Directional Wavelet based Features for Colonic Polyp Classification

Georg Wimmer[1], Michael Häfner[3], Shigeto Joshida[4], Toru Tamaki[5], Shinji Tanaka[4], Jens Tischendorf[2] and Andreas Uhl[1]

[1] University of Salzburg, Austria; [2] RWTH Aachen University Hospital;
[3] St. Elisabeth Hospital; [4] Hiroshima University Hospital; [5] Hiroshima University

Abstract

In this work, various wavelet based methods like the discrete wavelet transform, the dual-tree complex wavelet transform, the Gabor wavelet transform, curvelets, contourlets and shearlets are applied for the automated classification of colonic polyps. The methods are tested on 8 HD-endoscopic image databases, where each database is acquired using different imaging modalities (Pentax's i-Scan technology combined with or without staining the mucosa), 2 NBI high-magnification databases and one database with chromoscopy high-magnification images. To evaluate the suitability of the wavelet based methods with respect to the classification of colonic polyps, the classification performances of 3 wavelet transforms and the more recent curvelets, contourlets and shearlets are compared using a common framework. Wavelet transforms were already often and successfully applied to the classification of colonic polyps, whereas curvelets, contourlets and shearlets have not been used for this purpose so far. We apply different feature extraction techniques to extract the information of the subbands of the wavelet based methods. Most of the in total 20 approaches were already published in different texture classification contexts. Thus, the aim is also to assess and compare their classification performance using a common framework. Three of the 20 approaches are original. These three approaches extract Weibull features from the subbands of curvelets, contourlets and shearlets. Additionally, 5 state-of-the-art non wavelet based methods are applied to our databases so that we can compare their results with those of the wavelet based methods. It turned out that extracting Weibull distribution parameters from the subband coefficients generally leads to high classification results, especially for the dual-tree complex wavelet transform, the Gabor wavelet transform and the Shearlet transform. These three wavelet based transforms in combination with Weibull features even outperform the state-of-the-art methods on most of the databases. We will also show that the Weibull distribution is better suited to model the subband coefficient distribution than other commonly used probability distributions like the Gaussian distribution and the generalized Gaussian distribution.

DeVisOR - Detection and Visualization of Unexploded Ordnance Risks*

Sebastian Zambanini, Fabian Hollaus, and Robert Sablatnig

Computer Vision Lab, Institute of Computer-Aided Automation, TU Wien, Austria
{*zamba,holl,sab*}*@caa.tuwien.ac.at*

The project 'Detection and Visualization of Unexploded Ordnance Risks' (DeVisOR) is devoted to the analysis of historical aerial images. These images are currently investigated by experts in order to detect UneXploded Ordnances (UXO) [3]. For this purpose, the aerial images have to be georeferenced first which is accomplished by a manual registration of the images onto modern satellite images by means of a professional GIS software tool. Afterwards, the experts detect suspicious image regions by looking for characteristic shapes or patterns. Additionally, images captured at different time instances are compared in order to detect changes of the scene, which might stem from bombs or other events related to military operations.
A problem of this current practice is that its manual steps are tedious and taxing. Thus, analysis takes a long time and intense reviewing is necessary. An automated analysis could obviously solve the tasks faster and less tiresome. The DeVisOR project aims at developing tools that support the work of the experts by making use of methods originated from the fields of computer vision and visual analytics. The main computer vision tasks can be grouped into two categories: automated image registration and object detection.

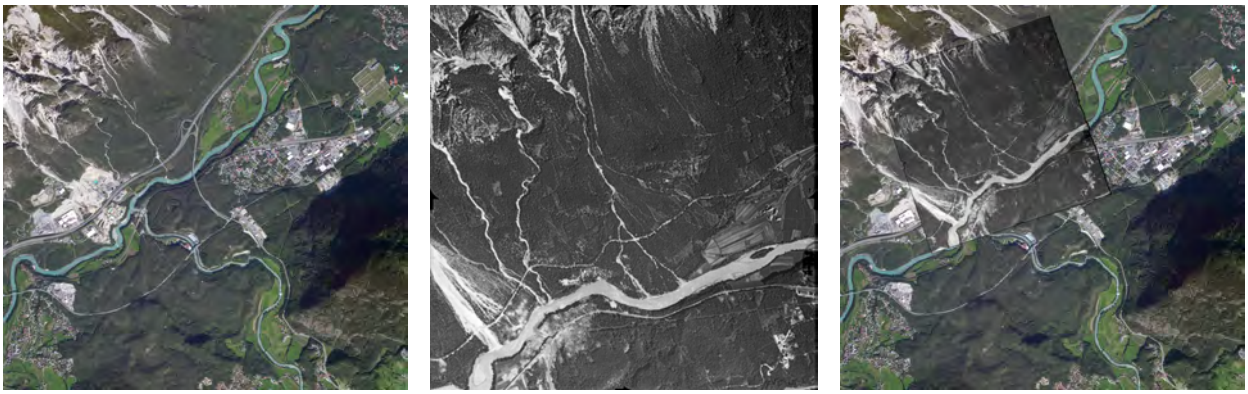
Image Registration

This task is concerned with the automatic georeferencing of the historical aerial images. By taking modern satellite image as reference, this task can be approached as a classical image registration problem [5], as illustrated in Figure 1. The main challenge are the strong changes in image content caused by the age differences of around 70 years between the old and new images that hinder the reliable identification of correspondences, especially in non-urban areas. Additionally, the historical images are partially in a poor condition, meaning they are affected by over- or underexposure, uneven illumination, low spatial resolution, blurring, sensor noise or cloud coverage. Consequently, a straightforward solution based on standard algorithms using keypoint matching [4] and robust transformation estimators [2] does not exist.

Object Detection

The second task is dedicated to the automated detection of military objects (e.g. bomb craters or trenches) and assignment of prediction probabilities to the objects found. The task is hindered by the low quality of the images investigated and their high variety. Due to the absence of large amounts of training data, we are planning to implement and evaluate semi-supervised and active learning procedures [1], which will also make use of techniques stemming from the field of visual analytics.

*This work is supported by Austrian Research Promotion Agency (FFG) under project grant 850695.



Reference satellite image (Ötztal region).

Historical aerial photo from May 1945.

Result of manual georeferencing.

Reference satellite image (Vienna, 21st district).

Historical aerial photo from November 1943.

Result of manual georeferencing.

Figure 1: Two examples illustrating the process of manually georeferencing historical aerial photos by image registration.[1]

References

[1] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney S Tan. Effective end-user interaction with machine learning. In *AAAI Conference on Artificial Intelligence*, pages 1529–1532, 2011.

[2] A. Ardeshir Goshtasby. Robust parameter estimation. In *Image Registration: Principles, Tools and Methods*, pages 313–341. Springer, 2012.

[3] Andrew E Hooper. Unexploded Ordnance (UXO): The Problem. *Detection and Identification of Visually Obscured Targets*, page 1, 1998.

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[1] Historical photos are provided by Luftbilddatenbank Dr. Carls GmbH.

Subpixel Localisation of Nanoparticles in Images *

Thomas Hoch[1], Matthias Dorfer[1], and Clemens Helmbrecht[2]

[1] SCCH, Austria
{*thomas.hoch,matthias.dorfer*}*@scch.at*

[2] Particle Metrix GmbH, Germany
helmbrecht@particle-metrix.de

Abstract
Nanoparticle Tracking Analysis (NTA) is an emerging technology for the quantification of particle size, concentration and zeta potential for particles in the size regime of 10 to 1000 nm. The technique allows the visualization of the Brownian motion of particles in liquid suspensions. It is frequently used in commercial and academic applications for the analysis of the physical and chemical properties of dispersions such as solubility, rheology and reactivity which are strongly influenced by the size of the respective particles. Hence, measuring the size of micro- or even nano- sized particles in dispersions plays a central role in chemical and biomedical industries.

With the NTA technique, particles dispersed in liquids are illuminated with an intensive light beam, e.g. from a laser. An image series of the light scattered by the particles is recorded with a sensitive digital camera with a magnification microscope attached to it. From the image series the Brownian motion of the particles is analyzed by first localizing the particle in each video frame, second tracking of the particles from frame to frame, and third computing the Mean Squared Displacement (MSD) along the track of each individual particle. Having the MSD one can estimate the particles diffusion coefficient and apply the Stoke-Einstein relationship to estimate the hydrodynamic size of individual particle. Current NTA systems use background segmentation method to differentiate the particles from background, mostly with fixed threshold approach. Fixed threshold works well for mono-modal dispersion since the brightness of the particles is evenly distributed. Poly-disperse particle solutions on the other hand show a high variation in the particle intensity because the reflected light intensity depends on the particle size and thus it is difficult to find a fixed threshold value.

We propose a new method for NTA which utilizes a multi-scale Laplacian of a Gaussian (LoG) detector on top of the background-subtraction model to localize the particles. Our approach uses an optimized thresholding method for each blob individually to compute a super-resolution position estimate. We show that our method finds more particles in the video with higher precision over the full size-range of tested solutions (20nm-500nm) in comparison to the fixed threshold approach. We further show that the increased efficiency in particle tracking and the higher precision in the localization of the particle center leads to particle size distributions that are narrower (having less variance). Thus, our method is in particular better suited for the analysis of mixtures of poly-disperse particle solutions if the size of the particles in the mixture solution is not too far apart.

*The research reported in this article has been partly supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

The 3D-PITOTI Project with a Focus on Multi-Scale 3D Reconstruction using Autonomous UAVs *

Christian Mostegel[1], Georg Poier[1], Christian Reinbacher[1], Manuel Hofer[1], Friedrich Fraundorfer[1], Horst Bischof[1] , Thomas Höll[2], Gert Holler[2] and Axel Pinz[2]

[1] Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*lastname*}@*icg.tugraz.at*

[2] Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Austria
{*firstname*}.{*lastname*}@*tugraz.at*

Abstract

In this talk, we showcase our outcome of the ambitious 3D-PITOTI project, which involves a multi-disciplinary team of over 30 scientists from across Europe. The project focuses on the 3D aspect of recording, storing, processing and visualizing prehistoric rock art in the UNESCO World Heritage site in Valcamonica, Italy. The rock art was pecked into open-air rock formations thousands of years ago and has an inherent 3D nature.

After a project overview, we present the results of the Graz University of Technology's contributions in 3D acquisition and processing with a focus on our novel autonomous UAV system. We elaborate the challenges of 3D reconstruction across vastly different scales, from a valley wide reconstruction down to individual peckings on the rock surface [1]. Within this context, we first present a novel 3D scanning device with sub-millimeter accuracy [2]. Aside from correctly scaled 3D information, the scanning device also provides the surface radiometry without the need for artificial shrouding [3]. Additionally, we point out one application for which this highly accurate 3D data has shown to be crucial: The interactive segmentation of the individually pecked figures [7, 8].

Finally, we present a novel autonomous UAV system for acquiring high-resolution images at a few meters distance [6, 5, 4]. The system optimizes scene coverage, ground resolution and 3D uncertainty, while ensuring that the acquired images are suitable for a specific dense offline 3D reconstruction algorithm. There are three main aspects that set this system apart from others. First, the system operates completely on-site without the need for a prior 3D model of the scene. Second, the system iteratively refines a surface mesh, predicts the fulfillment of requirements and can thus correct for initially wrong geometry estimates and imperfect plan execution. Third, the system uses the already acquired 2D images to predict the chances of a successful reconstruction with a specific offline 3D densification algorithm depending on the observed scene and potential camera constellations. We demonstrate the capabilities of our system in the challenging environment of the prehistoric rock art sites and then register the individual reconstructions of all scales in one consistent coordinate frame.

*The research leading to these results has received funding from the EC FP7 project 3D-PITOTI (ICT-2011-600545). We would like to thank all colleagues and the consortium of the 3D-PITOTI project for the fruitful collaboration.

References

[1] Craig Alexander, Axel Pinz, and Christian Reinbacher. Multi-scale 3d rock-art recording. *Digital Applications in Archaeology and Cultural Heritage*, 2(2-3):181 – 195, 2015. Digital imaging techniques for the study of prehistoric rock art.

[2] Thomas Höll, Gert Holler, and Axel Pinz. A novel high accuracy 3d scanning device for rock-art sites. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(5):285, 2014.

[3] Thomas Höll and Axel Pinz. Cultural heritage acquisition: Geometry-based radiometry in the wild. In *International Conference on 3D Vision (3DV)*, pages 389–397, Oct 2015.

[4] Christian Mostegel, Markus Rumpler, Friedrich Fraundorfer, and Horst Bischof. UAV-based Autonomous Image Acquisition with Multi-View Stereo Quality Assurance by Confidence Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016. [accepted for publication].

[5] Christian Mostegel, Markus Rumpler, Friedrich Fraundorfer, and Horst Bischof. Using Self-Contradiction to Learn Confidence Measures in Stereo Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. [accepted for publication].

[6] Christian Mostegel, Andreas Wendel, and Horst Bischof. Active monocular localization: Towards autonomous monocular exploration for multirotor MAVs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3848–3855, May 2014. [Best Student Paper Award - Finalist].

[7] Matthias Zeppelzauer, Georg Poier, Markus Seidl, Christian Reinbacher, Christian Breiteneder, Horst Bischof, and Samuel Schulter. Interactive segmentation of rock-art in high-resolution 3d reconstructions. In *2015 Digital Heritage*, volume 2, pages 37–44, Sept 2015. [Best Paper Award].

[8] Matthias Zeppelzauer, Georg Poier, Markus Seidl, Christian Reinbacher, Samuel Schulter, Christian Breiteneder, and Horst Bischof. Interactive 3d segmentation of rock-art by enhanced depth maps and gradient preserving regularization. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 0(0):1–28, 2016. [accepted for publication].

WS 1: Learning / Recognition

Semantic Labeling Enhanced by a Spatial Context Prior

Daniel Steininger, Csaba Beleznai

Austrian Institute of Technology, Austria
Daniel.Steininger.fl@ait.ac.at
Csaba.Beleznai@ait.ac.at

Abstract
Our observed visual world exhibits a structure, which implies that scene objects and their surroundings are not randomly arranged relative to each other but typically appear in a spatially correlated manner. Thus, the structural correlation can be exploited to make the visual recognition task predictable to a certain extent. Modeling relations between categories is, however, non-trivial, since categories are often represented at different granularities across distinct datasets. In this paper, we merge fine-level semantic descriptions into basic semantic classes which allows the generation of spatial contextual priors from a wide range of datasets. In this way, a contextual model is derived with the objective to employ the learned contextual prior to enhance visual recognition via improved semantic labeling. The prior is captured explicitly by computing occurrence and co-occurrence probabilities of specific semantic classes and class pairs from a diverse set of annotated datasets. We show improved semantic labeling accuracy by incorporating the contextual priors into the label inference process, which is evaluated and discussed on the Daimler Urban Segmentation 2014 dataset.

1. Introduction

Semantic segmentation of digital images links two core computer vision challenges: visual object recognition and segmentation. In recent years, great improvement in accuracies to both task domains has been demonstrated, mainly due to a transition from learned hand-crafted representations towards representations distributed within hierarchies and embedded into compositional schemes, enabling a rich generalization for a large number of object classes.

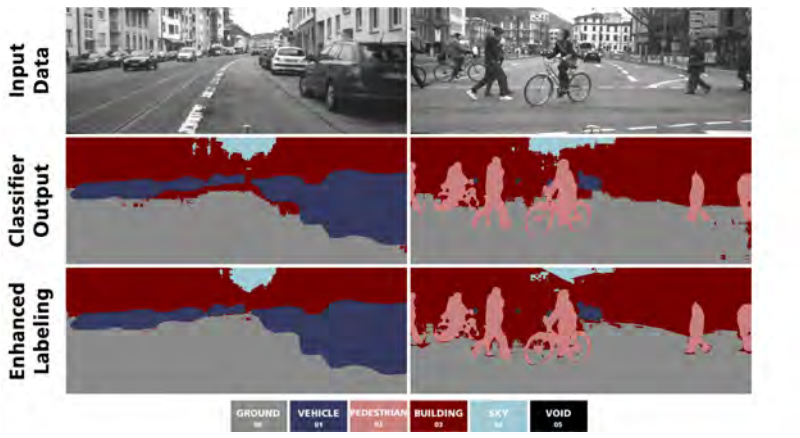


Figure 1. Semantic Labeling enhanced by a Spatial Context Prior and Conditional Random Field.

As representations and learning schemes have grown capable of accommodating the sheer variability in the data, this progress is also imposing new requirements on the employed datasets. Current learned models are often optimized for specific datasets they have been trained on, and their capture modalities are restricted by their implicit design. Real world scenarios are highly diverse, therefore, a single dataset solely represents a small fraction of all possible visual appearances. Although datasets have become more elaborate and diverse lately [17], class coverage, balancing and variability are still relevant issues to be tackled. Motivated by the diversity in the characteristics of prevailing datasets, in terms of number and granularity of annotated classes and scene-specific view attributes, we propose to capture the spatial relationship between various semantically labeled regions across several datasets. We demonstrate that the modeled spatial prior can enhance recognition accuracies leading to state-of-the-art results, as illustrated in Figure 1.

2. Related Work

Spatial context is an important type of information in the human cognitive process [12] when recognizing objects, especially in the presence of a cluttered background. Certain objects predominantly co-occur in the real world. Thus analyzing vast amounts of visual data can result in meaningful contextual statistics which can be used to robustify visual object recognition [5].

Pixel-wise semantic labeling is a relatively novel domain since large-scale object recognition with shared informative representations is a prerequisite for this task. Starting with manually selected low-level features, discriminatively trained Random Forests or Boosting have been used to perform classification patch-wise [16] or to additionally incorporate local structural information within the analysis patch [7]. Based on recent advances in deep learning, several frameworks [13, 18] have demonstrated significant improvements in the accuracy of per-pixel class estimates. Recently, multi-scale deep architectures have been proposed in order to represent local and global context by employing multiple input images at different resolutions [2], or combining feature maps from different layers of the convolutional architecture [6]. Both techniques aim to combine fine detail representations with relational information established at a coarse resolution level in order to generate accurate segment boundaries between labeled regions. The immense representational power of deep convolutional architectures captures rich details of the object classes to be represented and yields segmentation frameworks which surpass learned hand-crafted representations. Capturing spatial context within convolutional architectures, however, is linked with complexities in terms of training (augmented parameter space) and increased computational expense due to the computation of multiple scale-specific features.

Our proposed approach employs a previously learned spatial prior model as an additional step to switch class labels at locations where per-pixel estimates are ambiguous. We term our model as the *Explicit Priors* model. Per-pixel ambiguity is quantified from class posterior probabilities at the given pixel by examining the distance between first and second rank probabilities. Our method, while limited in representing spatial context at a wide range of spatial scales and orientations, yields a remarkable improvement at a negligible increase of computational complexity.

3. Methodology and Experimental Setup

The proposed approach for combining learned information from multiple datasets and thereby enhancing existing classifiers is based on the concept of *Explicit Priors*. By aggregating statistical data on the level of individual pixels and capturing spatial context, we generate additional cues for training

and classification, while remaining independent of the underlying machine learning algorithm. The method and its integration throughout the entire processing pipeline is described in this chapter and demonstrated on the Daimler Urban Segmentation 2014 dataset [14].

The dataset consists of image sequences captured by a camera mounted on a moving car. The images are provided without color information at a resolution of 1024x440 px, with every 10th frame of the sequences being annotated with pixel-wise segmentations. For a reasonable comparison, only the test sequences, as specified by the evaluation protocol, are considered. The dataset is supplemented with precomputed disparity maps and additional information, like time-stamps, vehicle speed and yaw rate. The ground truth distinguishes between two foreground (*Vehicle* and *Pedestrian*) and three background classes (*Ground*, *Sky* and *Building*). Within the test data 36.3% of all pixels are defined as *Void*. The frequency of occurrence of the labeled pixels is 54.1% for *Ground*, 14.8% for *Vehicle*, 4.6% for *Pedestrian*, 2.4% for *Sky* and 24.0% for *Building*, resulting in a background ratio of 80.6%.

3.1. Training

Dataset Analysis As a preliminary step for the training and classification process, an appropriate choice of input data with regard to the intended application scenario is a decisive aspect. For this purpose, a statistical analysis of multiple datasets was conducted according to the concept of *Explicit Priors*. The resulting data ranges from basic statistics, such as label frequency and the ratio of background to foreground classes, to more sophisticated aspects concerning occurrence distribution and spatial context. For each application scenario, this dataset analysis can be used to select a subset of additional cues for identifying appropriate datasets. For the demonstrated task, for instance, the most useful information was provided by the concept of *Location Bins*. By dividing the image dimensions into a coarse grid and capturing the spatial distribution of each class across the resulting cells over the entire dataset, probabilities for the occurrence of certain labels with regard to their location can be derived. The resulting representation provides clearly arranged patterns closely related to certain characteristics of the dataset, such as the method of image acquisition. In the case of Vehicles, for instance, the analysis clearly showed that images taken with a hand-held camera are mostly centered on the these objects, while for the datasets using a camera mounted on a car they are most often found in the lower half of the image. Comparing these statistics for candidate training datasets to the intended application scenario facilitates the evaluation of their compatibility.

Other available statistical measures proved to add less distinct cues for the given task, such as the analysis of co-occurrence, which provides a measure of probability for each combination of labels to appear in the same image. Since the application scenario only includes five labels arranged within a consecutive image sequence, the resulting correlation matrix did not show significant peaks. However, an adapted version in the form of *Local Label Neighborhood* (*LLN*), which limits the co-occurance measure to label transitions, was successfully applied, as described in detail in Section 3.2.

Based on the aggregated information of class frequency and *Location Bins*, the CamVid dataset [1] could be identified as an appropriate choice for training background classes, since it offers a background ratio of 80.9%, as well as a fitting spatial arrangement of class probabilities. The foreground classes, on the other hand, are trained on the PascalContext dataset [11], in particular the version including 33 categories, which contains 46% foreground pixels.

Classifier Setup Based on the selected datasets, two classifiers are applied to cover the background and foreground classes separately. The former classifier uses the pre-trained model pascal-fcn8s-tvg-

dag provided by Zheng et al. [18], which is evaluated on the foreground classes of the PascalContext33 dataset. The background classifier was trained using TextonBoost [8] on randomly sampled images of the CamVid dataset. For this purpose, feature descriptors based on filter banks, location and gradient orientations were applied for training a total of 950 Textons, which represents a compromise between computational complexity and accuracy. Since the test dataset consists of gray-scale images, the learning input is restricted to the intensity channel.

Label Aggregation and Mapping The main obstacle in aggregating multiple datasets during the training stage results from variations in the denomination of object classes. Furthermore, since in many cases not all labels of the training datasets are required for classifying the test images, and multiple labels of one dataset can relate to a single label of another, a generalized mapping strategy is a prerequisite for combining label information. For this purpose, an automatic method for label clustering was developed based on version 3.0 of the Wordnet database [10]. This knowledge representation was trained exclusively on lexical data and is capable of providing a similarity measure among semantic descriptions. Based on this, labels of the training dataset can be assigned to the final denominations by applying a threshold and giving preference to classes with higher similarity.

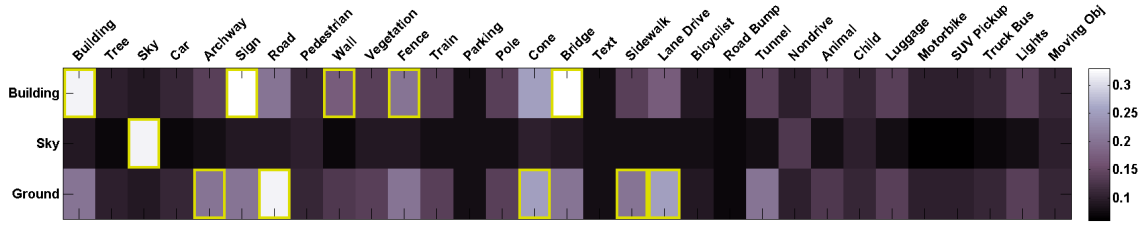


Figure 2. Label Mapping of CamVid (Columns) to Daimler (Rows) dataset based on Wordnet similiarity (selected labels are marked in yellow color).

In the case of the CamVid dataset this process resulted in a selection of eleven labels, as visualized in Figure 2, while the remaining ones are not required for the application task and therefore suppressed. The selected labels were assigned to the background classes *Building*, *Sky* and *Ground* of the final dataset based on the corresponding similarity. Analogously, the two foreground objects *Pedestrian* and *Vehicle* are assigned the PascalContext labels of *Pedestrian*, *Bicyclist*, *Child* and *Moving Object*, as well as *Car*, *Motorbike*, *SUV Pickup* and *Truck*, respectively.

3.2. Classification

The foreground and background classifiers are applied to each input image of the test set resulting in two complementary segmentations, which are further refined by applying the label mapping method described in Section 3.1. This step results in both images being segmented into the labels required by the test dataset. In order to further improve the segmentation quality of background classes, the two highest ranked labels of each pixel are retained, as well as the probability distance between them. This information is required for enhancing the results with *Local Label Neighborhood* priors and further refinement by inference based on a Conditional Random Field (CRF).

Local Label Neighborhood The concept of *Local Label Neighborhood* is based on statistically learning conditional probabilities of transitions between specific labels in vertical and horizontal direction. Each annotated pixel within the selected training images is evaluated to capture this prior based on spatial context. For the given task, this results in a measure of probability for each background class to be found on a specific side of either of the two foreground classes. The probabilities

extracted from the CamVid dataset using this method are weighted by the frequency of occurrence for each background class and aggregated into the final labels, as visualized in Table 1. The learned a-priori knowledge is used to resolve ambiguous classifications.

	Ground	*Sky*	*Building*
LLN Vehicle ↑	0.026	0.005	0.195
LLN Pedestrian ↑	0.025	0.002	0.222
LLN Vehicle ↓	0.383	0.000	0.004
LLN Pedestrian ↓	0.086	0.001	0.081

Table 1. *Local Label Neighborhood* learned on CamVid dataset.

The resulting statistics show the probability of encountering each background class above or below a label transition from each foreground class. For instance, the *Vehicle* prior in upward direction indicates a significant chance of detecting *Buildings* above the class and the prior in downward direction increases the probability of detecting *Ground* below it. Using this information, areas between foreground classes and image borders in vertical direction are marked as candidates for the corresponding background label based on the probability indicated by the prior. If the candidate labels correspond to the second-ranked label for a pixel and the probability distance to the current class is sufficiently low, the second rank is recovered and replaces the first.

Conditional Random Field In order to further increase segmentation accuracy, especially in areas of label transitions, a framework [8] for inference based on CRF is applied with empirically determined parameters. As an input, the existing intermediate background segmentation is integrated in the form of a unary potential with a globally defined confidence of 80%. Additionally, two pairwise potentials, based on label compatibility and intensity information within a defined radius, are added, the latter weighted four times higher than the former. After conducting the inference process in five iterations, the eventual segmentation is combined with the foreground classes.

4. Experiments and Discussion

Semantic Labeling was performed on the test sequences of the Daimler Urban Segmentation 2014 dataset with and without the learned spatial context prior. In order to compare the results to previously published methods, the Intersection-over-Union (*IoU*) metric is used according to the official Pascal VOC definition [3],

$$IoU_{l_i} = \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{1}$$

where $L = \{l_1, ..., l_k\}$ is a set of labels and TP_i, FP_i and FN_i are the true positive, false positive and false negative detections corresponding to label l_i, is used. The detailed results are shown in Table 2. Additionally, we show the average IoU over all classes, as well as a separate average value for the dynamic classes *Vehicle* and *Pedestrian*. The global per-pixel accuracy (PPA) represents the ratio of correctly classified pixels to the total number of annotated pixels in the test dataset. Each column shows the results for the baseline method and its enhancement with the proposed *LLN* and CRF, which are compared to state-of-the-art methods. The best-performing results are displayed in bold numbers.

	Ground	*Vehicle*	*Pedestrian*	*Sky*	*Building*	*Avg*	Avg_{dyn}	*PPA*
Stixmantics [14]	93.8	78.8	66.0	75.4	89.2	80.6	72.4	92.8
ALE [14]	94.9	76.0	73.1	**95.5**	90.6	86.0	74.5	**94.5**
Darwin pw. [4]	95.7	68.7	21.2	94.2	87.6	73.5	44.9	-
PN-RCPN [15]	**96.7**	79.4	68.4	91.4	86.3	84.5	73.8	**94.5**
Layered Ip. [9]	96.4	83.3	71.1	89.5	**91.2**	**86.3**	77.2	-
BL	92.9	**85.5**	**75.4**	54.2	80.1	77.6	**80.5**	92.8
BL | LLN	92.9			54.2	81.3	77.9		93.0
BL | LLN | CRF	94.8			74.1	85.1	83.0		**94.5**

Table 2. Intersection-over-Union measures and Per-Pixel Accuracy (BL: baseline method, LLN: Local Label Neighborhood, CRF: Conditional Random Field).

Compared to recently published approaches, the proposed method leads to an improved segmentation of dynamic classes by 3.3%. The concept of Label Aggregation applied to a pre-trained model proves to be an appropriate choice for both labels. The classification of background classes, on the other hand, is quite competitive for the *Ground* class with a distance of 1.9% to the leading method, while being slightly inferior to the others concerning *Building* and *Sky*. However, these results are still promising, considering several influencing factors. Firstly, the proposed method is presently based exclusively on intensity information, while the other algorithms, except [15], incorporate additional cues such as depth and motion data. However, this limitation can still be partially compensated by the application of *LLN* and CRF. While *LLN* leads to an increase 0.3% concerning the average IoU, CRF contributes an additional 5.1%. For the PPA, improvements of 0.2% and an additional 1.5% can be achieved. An example of the overall results is provided in Figure 3.

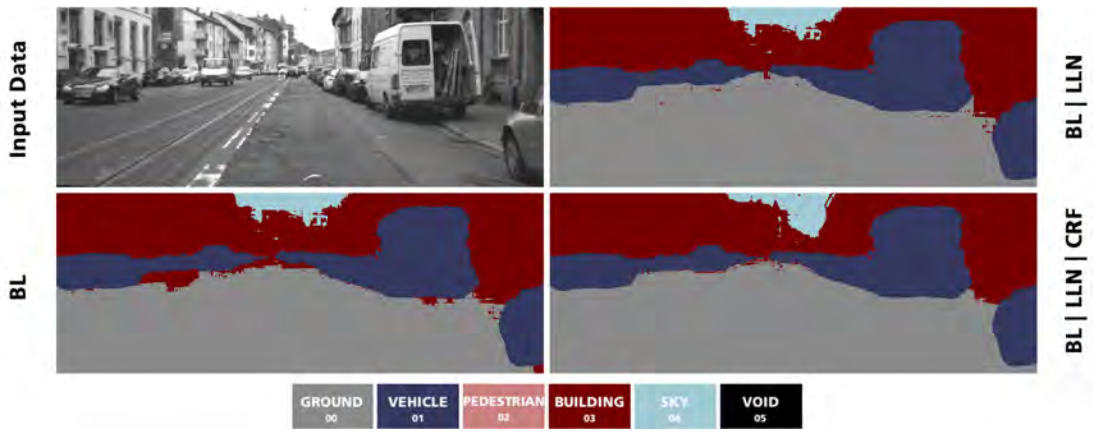


Figure 3. Improvement of segmentation quality of background classes (BL: baseline method, LLN: Local Label Neighborhood, CRF: Conditional Random Field).

Please note that the lowest accuracy corresponds to the *Sky* class, which has a frequency of occurrence of solely 2.4% in the testing dataset. Therefore, its influence on the PPA is almost negligible, which leads to an accuracy equal to the currently best results.

More detailed insights can be retrieved by analyzing precision and recall measures for each class, as displayed in Table 3. Both values present highly promising results for the *Ground* class, which is the most frequent background class. The remaining two background classes show higher inter-dependency. While *Building* offers a high recall but lower precision value, *Sky* shows the opposite characteristics, which indicates that the *Building* class tends to inaccurate over-segmentation into *Sky*

	Ground	*Vehicle*	*Pedestrian*	*Sky*	*Building*	*Avg*
BL	97.8 | 95.0			72.9 | 67.9	82.0 | 97.1	89.6 | 84.8
BL | LLN	97.2 | 95.5	**98.5** | **86.7**	**97.1** | **77.2**	72.9 | 67.9	83.5 | 96.9	89.8 | 84.8
BL | LLN | CRF	**97.7** | **96.9**			**89.1** | **81.5**	**86.3** | **98.4**	**93.7** | **88.1**

Table 3. Precision (left) and recall (right) of each label class.

regions. Concerning the influence of *LLN*, the *Building* class reaches an increase in precision of 1.5% combined with an insignificant decrease of recall. Simultaneously, the optimization leads to a decrease in precision for the *Ground* class, while increasing its recall. It can be concluded that the method successfully recovers misclassified *Ground* pixels originally labeled as *Building*. CRF further increases the average precision and recall by an additional 3.9% and 3.3%, respectively.

5. Conclusions

This paper introduces a concept to capture spatial context between labeled regions for diverse datasets annotated at different semantic granularity, referred to as *Explicit Priors*, which was successfully applied to enhance the entire training and classification process of semantic segmentation demonstrated on the Daimler Urban Segmentation 2014 dataset. The approach provides a generalized way to select an appropriate subset of multiple training datasets and to efficiently combine their labels to fit a given application scenario. The segmentation quality of foreground classes is comparable to, and in terms of certain measures even surpasses, state-of-the-art methods. The results for the background classes proved to be competitive as well. Their relatively high precision, combined with lower recall correspond to a classification accuracy of certain labels slightly inferior to currently leading methods. Further improvements concerning background labeling were achieved by applying priors based on *Local Label Neighborhood* as well as inference using CRF. In order to exploit additional potentials, the next step would be to integrate complimentary modalities, such as depth and motion cues.

Acknowledgments

This work is supported by the research initiative 'Mobile Vision' with funding from the Austrian Federal Ministry of Science, Research and Economy and the Austrian Institute of Technology.

References

[1] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[2] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[3] Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[4] Stephen Gould. Darwin: A framework for machine learning and computer vision research and development. *The Journal of Machine Learning Research*, 13(1):3533–3537, 2012.

[5] Michelle R. Greene, Christopher Baldassano, Andre Esteva, Diane M. Beck, and Li Fei-Fei. Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, 145(1):82, 2016.

[6] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[7] Peter Kontschieder, S. Rota Bulò, Horst Bischof, and Marcello Pelillo. Structured class-labels in random forests for semantic image labelling. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2190–2197, 2011.

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011.

[9] Ming-Yu Liu, Shuoxin Lin, Srikumar Ramalingam, and Oncel Tuzel. Layered interpretation of street view images. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[10] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[11] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[12] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[13] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[14] Timo Scharwächter, Markus Enzweiler, Uwe Franke, and Stefan Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 533–548. Springer, 2014.

[15] Abhishek Sharma, Oncel Tuzel, and David W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 530–538, 2015.

[16] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[17] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[18] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H.S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

Tattoo Detection for Soft Biometric De-Identification Based on Convolutional Neural Networks*

Tomislav Hrkać[1], Karla Brkić[1], and Zoran Kalafatić[1]

[1] Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
{*tomislav.hrkac, karla.brkic, zoran.kalafatic*}*@fer.hr*

Abstract

Nowadays, video surveillance is ubiquitous, posing a potential privacy risk to law-abiding individuals. Consequently, there is an increased interest in developing methods for de-identification, i.e. removing personally identifying features from publicly available or stored data. While most of related work focuses on de-identifying hard biometric identifiers such as faces, we address the problem of de-identification of soft biometric identifiers – tattoos. We propose a method for tattoo detection in unconstrained images, intended to serve as a first step for soft biometric de-identification. The method, based on a deep convolutional neural network, discriminates between tattoo and non-tattoo image patches, and it can be used to produce a mask of tattoo candidate regions. We contribute a dataset of manually labeled tattoos. Experimental evaluation on the contributed dataset indicates competitive performance of our method and proves its usefulness in a de-identification scenario.

1. Introduction

In the last decade, video surveillance has spread to almost all aspects of daily life. Storing the recorded surveillance data in its unprocessed form poses a privacy risk to law-abiding individuals, as their whereabouts and activities can be exposed without their consent. Privacy concerns are aggravated by the development of various video retrieval techniques [17, 26, 16] that enable searching for content in large volumes of video data, as well as by the development of techniques for person re-identification across different video sequences [1, 8]. In order to minimize privacy risks, many jurisdictions implement strict regulations for the protection of personal data (see e.g. the Data Protection Directive of the European Union[1]). For video sequences, protection of personal data entails obfuscating or removing personally identifying features of the recorded individuals, usually in a reversible fashion so that law enforcement can access them if necessary.

The process of removing personally identifying features from data is called de-identification. One of the most commonly used de-identification techniques, used in commercial systems such as Google Street View, involves detecting and blurring the faces of recorded individuals. However, this approach ignores soft biometric and non-biometric features like clothing, hair color, birthmarks or tattoos, that can be used as cues to identify the person [6, 20]. In this paper, we propose a method for detecting

*This work has been supported by the Croatian Science Foundation, within the project "De-identification Methods for Soft and Non-Biometric Identifiers" (DeMSI, UIP-11-2013-1544). This support is gratefully acknowledged.

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046

tattooed skin regions that can be used in an advanced de-identification pipeline to obfuscate or remove tattoos. We train a convolutional neural network that acts as a patch classifier, labeling each patch of an input image as either belonging to a tattoo or not.

2. Related work

Current research in de-identification is mainly concerned with de-identifying hard biometric features, especially the face [9]. Considerably less volume of research is devoted to soft and non-biometric features [20]. Tattoo detection is typically studied not in the context of de-identification, but in forensic applications. There the goal is to build a content-based image retrieval system for tattoos that would help law enforcement in finding suspects and other persons of interest, e.g. persons associated with a particular gang etc. [6, 12, 10]. For instance, Jain et al. [12] propose a content-based image retrieval system intended to be used by law enforcement agencies. The query image is a cropped tattoo, which is then segmented, represented using color, shape and texture features and matched to the database. Han and Jain [10] take the concept further by proposing a content-based image retrieval system for sketch-to-image-matching, where a sketch of the tattoo is matched to real tattoo images. Their system uses SIFT descriptors to model shape and appearance patterns in both the sketch and the image, and matches the descriptors using a local feature-based sparse representation classification scheme. Kim et al. [13] propose combining local shape context, SIFT descriptors and global tattoo shape for tattoo image retrieval. Their descriptor is robust to partial shape distortions and invariant to translation, scale and rotation.

The methods used in content-based image retrieval systems often assume that tattoo images are cropped, which limits their potential use in other scenarios. Heflin et al. [11] consider detecting scars, marks and tattoos "in the wild", i.e. in uncropped images, where a tattoo can appear anywhere in the image (or not appear at all) and be of arbitrary size. They propose a method for tattoo detection where tattoo candidate regions are detected using graph-based visual saliency. Further processing of the candidate regions utilizes the GrabCut algorithm [21], image filtering and the quasi-connected components technique [4] to obtain the final estimate of the tattoo location.

Wilber et al. [25] propose a mid-level image representation called Exemplar Codes and apply it to the problem of tattoo classification. Exemplar codes are feature vectors that consist of normalized outputs of simple linear classifiers. Each linear classifier measures the similarity between the input image and an exemplar, i.e. a training image that best captures some property of the tattoo. Decision score outputs from individual linear classifiers are used to estimate probabilities using extreme value theory [23], thus forming exemplar code feature vectors. A random forest classifier is trained on exemplar codes, enabling multi-class tattoo recognition.

Because of great variability of tattoo designs, individual skin color and lighting conditions in real-world tattoo images, as well as the fact that the tattoos resemble many different real world objects, it is very difficult to devise good hand-crafted features suited for differentiating between tattoos and background [19]. In recent times, however, convolutional neural networks (CNNs) were shown to be able to automatically learn good features for many classification tasks [15]. We therefore propose to apply a deep convolutional neural network to the difficult problem of tattoo detection. In seminal work by Krizhevsky et al. [14], convolutional neural networks were proven to be extremely successful on the ImageNet dataset. According to LeCun et al. [15], this success can be attributed to several factors: efficient use of GPUs for network training, use of rectified linear units, use of dropout regularization and augmenting the training set with deformations of the existing images. Convolutional

networks have already been successfully applied to the problem of scene labelling [7] and semantic segmentation [18].

In contrast to related work, in this paper we take a bottom-up approach. Our CNN-based model operates at the level of small image patches and enables classifying each patch as either belonging to a tattoo or not. Our approach can be used on arbitrary images to obtain a low-level estimate of candidate tattooed regions.

We propose this approach with our target application of de-identification in mind. In a de-identification pipeline, the detected candidate tattoo regions can be removed or averaged to remove personally identifying information. We place much greater importance on correctly detecting all tattooed regions than on eliminating false positive detections, as false positives can be eliminated in subsequent stages, e.g. by combining our method with a person detector (e.g. [5]).

3. Our method

Our proposed method for tattoo detection is based on image patch labeling using a convolutional neural network. We do not detect a tattoo as a global entity. Rather, we use the sliding window approach and at each window position we extract a patch of the size $N \times N$. The patch is then classified as either tattoo or background. The output of our method consists of masked image regions that are tattoo candidates.

Convolutional neural networks typically consist of several convolutional layers, followed by one or more fully connected layers. Convolutional layers are in charge of learning good features and they are characterized by (i) local receptive fields (i.e. the neuron in the convolutional layer is not connected to the outputs of all the neurons from the previous layer, but only to the ones in its local neighborhood), and (ii) shared weights, reflecting the intuition that the features are computed in the same way at different image locations. After the convolutional layers, the so-called pooling layers are typically inserted in order to reduce the dimensionality of feature space for subsequent steps. Fully connected layers perform the task of classification and contain the majority of learned weights.

The architecture of our network is broadly inspired by the successful VGGNet model, proposed in 2014 by Simonyan and Zisserman [24]. The VGGNet is characterized by a very homogeneous architecture that only performs 3×3 convolutions and 2×2 pooling from the beginning to the end. However, our model modifies it to accommodate smaller input images and smaller number of output classes. The simplified network, with fewer and smaller layers is faster to train and it proved adequate for our purposes. The proposed network architecture is shown in Fig. 1.

The input to the network is an $N \times N$ color image (we assumed the RGB color model). The image has to be classified either as belonging to the tattoo or not, depending on whether its center lies inside the polygon that demarcates the tattoo.

The network consists of eight layers (not counting the input layer, i.e. the image itself). The first two layers are convolutional layers with 32 feature maps with 3×3 filters and ReLU activation units. The third layer is a max-pooling layer that reduces the feature map dimensions by 2×2. The fourth and the fifth layers are again convolutional layers with ReLU activation units, but with 64 feature maps (again with 3×3 filters). The sixth layer is another max-pooling layer, once more reducing the input dimension by 2×2. The seventh layer is a fully connected layer consisting of 256 neurons. The final,

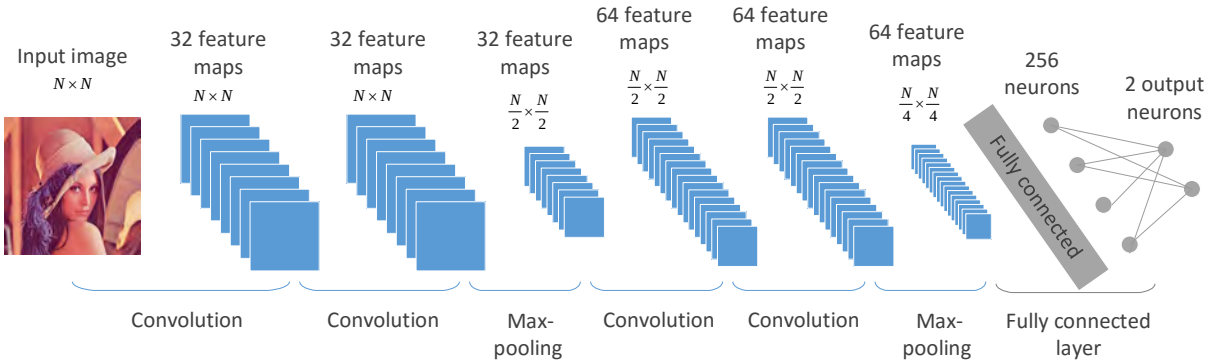


Figure 1: The architecture of the proposed ConvNet model.

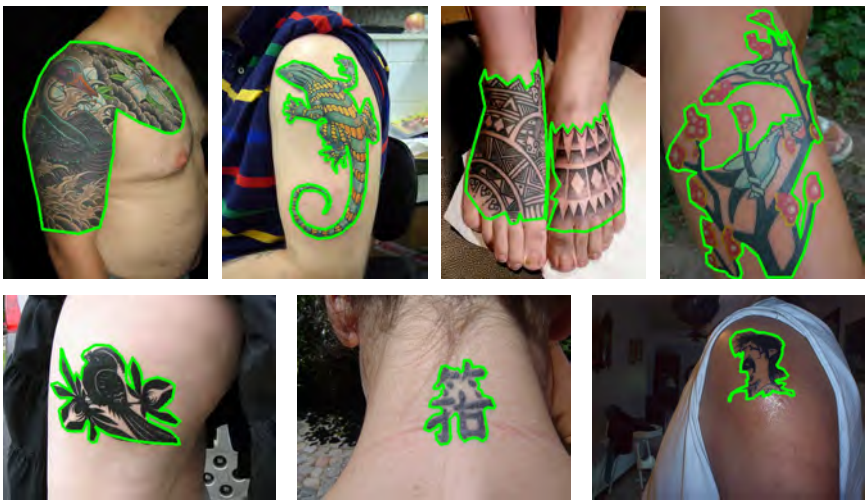


Figure 2: Examples of annotated tattoo images.

eighth layer consists of two neurons with the Softmax activation function, corresponding to the two output classes. Dropout, with the dropout ratio set to 0.5, is applied to the fully connected layer.

We implemented the described network in Python, using Theano [2, 3] and Keras [2] libraries.

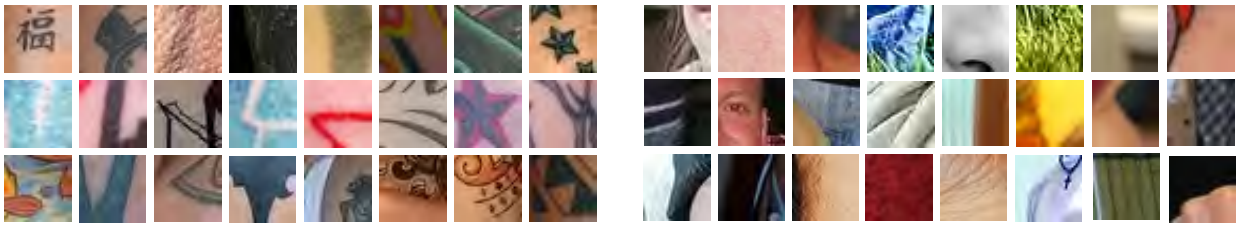
4. Experiments

Given the relatively modest volume of work on tattoo detection, there are no readily available tattoo detection datasets. Recently, a dataset called Tatt-C has been published [19], but it cannot be freely downloaded. Hence, to facilitate the development and testing of our method we have assembled our own dataset[3] by collecting and manually labeling 890 tattoo images from the ImageNet database [22].

Each of the collected images contains one or more tattoos. We annotated each tattoo using a series of connected line segments. Example annotated images from our dataset are shown in Fig. 2. We attempted to closely capture the outline of each tattoo, which can be a challenging task, as tattoos can have highly irregular edges.

[2]https://github.com/fchollet/keras, accessed March 2016.

[3]The dataset is available at http://www.fer.unizg.hr/demsi/databases_and_code/tattoo_dataset.



(a) tattoo patches (b) background patches

Figure 3: Example extracted patches from our dataset (patch size 32×32).

4.1. Training the network

For training, we constructed a training set by randomly sampling a number of image patches of predefined size from each annotated tattoo image in our dataset. This procedure was done both for positive and negative samples, i.e. for patches that do and do not contain tattoos. Examples of extracted patches can be seen in Fig. 3.

The training of the network was carried out by optimizing the mean squared error loss function, using stochastic gradient descent with momentum. We used the mini-batch of 32 samples and the momentum was set to 0.9. The learning rate was set to 0.1. The training was performed for maximaly 40 epochs, with early stopping based on validation loss. The duration of the training varied greatly with the size of the patches, in our case from 10 minutes for smallest patches to 13 hours for the largest.

4.2. Performance evaluation

The set of all extracted patches totalled 22700 images (11359 positive and 11341 negative samples). This set was divided into sets for training (containing 15134 samples, out of which 7573 positive and 7560 negative), and validation and testing (both of the same size of 3783 samples, out of which 1893 positive and 1890 negative). We have ensured that all patches extracted from the same image end up in only one of the sets (either training, validation or testing), in order to avoid mixing training and testing data.

We trained and evaluated the network for different patch sizes (8×8, 12×12, 16×16, 24×24, 32×32 and 48×48) to determine the optimal patch size. The larger patches presumably provide more information about context, but the network that utilizes them is slower to train and test.

The test set was used for evaluation. The results are summarized in Table 1. The accuracy was calculated as a total number of misclassifications (false positives and false negatives) divided by the test set size. As we can see, the results improve in terms of accuracy with the increase in image patch size, up to the largest considered size (48×48) that gives slightly worse results than most of the smaller patch sizes. The difference in accuracy is not very pronounced; i.e. we can say that results for all the patch sizes are similar. The other thing that can be noticed is that the improvement in performance with the increase in patch size comes mainly from reducing the number of false positives, while at the same time the number of false negatives rises.

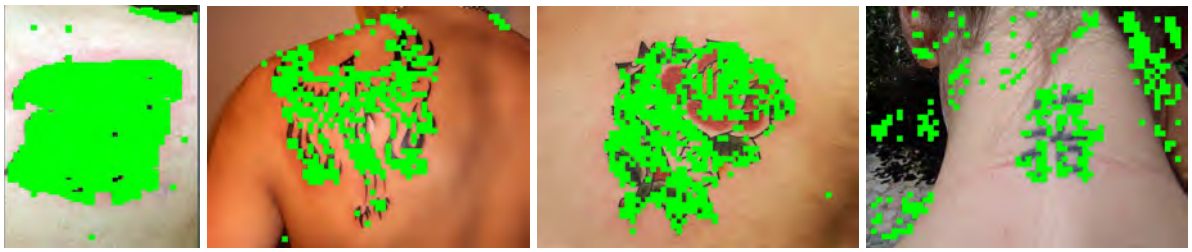
We have done a preliminary qualitative evaluation of the performance of the network in a sliding window setting. Some results are shown in Fig. 4. These examples are relatively simple, with homo-

Patch size	8×8	12×12	16×16	24×24	32×32	48×48
False negatives	152 (8.03%)	229 (12.10%)	187 (9.88%)	213 (11.25%)	248 (13.10%)	290 (15.32%)
False positives	593 (31.37%)	418 (22.12%)	444 (23.49%)	436 (23.07%)	337 (17.83%)	408 (21.59%)
Accuracy	0.8031	0.8290	0.8332	0.8283	0.8454	0.8155

Table 1: Evaluation of the network performance on different patch sizes

geneous skin around the tattoo and simple background. We see that many tattoo patches are correctly detected, but there are also some misclassifications. In more difficult examples with more background containing textured objects, the number of false positives rises. In the context of de-identification, this problem could be addressed by combining this detector with other stages of a de-identification pipeline, e.g. by eliminating detections outside of candidate person locations.

(a) the original images



(b) labeled tattoo patches

Figure 4: The output of the network on full images.

5. Conclusion and outlook

We addressed the challenging problem of tattoo detection for soft biometric de-identification. Instead of hand-crafting image features, we applied deep learning. We trained and evaluated a deep convolutional neural network using the dataset of positive and negative patches generated from a subset of ImageNet tattoo images annotated by hand. Our findings indicate that using a convolutional neural network to classify small image patches can be a reliable way to detect candidate tattoo regions in an image. Patch sizes should be kept small, up to 32×32 patches, in order to obtain best accuracy, good foreground-background segmentation and minimize false negatives.

In our future work, we plan to combine this method with other stages of a de-identification pipeline in order to solve the problem of false positives. As our qualitative analysis shows that the majority of false positives are in the surroundings rather than on the person, one possibility is to run the method only on the outputs of a person detector. We also plan to quantitatively evaluate the performance of our network on full tattoo images (as opposed to patches), and investigate whether this performance could be improved by merging the detections into blobs.

References

[1] D. Baltieri, R. Vezzani, and R. Cucchiara. Mapping appearance descriptors on 3d body models for people re-identification. *International Journal of Computer Vision*, 111(3):345–364, 2014.

[2] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio. Theano: new features and speed improvements. In *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, 2012.

[3] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010.

[4] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann. Into the woods: Visual surveillance of non-cooperative and camouflaged targets in complex outdoor settings. In *Proceedings of the IEEE*, pages 1382–1402, 2001.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.

[6] J. eun Lee, A. K. Jain, and R. Jin. Scars, marks and tattoos (smt): Soft biometric for suspect and victim identification. In *In Proc. Biometric Symposium, Biometric Consortium Conference*, pages 1–8, 2008.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, Aug 2013.

[8] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni. Modeling feature distances by orientation driven classifiers for person re-identification. *Journal of Visual Communication and Image Representation*, 38:115 – 129, 2016.

[9] R. Gross, L. Sweeney, J. F. Cohn, F. De la Torre, and S. Baker. *Protecting Privacy in Video Surveillance*, chapter Face De-identification, pages 129–146. Springer Publishing Company, Incorporated, 2009.

[10] H. Han and A. K. Jain. Tattoo based identification: Sketch to image matching. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8, June 2013.

[11] B. Heflin, W. Scheirer, and T. E. Boult. Detecting and classifying scars, marks, and tattoos found in the wild. In *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*, pages 31–38, Sept 2012.

[12] A. K. Jain, J.-E. Lee, and R. Jin. *Advances in Multimedia Information Processing – PCM 2007: 8th Pacific Rim Conference on Multimedia, Hong Kong, China, December 11-14, 2007. Proceedings*, chapter Tattoo-ID: Automatic Tattoo Image Retrieval for Suspect and Victim Identification, pages 256–265. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[13] J. Kim, A. Parra, J. Yue, H. Li, and E. J. Delp. Robust local and global shape context for tattoo image matching. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 2194–2198, Sept 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 05 2015.

[16] Y. Li, R. Wang, Z. Huang, S. Shan, and X. Chen. Face video retrieval with image query via hashing across euclidean space and riemannian manifold. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[17] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2657–2664, June 2014.

[18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3431–3440, June 2015.

[19] M. Ngan and P. Grother. Tattoo recognition technology - challenge (tatt-c): an open tattoo database for developing tattoo recognition research. In *Identity, Security and Behavior Analysis (ISBA), 2015 IEEE International Conference on*, pages 1–6, March 2015.

[20] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross. Soft biometrics for surveillance: an overview. In *Machine Learning: Theory and Applications*, 31, pages 327–352. Elsevier, 2013.

[21] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, Aug. 2004.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[23] W. Scheirer, A. Rocha, R. Micheals, and T. Boult. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*, chapter Robust Fusion: Extreme Value Theory for Recognition Score Normalization, pages 481–495. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[25] M. J. Wilber, E. Rudd, B. Heflin, Y.-M. Lui, and T. E. Boult. Exemplar codes for facial attributes and tattoo recognition. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 205–212, March 2014.

[26] G. Ye, W. Liao, J. Dong, D. Zeng, and H. Zhong. *MultiMedia Modeling: 21st International Conference, MMM 2015, Sydney, NSW, Australia, January 5-7, 2015, Proceedings, Part II*, chapter A Surveillance Video Index and Browsing System Based on Object Flags and Video Synopsis, pages 311–314. Springer International Publishing, Cham, 2015.

WS 2: Signal & Image Processing / Filters

3-D Shape Recovery of the Left Heart Chamber from Biplane X-Ray Projections Using Anatomical A-Priori Information Learned from CT

Roland Swoboda[1,2], Josef Scharinger[2], and Clemens Steinwender[3]

[1] Research Center Hagenberg
University of Applied Sciences Upper Austria, Austria
roland.swoboda@fh-hagenberg.at

[2] Department of Computational Perception
Johannes Kepler University Linz, Austria
josef.scharinger@jku.at

[3] Clinic of Internal Medicine I
General Hospital Linz, Austria
clemens.steinwender@akh.linz.at

Abstract

Recovering the 3-D shape of the left heart chamber from bi-planar 2-D x-ray projection images is a challenging task since only sparse and noisy data is available for reconstruction. In this work, a 3-D statistical shape model (SSM) of the left ventricular (LV) anatomy is learned from high-resolution CT data and utilized as a-priori information to solve the under-determined and ambiguous reconstruction problem. A 2-D/3-D registration method fits the SSM to the x-ray images of the patient by calculating simulated projections of the SSM and minimizing the difference between simulated and given projections. The presented approach is evaluated using simulated and real patient data. For patients where both projection images and CT data are available, the reconstructed LV is compared to the true shape known from CT. Our results show a good correspondence between recovered and true shapes. Using a SSM as anatomical a-priori information for reconstruction helps in limiting the space of possible solutions and allows to generate statistically plausible shapes.

1. Introduction

Cardiac diseases are one of the most common causes of death in the industrialized world today. In the case of acute myocardial infarction, for instance, interventional x-ray angiography is state-of-the-art for both treatment and diagnosis. To evaluate the viability of myocardium after infarction, a catheter is advanced into the left heart chamber (ventricle) and contrast agent is injected to opacify the ventricular cavity during radiation. Bi-planar cine-angiographic equipment is used to acquire two x-ray image sequences simultaneously from standard right anterior oblique (RAO) and left anterior oblique (LAO) views, see Fig. 1.

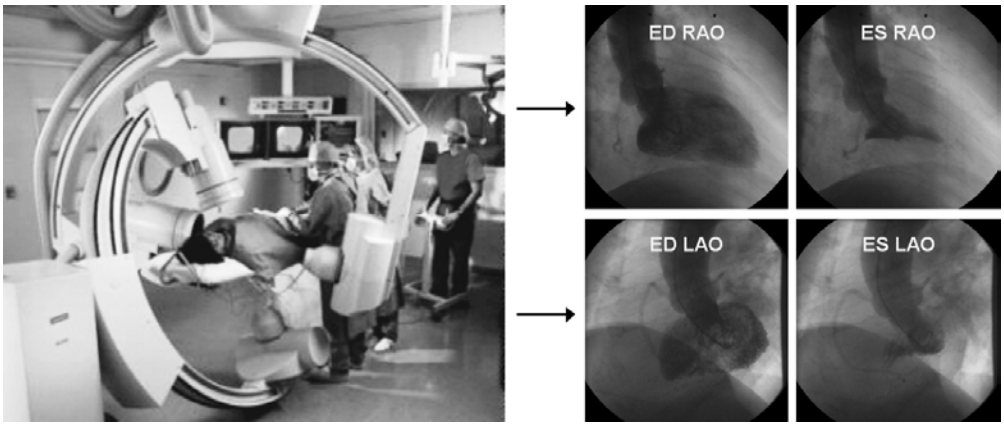


Figure 1. Biplane cine-angiographic x-ray equipment used in the catheter lab to acquire images for quantitative left ventricle analysis.

The gold standard for quantitative left ventricle analysis in the catheter lab is based on the evaluation of end-diastolic (ED) and end-systolic (ES) endocardial contour information gathered from these 2-D projection images. The ED and the ES volume are calculated (by applying e.g. the Area-Length method) and used to determine ejection fraction (EF), i.e. the volume that is squeezed out during contraction. Contour information is further utilized by wall motion analysis methods (like e.g. the Centerline method) to quantify myocardial viability. However, since 3-D information is lost due to projection, volumetric diagnostic parameters, like EF, can only be approximated and wall motion is only evaluable for LV surface areas with the boundary visible in the projection image. Novel approaches aim at reconstructing the spatio-temporal shape of the LV to perform analysis in 3-D [10].

2. Related Work

In classical computed tomography (CT), hundreds of projections are acquired by a fast rotating x-ray gantry. Analytical and algebraic reconstruction techniques exploit this dense information to yield voxel values that vary within a continuous range. However, these techniques typically fail if merely two (noisy) projections are available. C-arm CT is a relatively young and hybrid type of imaging modality, where the C-arm is rotated during acquisition to increase the number of projections. Techniques known from CT can then be utilized to address the reconstruction problem [8]. In the catheter lab, however, the application of C-arm CT is challenged by the higher amount of x-ray dose and bolus compared with conventional x-ray angiography (XA), and the slower rotational speed of the C-arm compared with classical CT when imaging the rapidly moving heart. Whether C-arm CT will substitute XA as a routine method in future remains to be seen [9].

Unlike classical (continuous) CT, discrete tomography focuses on reconstruction problems where only a small number of projections – as small as two – are available and the object's intensity levels are limited, i.e. discrete, and known a-priori [3]. Using additional a-priori information is crucial when trying to solve such under-determined and ambiguous problems, since this can reduce the space of possible solutions and improve the ability to deal with noisy projection data. Some of the early approaches published in the field of 3-D LV shape recovery from XA rely on the assumption that ventricular cross-sections follow certain geometric priors (like connectedness, convexity, symmetry, roundness, etc.), however, this is usually too restrictive in practice. In the work of Prause and Onnasch [7], digitized post-mortem human LV casts are used as a-priori information. Other approaches often do not incorporate anatomical a-priori information at all [5], [6].

The novelty of our approach is that anatomical a-priori information is learned from high-resolution CT data and modeled as a SSM, which is then fit to the angiograms by a 2-D/3-D registration method. The application of SSMs for recovering shape from angiography has been successfully demonstrated by other authors for hard-tissue objects like the pelvis [4] or the vertebrae [1], but not yet for non-rigid contrast-enhanced soft-tissue objects like the LV. This paper is a refinement of our previous work [12]. For the sake of comprehensibility, parts of Sec. 3 and 4 are based thereon.

3. Methods

3.1. Statistical Shape Models

In order to build a 3-D SSM [2], a set of segmentations of the target shape is required. The contour of each shape S_i is described by n landmarks, i.e. points of correspondence that match between shapes, and represented as a vector of coordinates: $x_i = (x_1, ..., x_n, y_1, ..., y_n, z_1, ..., z_n)_i^T$. All n_s shape vectors form a distribution in a $3n$-dimensional space. This distribution is approximated by $x = \bar{x} + \Phi b$, with $\bar{x} = \frac{1}{n_s}\sum_{i=1}^{n_s} x_i$ being the mean shape vector and b being the shape parameter vector. By varying b, new instances of the shape class are generated. Φ is obtained by performing a principal component analysis (PCA) on the covariance matrix $C = \frac{1}{n_s-1}\sum_{i=1}^{n_s}(x_i - \bar{x})(x_i - \bar{x})^T$. PCA yields the principal axes of this distribution; the eigenvalues give the variances of the data in the direction of the axes (= eigenvectors). To reduce noise and dimensionality only those eigenvectors with the largest t eigenvalues are used. t denotes the number of the most significant modes of variation (MOV) and is chosen so that a fraction f of the total variation is retained, $\sum_{j=1}^{t}\lambda_j \geq f\sum\lambda_j$. Prior to statistical analysis, location, scale and rotational effects must be removed from the training shapes to obtain a compact model. Commonly, Procrustes analysis is applied to minimize $D = \sum|x_i - \bar{x}|^2$, the sum of squared distances (SSD) of each shape to the mean.

3.2. Modeling of Anatomical A-Priori Information

A Siemens Somatom Sensation Cardiac 64 multi-slice CT is used to acquire 20 data sets at 65% of the heart phase (R-R peaks) with an effective slice thickness of 0.5 mm and an average in-plane resolution of 0.33 mm. The size of the image mask in the transversal plane is 512×512 pixels; the number of slices varies between 220 and 310. The endocardial LV surface is manually segmented by experts in cardiology. Contours are specified in each fifth axial slice by interactively setting control points of a cardinal spline; intermediate contours are interpolated. The surface of an LV is represented as a stack of contours. Details like the atrial concavity, the apex and the aortic valve region are retained during segmentation to obtain an accurate model of the anatomy. Point correspondence among the training shapes is established based on back-propagation of the landmarks on a mean shape [11]. After segmentation, landmark extraction and removing location, scale and rotational effects, the SSM is built as outlined in Sec. 3.1. The first three MOV of the final model are illustrated in Fig. 2.

3.3. Left Ventricular Shape Recovery

In discrete tomography, a common strategy for solving the under-determined and ambiguous reconstruction problem is to use numeric optimization [3]. As an exact solution will usually not be available, the projections of the recovered object need only be approximately equal to the given projection data. In this work, a 2-D/3-D registration approach is followed to minimize the difference between the given projections and the simulated projections derived from the SSM. To transform the SSM from

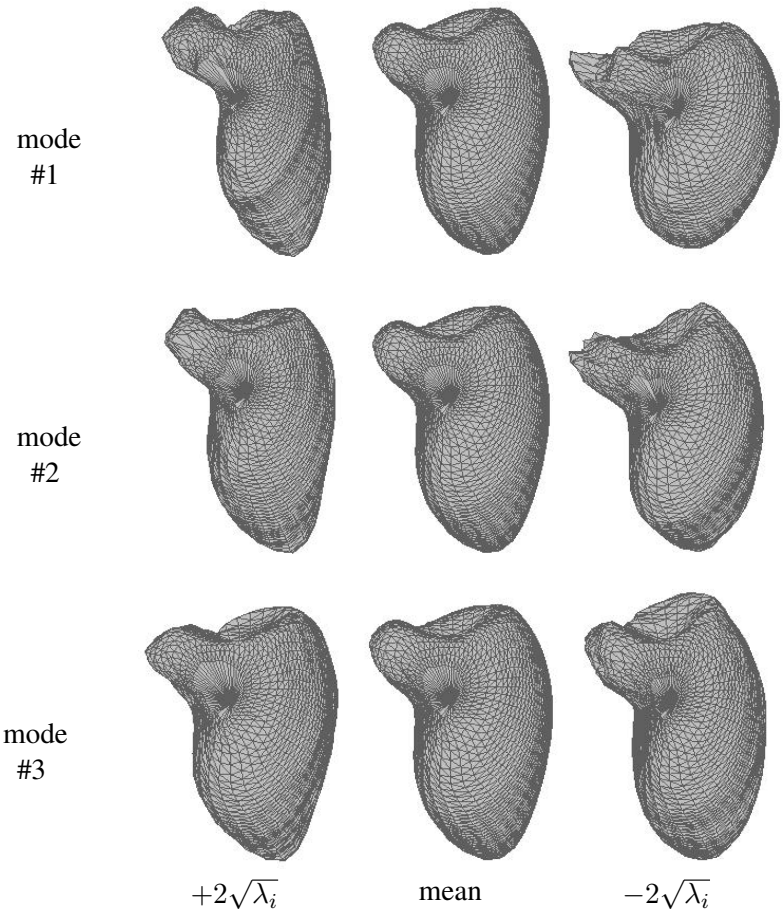


Figure 2. First three modes of variation of the LV SSM.

model space to image space the following equation is used: $y = R\left((\bar{x} + \Phi b)\, s + T\right)$. Both shape parameter vector b and the parameters for pose $p = \{R, s, T\}$, i.e. rotation matrix R, scale factor s and translation vector T, have to be found so that the registration error is minimized. Unlike [4] and [1], we derive R from Euler angles to reduce the dimensionality of the registration problem. Orientation in 3-D space is thus described using 3 angles, i.e. $R_{\alpha,\beta,\gamma}$, instead of a 3×3 matrix. To generate statistically plausible shapes [2], b is constrained by $\pm 2\sqrt{\lambda_i}$. In contrast to [4] and [1], we exploit the training data to derive constraints for p. The training instances in model space are transformed to image space and the range of the pose vector components is analyzed. Note that this can be regarded as additional a-priori information. To minimize our cost function, the Nelder-Mead algorithm is applied. Experiments showed that optimizing pose and shape sequentially is more efficient than optimizing both simultaneously.

3.3.1. Cost Function

Our cost function depends on the shape and the pose parameter vector and incorporates both contour and densitometric information derived from the given projections P_i and the simulated projections $P_i'(b,p)$: $\epsilon(b,p) = \sum_{i=1}^{n_P} \left(\omega_C \epsilon_C(P_i, P_i'(b,p)) + \omega_D \epsilon_D(P_i, P_i'(b,p))\right)$. Contour-related error ϵ_C is obtained by equiangular sampling of the given and the simulated contour and by calculating the SSD for the sampled points. As density-related error ϵ_D, the sum of squared difference metric is used. Total error ϵ is defined as the weighted sum of ϵ_C and ϵ_D over all $n_P = 2$ projections.

3.3.2. Extraction of Contour and Densitometric Information

In the case of in-vivo angiograms, the endocardial contour is segmented by experts in cardiology prior to reconstruction. Densitometric information is derived by means of digital subtraction angiography. From the initial frames of an angiographic sequence showing no contrast agent, a mask is deduced. Logarithmic subtraction of mask and current frame is performed due to the exponential attenuation of x-rays. To reduce noise and the inhomogeneous saturation of contrast agent within the ventricle, two frames before and after a frame are used for averaging. In the case of simulated angiograms, contour information is extracted by border detection, whereas densitometric information is measured directly.

3.4. Simulation of Angiographic Projections

Both the presented reconstruction approach and the following evaluation strategy require the simulation of projections. Our model of the bi-planar angiographic device calculates the exact position of the x-ray sources and the image intensifier planes for the projections. For a given viewing direction, shape and pose parameter vector, a simulated projection of the SSM in image space is obtained in two steps. First, the polygonal model is converted into a 3-D binary image, V, whose values denote the presence/absence of contrast agent. Then, a projection is derived using ray-casting. Since densitometric information is expected to be linear for reconstruction, an exponential attenuation of x-rays has not been incorporated into the simulation process.

4. Results

The presented methods are implemented and evaluated using Matlab and the Image Segmentation and Registration Toolkit (ITK) C++ library. To quantify the difference between original and recovered shape, two geometric and three volumetric similarity metrics are defined for comparing the polygonal models and the binary image representations, respectively. An exemplary reconstruction result of the performed leave-one-out experiments is illustrated in Fig. 3.

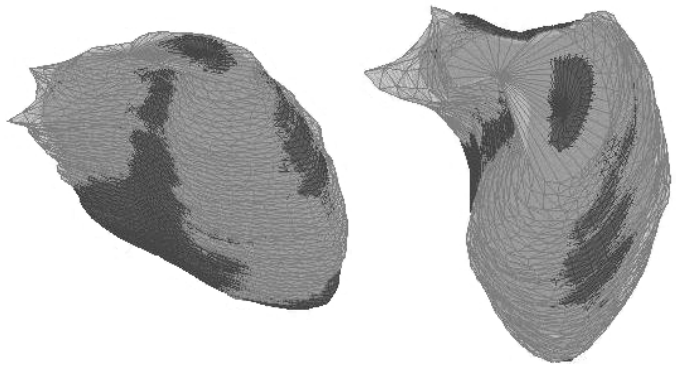


Figure 3. Reconstruction example showing original shape (bright) and recovered shape (dark).

4.1. Similarity Metrics

Similarity of two polygonal models S_1 and S_2 is measured based on a given distance metric d: $sim_d(S_1, S_2) = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n} d(p_i, S_2) + \frac{1}{m}\sum_{j=1}^{m} d(q_j, S_1))$, $p_{i=1,\ldots,n} \in S_1, q_{j=1,\ldots,m} \in S_2$. Distance metric d_{min} is defined as the Euclidean distance between point p_i and its closest point on S_2: $d_{min}(p_i, S_2) =$

$\min_{q_j \in S_2} |p_i - q_j|$. Distance metric d_{ortho} denotes the Euclidean distance between p_i and the point obtained by intersecting S_2 with the surface normal at p_i: $d_{ortho}(p_i, S_2) = |p_i - \mathrm{surfn}(p_i) \cap S_2|$.

Let $|V|$ denote the volume of a 3-D binary image V. Volume conformity is measured by calculating the difference of volumes (DOV): $sim_{DOV} = 1 - abs(|V_{orig}| - |V_{rec}|)/|V_{orig}|$. To assess shape conformity, the volume of differences (VOD) metric is used: $sim_{VOD} = 1 - |\mathrm{xor}(V_{orig}, V_{rec})|/|V_{orig}|$. An alternative metric for shape conformity, derived from kappa statistic, quantifies the overlap between two binary masks: $sim_\kappa = 2|V_1 \cup V_2|/(|V_1| + |V_2|)$.

4.2. Evaluation based on Simulated Data

Evaluation with simulated data is performed based on leave-one-out experiments. From the 20 segmented CT data sets, all but one are used to learn a SSM. Simulated angiograms from RAO and LAO view are calculated for the left-out data set as described in Sec. 3.4, and from these angiograms shape is recovered by fitting the learned SSM. The recovered shape is compared with the segmented shape of the left-out data set using the defined similarity metrics. This procedure is repeated for each data set. The DOV metric in Tab. 1 shows that the original volume is approximated at high accuracy. This is essential for assessing volume-based diagnostic parameters, like EF. Concerning shape conformity we can see that a high overlap between the two shapes is achieved, although the VOD is still improvable. The distance metrics d_{min} and d_{ortho} are near the mean reconstruction error of 2.3 mm [11].

Sim. Metric	Mean	Std.	Min.	Max.
d_{min} (mm)	2.61	0.65	1.65	3.53
d_{ortho} (mm)	2.49	0.77	1.38	3.72
DOV (%)	94.56	3.55	87.35	98.73
VOD (%)	78.17	5.30	68.88	84.91
κ (%)	87.12	2.53	82.54	90.18

Table 1. Evaluation of LV shape recovery from simulated angiograms.

4.3. Evaluation based on Real Patient Data

For three patients, a corresponding CT image is available for the RAO/LAO in-vivo angiograms. Note that this allows an accurate evaluation of our approach since the true 3-D LV shape is exactly known from CT. Evaluation based on the three in-vivo angiograms is performed as follows: 1) a SSM is learned from 19 of the 20 data sets, with the CT data set corresponding to the angiograms being excluded, 2) the model is fit to interpolated angiographic RAO/LAO frames of a single cardiac cycle showing the LV at 65% of the heart phase, and 3) the recovered shape is compared with the true 3-D shape of the excluded CT data set using the defined similarity metrics. The angiograms are acquired using a Siemens Bicor and a Siemens AXIOM Artis dBC system, capturing images of 512×512 pixels and 8-bit gray level depth at a frame-rate of 25 fps. For temporal registration with CT data in step 2, the ECG information accompanying the angiograms is utilized. The results for three in-vivo angiograms are given in Tab. 2. Our experiments indicate that values similar to the evaluation with simulated data are achieved, although the number of data sets is relatively small. The best shape conformity is achieved for example #2. For example #3, the reconstruction yields suboptimal results.

Sim. Metric	#1	#2	#3	Mean	Std.
d_{min} (mm)	2.43	2.32	2.95	2.57	0.34
d_{ortho} (mm)	2.36	2.05	3.36	2.59	0.68
DOV (%)	98.01	92.87	82.11	91.00	8.11
VOD (%)	74.72	80.13	68.12	74.32	6.01
κ (%)	87.49	90.41	79.75	85.88	5.51

Table 2. Evaluation of LV shape recovery from three in-vivo angiograms.

5. Discussion and Conclusion

In this work, a new method for recovering the LV from contrast-enhanced bi-planar cine-angiographic x-ray images has been proposed. The novelty of our approach is that a-priori information about the LV anatomy is learned from high-resolution CT images, modeled as a SSM and utilized for reconstruction. A 2-D/3-D registration technique is applied to fit the SSM to angiographic projections.

When only two (noisy) projections are available, the reconstruction problem usually becomes underdetermined and ambiguous. In such cases, the incorporation of a-priori information plays an important role, since this can limit the space of possible solutions and improve the ability to deal with noisy data. In contrast to [7], anatomical a-priori information is derived from data of in-vivo instead of post-mortem subjects; other approaches often do not utilize this kind of information at all. Although only one bi-planar acquisition is used for reconstruction, our approach is generally not limited by the number of projections. However, since additional acquisitions increase the amount of radiation and bolus, this number is usually kept to a minimum.

Using a SSM for reconstruction allows to generate statistically plausible and patient specific shapes. Unlike other 3-D LV SSMs often found in literature, anatomical areas like the apex, the atrial concavity and the aortic valve region are preserved in our model. This is necessary to generate complete contour and densitometric information; otherwise, additional errors are introduced in the reconstruction process. Further note that these areas typically overlap with the ventricular cavity in projection images and are therefore hard to recover without prior knowledge.

Evaluation with both simulated data and real patient data shows promising results. The LV volume is recovered at high accuracy. This is important for assessing volumetric diagnosis parameters, like EF. Concerning shape conformity, the overlap between original and recovered volume is high, though there is still place for minor improvements. Future work will focus on improving the model fitting process and on evaluating our approach with more in-vivo angiograms.

References

[1] S. Benameur, M. Mignotte, S. Parent, H. Labelle, W. Skalli, and J. de Guise. 3D/2D Registration and Segmentation of Scoliotic Vertebrae using Statistical Models. *Computerized Medical Imaging and Graphics*, 27:321–337, 2003.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active Shape Models - Their Training and Application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[3] G. T. Herman and A. Kuba. Discrete Tomography in Medical Imaging. *Proceedings of the IEEE*, 91:1612–1626, 2003.

[4] H. Lamecker, T. H. Wenckebach, and H.-C. Hege. Atlas-Based 3D-Shape Reconstruction from X-Ray Images. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 371–374. IEEE Computer Society, 2006.

[5] R. Medina, M. Garreau, J. Toro, H.L. Breton, J.-L. Coatrieux, and D. Jugo. Markov Random Field Modeling for Three-Dimensional Reconstruction of the Left Ventricle in Cardiac Angiography. *IEEE Transactions on Medical Imaging*, 25:1087–1100, 2006.

[6] M. Moriyama, Y. Sato, H. Naito, M. Hanayama, T. Ueguchi, T. Harada, F. Yoshimoto, and S. Tamura. Reconstruction of Time-Varying 3-D Left-Ventricular Shape from Multiview X-Ray Cineangiocardiograms. *IEEE Transactions on Medical Imaging*, 21:773–785, 2002.

[7] G. P. M. Prause and D. G. W. Onnasch. Binary Reconstruction of the Heart Chambers from Biplane Angiographic Image Sequences. *IEEE Transactions on Medical Imaging*, 15:532–546, 1996.

[8] M. Prummer, J. Hornegger, G. Lauritsch, L. Wigstrom, E. Girard-Hughes, and R. Fahrig. Cardiac C-Arm CT: A Unified Framework for Motion Estimation and Dynamic CT. *IEEE Transactions on Medical Imaging*, 28:1836–1849, 2009.

[9] J. Rieber, C. Rohkohl, G. Lauritsch, H. Rittger, and O. Meissner. Application of C-Arm Computed Tomography in Cardiology. *Der Radiologe*, 49:862–867, 2009.

[10] R. Swoboda, M. Carpella, W. Backfrieder, C. Steinwender, C. Gabriel, and F. Leisch. From 2D to 4D in Quantitative Left Ventricle Wall Motion Analysis of Biplanar X-Ray Angiograms. In *Computers in Cardiology 2005*, pages 977–980. IEEE Computer Society Press, 2005.

[11] R. Swoboda and J. Scharinger. A 3-D Statistical Shape Model of the Left Ventricle - Geometric Prior Information for Recovering Shape from Projective Bi-Planar X-Ray Images. In *Challenges in Biosciences: Image Analysis and Pattern Recognition Aspects*, pages 53–62. books@ocg.at, 2008.

[12] R. Swoboda, J. Scharinger, and C. Steinwender. Model-Based 3-D LV Shape Recovery in Biplane X-Ray Angiography: A-Priori Information Learned from CT. In *2015 Computing in Cardiology Conference (CinC)*, pages 101–104, 2015.

Robust Blind Deconvolution Using Convolution Spectra of Images

Philipp Moser[1] and Martin Welk[1]

Biomedical Image Analysis Division
University for Health Sciences, Medical Informatics and Technology (UMIT),
Hall/Tyrol, Austria
philipp.moser@hotmail.com, martin.welk@umit.at

Abstract
We present a method for blind image deconvolution that acts by alternating optimisation of image and point-spread function. The approach modifies a variational model recently published by Liu et al. which combines a quadratic data term with a total variation regulariser for the image and a regulariser for the point-spread function that is constructed from convolution eigenvalues and eigenvectors of the blurred input image. We replace the image estimation component with a robust modification of Richardson-Lucy deconvolution, and introduce a robust data term into the point-spread function estimation. We present experiments on images with synthetic and real-world blur that indicate that the modified method has advantages in the reconstruction of fine image details.

1. Introduction

Blur is, second to noise, one of the major sources of degradations in digital images. Its removal has therefore been a subject of intense investigation since the beginnings of digital image processing. If for each location the intensity is smeared across a neighbourhood in the same way, this *spatially invariant* blur is mathematically modelled by a convolution with a kernel, called *point-spread function (PSF)*. Describing the observed blurred image f, the unobservable sharp image u and the PSF h as functions from suitable function spaces, and assuming additive noise n, one has

$$f = u * h + n \; . \tag{1}$$

Spatially variant blur can be modelled similarly using Fredholm integral operators. The mathematical problem of approximate inversion of these blur operations is termed deconvolution.

In this paper, we focus on the spatially invariant case. In *non-blind* deconvolution problems, both the blurred image f and the PSF h are available as input; in contrast, *blind* deconvolution aims at recovering the PSF h along with the sharp image u from the input image f. Both kinds of problems are severely ill-posed; in particular, deconvolution algorithms are highly sensitive to noise.

Non-blind deconvolution can nowadays be performed efficiently with favourable quality, with methods ranging from the time-proven Wiener filter [15] and Richardson-Lucy deconvolution [6, 10] up to the performant iterative algorithm by Krishnan and Fergus [4], to name a few representatives. Recently also neural network techniques have been used, see [12, 17].

For blind deconvolution, a straightforward approach proceeds in two steps: first, estimating the PSF,

and second, performing non-blind deconvolution with that PSF, see e.g. [3, 16]. Merging both steps, u and h can be estimated simultaneously by minimising a joint energy functional such as

$$E[u,h] := \int_\Omega (f - u * h)^2 \,\mathrm{d}\boldsymbol{x} + \alpha R_u[u] + \beta R_h[h] \;, \tag{2}$$

see e.g. [2, 11, 18], which combines the *data term* that integrates the squared model error $(f - u * h)^2$ over the image domain Ω with regularisation functionals R_u and R_h for the image and PSF, respectively, using regularisation weights α, β. Note that there is a formal symmetry between u and h in the data term, coming from the blur model (1); however, this symmetry usually does not extend to R_u and R_h – regularisers that work well for images do generally not perform favourably in PSF estimation, and vice versa. This is because the regularisers express model requirements for sharp images and for PSFs, respectively, and these model requirements differ substantially. For example, sharp edges are important for u which makes total variation regularisers a good candidate, whereas for h rather locality and sparse support may be meaningful requirements.

Motivated by the separation of regularisers in (2), one often separates u and h again in the minimisation, by using iterative methods that alternatingly update u and h. Each cycle then comprises an image estimation step, which is a non-blind deconvolution, and a PSF estimation step. Whereas the latter can formally be considered as non-blind deconvolution of the blurred image with respect to the sharpened image as convolution kernel, the dissimilarity of regularisers in fact often implies that substantially different algorithms have to be used for image and PSF estimation.

A refinement of (2) results from applying to the squared model error $(f - u * h)^2$ a function Φ with less-than-linear growth, yielding a sub-quadratic data term $\int_\Omega \Phi((f - u * h)^2) \,\mathrm{d}\boldsymbol{x}$. Data terms of this kind have been proven useful in various image processing tasks in order to reduce sensitivity to (particularly, heavy-tailed) noise and measurement errors as well as to minor deviations from the data model, and are therefore known as *robust data terms*, see e.g. [1, 19] in the deconvolution context. A similar modification of the objective function underlying the Richardson-Lucy deconvolution (the information divergence) has been introduced in [13], leading to a non-blind deconvolution method called robust and regularised Richardson-Lucy deconvolution (RRRL).

Our contribution. In this paper, important parts of which are based on the thesis [7], we review a recent blind deconvolution approach from [5] that is based on alternating minimisation of an energy in the sense of (2) with a PSF regulariser constructed from so-called convolution eigenvalues and eigenvectors. We then modify both the PSF and image estimation components of this approach by using robust data terms. To this end, we adopt in the image estimation component the RRRL method from [13]; regarding the PSF estimation component, we introduce a subquadratic data term. The so modified PSF estimation component has to the best of our knowledge not been studied before. We present experiments on a proof-of-concept level that support the conclusion that our modified method achieves higher reconstruction quality than its predecessor.

2. PSF Estimation Using Spectra of Convolution Operators

In this section, we review the approach from [5] which forms the basis for our further work in this paper. As the construction of the regulariser R_h from [5] relies on spectral decompositions formulated in matrix language, we switch our notations to use discrete images from here on.

Given a unsharp discrete grey-value image $\boldsymbol{f} = (f_{i,j})_{i,j}$, the sharp image $\boldsymbol{u}$ and the PSF $\boldsymbol{h}$ are sought as minimisers of the function (a discrete version of the functional (2))

$$E(\boldsymbol{u}, \boldsymbol{h}) := \sum_{i,j} (f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}]_{i,j})^2 + \alpha R_u(\boldsymbol{u}) + \beta R_h(\boldsymbol{h}) \tag{3}$$

where in the discretised data term $[\boldsymbol{u} * \boldsymbol{h}]_{i,j}$ denotes the sampling value of the discrete convolution $\boldsymbol{u} * \boldsymbol{h}$ at location (i, j), and the regularisers R_u, R_h are still to be specified.

For the image, [5] use a total variation regulariser, which is common in literature, and known to produce favourable results in non-blind deconvolution. In discretised form it reads as $R_u(\boldsymbol{u}) = \sum_{i,j} \|[\boldsymbol{\nabla u}]_{i,j}\|$ where $[\boldsymbol{\nabla u}]_{i,j}$ denotes a discretisation of $\boldsymbol{\nabla} u$ at location (i, j). The central innovation of [5] lies in the PSF regulariser R_h which is built from *convolution eigenvalues and eigenvectors,* i.e. singular values and singular vectors of a linear operator associated with the image $\boldsymbol{f}$.

Note first that any discrete image $\boldsymbol{w}$, acting by convolution $\boldsymbol{w} * \boldsymbol{h}$ on the PSF, yields a linear operator on $\boldsymbol{h}$. In the discrete setting, we assume that the support of $\boldsymbol{h} = (h_{i,j})_{i,j}$ has size $m_x \times m_y$, which is embedded in a larger area $s_x \times s_y$ ([5] suggests $s_{x,y} \approx 1.5\, m_{x,y}$), and the discrete image $\boldsymbol{w}$ is of size $n_x \times n_y$. By discrete convolution with zero-padding, one has a linear operator $\boldsymbol{A}^{\boldsymbol{w}}_{s_x,s_y} : \boldsymbol{h} \mapsto \boldsymbol{w} * \boldsymbol{h}$ mapping $\mathbb{R}^{s_x \times s_y}$ to $\mathbb{R}^{(s_x+n_x-1)\times(s_y+n_y-1)}$ which has $s_x s_y$ (right) singular values $\sigma_k(\boldsymbol{w})$ with singular vectors $\boldsymbol{v}_k(\boldsymbol{w}) \in \mathbb{R}^{s_x \times s_y}$, which are called the convolution eigenvalues and eigenvectors of $\boldsymbol{w}$.

Theoretical analysis in [5] has brought out that, for meaningful convolution kernels $\boldsymbol{h}$, the convolution eigenvalues of $\boldsymbol{w} * \boldsymbol{h}$ are significantly smaller than those of $\boldsymbol{w}$; moreover, it is shown in [5] that particularly the convolution eigenvectors with smallest convolution eigenvalues of $\boldsymbol{w} * \boldsymbol{h}$ are almost convolution-orthogonal to $\boldsymbol{h}$, i.e. $\|\boldsymbol{v}_k(\boldsymbol{w} * \boldsymbol{h}) * \boldsymbol{h}\|$ is almost zero for those k for which $\sigma_k(\boldsymbol{w})$ is small enough. This motivates that for a given blurred image $\boldsymbol{f}$ the underlying PSF h can be sought as a minimiser of $\sum_{k=1}^{m_x m_y} \|\boldsymbol{v}_k(\boldsymbol{f}) * \boldsymbol{h}\|^2 / \sigma_k(\boldsymbol{f})^2$ where placing the squared convolution eigenvalues in the denominator ensures the higher influence of the convolution eigenvectors with smallest eigenvalues, and avoids introducing a threshold parameter to single out the "small" convolution eigenvalues.

An additional degree of freedom in the procedure is that the image $\boldsymbol{f}$ can be preprocessed by some linear filter L. Since convolution itself is a linear filter, and therefore commutes with any linear filter L, the above reasoning about singular values remains valid in this case; at the same time, a suitable choice of L allows to reweight the influence of different parts of $\boldsymbol{f}$ on the PSF estimation. Based on the well-known fact from literature, see e.g. [16], that edge regions are particularly well-suited to estimate blur, [5] suggest the use of a Laplacean-of-Gaussian (LoG) filter, thus leading to the final formulation of the objective function

$$R_h(\boldsymbol{h}) := \sum_{k=1}^{s_x s_y} \frac{\|\boldsymbol{v}_k(L(\boldsymbol{f})) * \boldsymbol{h}\|^2}{\sigma_k(L(\boldsymbol{f}))^2} \tag{4}$$

where L is a LoG operator. Whereas the extended support size $s_x \times s_y$ for $\boldsymbol{h}$ is used in R_h, its minimisation is constrained to PSF $\boldsymbol{h}$ of the actual support size $m_x \times m_y$.

Using R_h alone as objective function would already allow to estimate the PSF fairly accurate. However, as discussed in [5] such a proceeding tends toward some over-sharpening of the image with visible artifacts. In order to achieve a good joint reconstruction of the sharp image and PSF that also takes into account regularity constraints on the image expressed by R_u, and improves the treatment of images with moderate noise, [5] insert R_h instead as PSF regulariser into (3).

This joint energy functional is then minimised by an alternating minimisation. In the PSF estimation step, $R_h(\boldsymbol{h})$ is represented as a quadratic form, $R_h(\boldsymbol{h}) = \sum_{i,j,i',j'} H_{i,j;i',j'} h_{i,j} h_{i',j'}$, with the coefficient matrix (Hessian) $\boldsymbol{H} = (H_{i,j;i',j'})_{i,j,i',j'}$ given by

$$\boldsymbol{H} = \sum_{k=1}^{s_x s_y} \frac{\boldsymbol{A}^{\boldsymbol{v}_k(L(\boldsymbol{f}))\mathrm{T}}_{m_x,m_y} \boldsymbol{A}^{\boldsymbol{v}_k(L(\boldsymbol{f}))}_{m_x,m_y}}{\sigma_k(L(\boldsymbol{f}))^2} \ , \tag{5}$$

and this is combined with the data term from (3) to establish a quadratic minimisation problem for $\boldsymbol{h}$.

In our re-implementation of the PSF estimation from [5], this quadratic minimisation problem is solved via the corresponding linear equation system and an LU decomposition [9, pp. 52p.], followed by a projection step that eliminates negative entries in $\boldsymbol{h}$ and normalises $\boldsymbol{h}$ to unit total weight. As a refinement of the projection step, it turned out useful to cut off even small positive entries in $\boldsymbol{h}$ by a threshold, thus additionally enforces sparsity of the PSF. Experiments indicate that the threshold is best adapted as a multiple of some quantile, e.g. 0.1 times the $95\,\%$-quantile of the entries of $\boldsymbol{h}$.

The image estimation step that alternates with PSF estimation comes down to a TV-regularised non-blind deconvolution problem for which several approaches exist. In [5] the method from [4] is used.

3. Robust Image and PSF Estimation

As demonstrated in e.g. [1, 8, 14, 13], robust data terms allow to achieve favourable deconvolution results even with imprecise estimates of the PSF or slight deviations from the spatial invariant blur model. While the latter is generally relevant in deconvolution of real-world images, robustness to imprecise PSF estimates is particularly useful in blind deconvolution. This makes it attractive to incorporate robust data terms into the framework of [5], which is our goal in this section.

Due to the alternating minimisation structure of the method, we consider the two steps separately. We start with the image estimation, which is tantamount to non-blind deconvolution. Thus, we simply have to replace the TV deconvolution model with a suitable robust approach. In this work, we choose RRRL [13] for this purpose, which is a fixed point iteration associated to the energy function

$$E(\boldsymbol{u}) = \sum_{i,j} \Phi\left([\boldsymbol{u} * \boldsymbol{h}]_{i,j} - f_{i,j} - f_{i,j} \ln \frac{[\boldsymbol{u} * \boldsymbol{h}]_{i,j}}{f_{i,j}} \right) \mathrm{d}\boldsymbol{x} + \alpha R_u(\boldsymbol{u}) \ . \tag{6}$$

We prefer this method for efficiency reasons; note that the non-blind deconvolution step is needed in each iteration of the alternating minimisation. RRRL is known to evolve fast toward a good solution during the first few iterations, see also [14], whereas methods based on approaches as in [1] tend to require more iterations. Following [13], the data term penaliser in the RRRL method is chosen as $\Phi(z) = 2\sqrt{z}$, whereas the image regulariser R_u is chosen as total variation as in Section 2.

For the PSF estimation, we insert a penaliser function Φ as mentioned above into the discretised data term from (3), which is then combined with the unaltered regulariser R_h from (4) to yield a (partial) discrete energy function for the estimation of $\boldsymbol{h}$:

$$E(\boldsymbol{h}) = \sum_{i,j} \Phi\big((f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}]_{i,j})^2 \big) \, \mathrm{d}\boldsymbol{x} + \alpha R_h(\boldsymbol{h}) \ . \tag{7}$$

Unlike its counterpart in Section 2., this energy function is no longer quadratic. Equating the gradient (i.e. the derivatives w.r.t. $h_{i,j}$) to zero now yields a system of nonlinear equations for the PSF entries.

Figure 1. Blind deconvolution of a synthetically blurred image. (a) Input image, 289×289 pixels, blurred with the PSF, 13×13 pixels, shown as insert. From [5], adapted. – (b) Reconstructed image and PSF (inserted) by the method from [5], $m_x = m_y = 13$, $\beta = 5050$, $K = 200$. – (c) Same as (b) but with RRRL used in the image estimation step, $\alpha = 0.0018$, $K_u = 30$. – (d) Same as (b) but with RRRL for image estimation, and the nonlinear PSF estimation method from Section 3., $\alpha = 0.0018$, $\beta = 5050$, $K_u = 30$, $K_h = 20$. – For τ, the quantile criterion (see Section 2.) was used in (c, d) and yielded values in the range $0.11 \ldots 0.12$. (b)–(d) from [7], adapted.

By a standard procedure of lagged weights (analogous to the lagged diffusivity method or Kačanov method) we transform the nonlinear equation system into a sequence of linear ones. Note that the nonlinearities result from the terms $\partial_{h_{p,q}} \Phi\big((f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}]_{i,j})^2\big) = -2\Phi'\big((f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}]_{i,j})^2\big)(f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}]_{i,j}) u_{i-p,j-q}$. Starting with some initial approximation $\boldsymbol{h}^0$ for $\boldsymbol{h}$, we proceed therefore for $l = 0, 1, 2, \ldots$ as follows: Compute the weights $\varphi_{i,j}^l := \Phi'\big((f_{i,j} - [\boldsymbol{u} * \boldsymbol{h}^l]_{i,j})^2\big)$ and replace $\Phi'(\,\cdot\,)$ in the equation system with the fixed $\varphi_{i,j}$. This gives a linear equation system for $\boldsymbol{h}$. Applying LU decomposition as in Section 2. one computes the solution $\boldsymbol{h}^{l+1}$ of this system, which is the starting point for the next iteration. A more spelled-out derivation of the sequence of linear equation systems is found in [7]. Experimental evidence in [7], see also Section 4., confirms the quick convergence of the sequence $(\boldsymbol{h}^l)$; in practical cases, often $10 \ldots 20$ iterations are sufficient.

To end the description of our robust blind deconvolution method, we summarise its parameters which will also be referred to in Section 4. The original method from [5] and the version with RRRL and linear PSF estimation use obvious subsets of these parameters. We start by the model parameters. First, there are the PSF sizes m_x, m_y that need to be chosen somewhat larger than the actual PSF. For the sample sizes s_x, s_y we adopt the heuristic choice $s_{x,y} \approx 1.5\, m_{x,y}$ from [5]. Regarding the image regularisation weight α in (6), a continuation strategy that starts with a larger α in the first iterations of the alternating minimisation and reduces α during the alternating minimisation process helps to speed up convergence; the final values of α lie in the range $\alpha \approx 0.001 \ldots 0.002$ proposed in [13]. The PSF regularisation weight β is set manually; if it is too small, the blur will be underestimated (with a point kernel as extreme); too large β leads to oversharpening, compare [5]. Finally, there is the threshold τ for the PSF entries. The essential numerical parameters are three iteration counts: K for the alternating minimisation, K_u for RRRL, and K_h for the iterated linearisation of the nonlinear equation system in the PSF estimation.

4. Experiments

As a proof of concept, we present two experiments here; further experiments can be found in [7]. Our first experiment is based on a synthetically blurred image, Fig. 1(a), that was already used in [5] to exemplify the method reviewed in Section 2. The result of this method is shown in Figure 1(b). Frame (c) has been obtained by replacing the image estimation component with RRRL, whereas in frame (d) also the robust PSF estimation from Section 3. has been employed. Comparing (b) and (c), it is

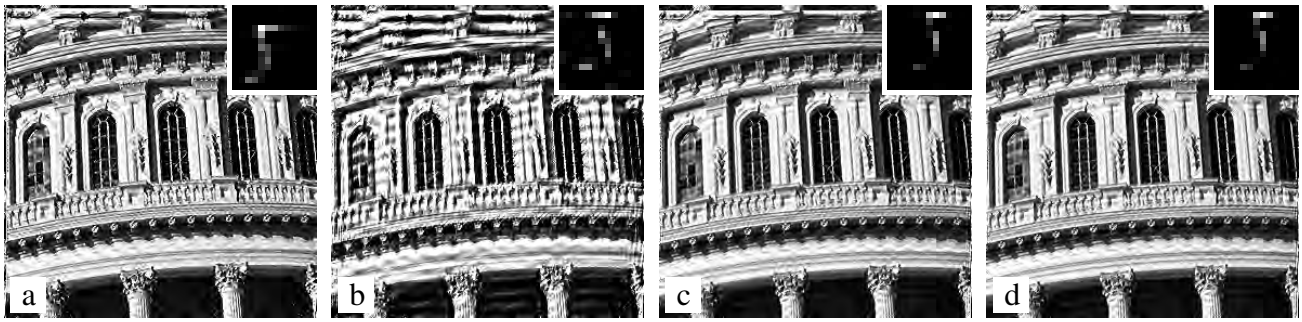


Figure 2. Blind deconvolution of the synthetically blurred image from Fig. 1(a) using RRRL for image estimation, and the nonlinear PSF estimation method from Section 3. with different numbers of linearisation iterations. (a) 1 iteration. – (b) 2 iterations. – (c) 5 iterations. – (d) 8 iterations. – From [7], adapted.

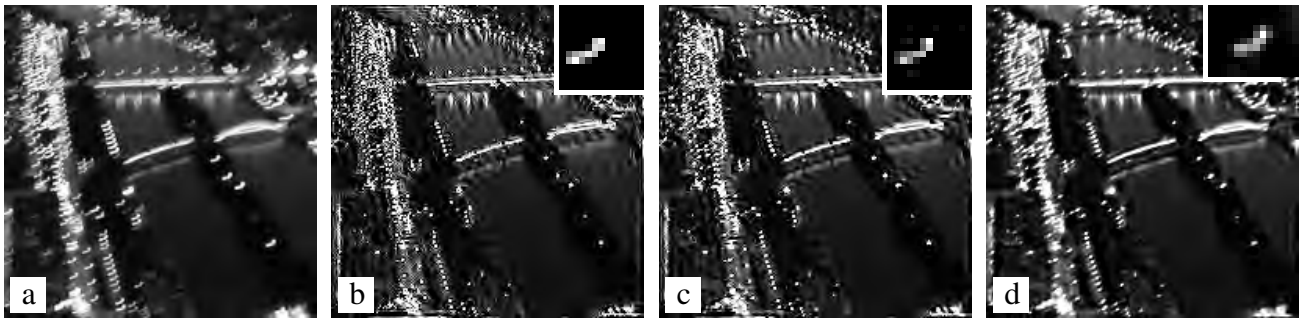


Figure 3. Blind deconvolution of an image blurred during acquisition. (a) Clipping from a photograph (Paris from Eiffel tower at dusk) blurred by camera shake, 200×200 pixels. – (b) Reconstructed image and PSF, 13×13 pixels (inserted), using RRRL for image estimation and PSF estimation according to [5], $m_x = m_y = 13$, $\alpha = 0.002$, $\beta = 260$, $\tau = 0.1$, $K = 300$, $K_u = 20$. – (c) Same as (b) but with nonlinear PSF estimation Section 3., $m_x = m_y = 13$, $\alpha = 0.0105$, $\beta = 255$, $\tau = 0.1$, $K = 300$, $K_u = 20$, $K_h = 8$. – (d) Non-blind RRRL deconvolution result with the manually tuned PSF (shown as insert) from [13], $\alpha = 0.002$, $K_u = 30$. The PSF, 14×11 pixels, has been generated from an impulse response.

evident that introducing robust image estimation brings about a slight gain in reconstruction of small detail, but also an amplification of artifacts is observed which may be attributable to the mismatch between the data terms underlying the PSF estimation (non-robust) and non-blind deconvolution (robust). Using robust estimation methods for both (d) leads to a result with visible gain in sharpness and fewer artifacts. In particular, fine details of the columns between the windows are reconstructed sharper in (d) than in (b). Regarding the visible translation by approx. 2 pixels between (d) and the two other results, it should be noted that shifting the PSF and image in opposing directions is an inherent degree of freedom of the convolution model (1). Note that this also poses a difficulty for quantitative evaluation of blind deconvolution methods: quantitative error measurements cannot be done without a registration step whose influence on the error values needs additional analysis. Since this is not feasible within the present paper, we restrict ourselves to a visual assessment at this point.

As discussed in Section 3. the non-linear system of equations arising in the PSF estimation is solved iteratively by linearisation. In Fig. 2 we demonstrate the evolution of estimated PSF and image with increasing number of linearisation iterations. With a single iteration, frame (a), the result is almost identically to the linear PSF estimation from Fig. 1(c). Additional iterations first lead to some artifacts, frame (b), which are apparently caused by the fact that the non-linear method places the PSF in this example at a translated position. With more iterations, the reconstruction quickly stabilises at the refined result, frame (d), which is numerically converged and corresponds to Fig. 1(d).

In our last experiment, Fig. 3, we consider an image blurred by camera shake. The test image, Fig. 3 (a), is clipped from a test image used in [13] to demonstrate the non-blind RRRL method. In [13] it was used in conjunction with a PSF manually generated from an impulse response (the image of a street light); we reproduce this experiment from [13] in frame (d) for reference. In frames (b) and (c) we show blind deconvolution results: frame (b) again combines RRRL for image estimation with the linear PSF estimation from [5], whereas (c) employs also the robust PSF estimation from Section 3. It is evident that in (c) the estimated PSF has become sharper, and artifacts in the deblurred image have been reduced, although the reconstruction quality is still not quite on par with the non-blind result (d).

5. Summary and Outlook

In this paper we have shown how a recent blind image deconvolution approach by alternating minimisation of a joint energy functional [5] can be improved by introducing robust methods for PSF and image estimation. For image estimation we used RRRL [13], whereas for PSF estimation a modification of the method from [5] has been used that is, to best of our knowledge, new. The viability of the approach has been demonstrated on synthetic and real-world blurred images.

A weakness of this combination of methods is that the robust data terms used in the image and PSF estimation differ, compare (6), (7), and cannot be cast into a joint energy functional. This is a pragmatic decision justified by the efficiency of RRRL and the fact that, as demonstrated in [13], its results in non-blind deconvolution are largely comparable with those of a method in the sense of [1] whose data term is compatible with (7). Notwithstanding, it will be a goal of future work to reformulate the robust model such that PSF and image estimation can be expressed in a unified functional. It is expected that an exact match of data terms will also further reduce artifacts in the blind deconvolution results.

The present paper is restricted to grey-value images; an extension to multi-channel (colour) images will be detailed in a forthcoming publication. Future work might also address strategies for the choice of parameters as well as efficiency improvements of the algorithm. In order to further study the practical applicability of the method, experimental validation using larger sets of images will be important, including quantitative comparisons. Moreover, we have focussed in this work on the ability of robust data terms to cope with imprecise PSF estimation and model violations, but largely ignored their potential in treating strong noise. Experiments on noisy blurred images will deepen insight into this aspect.

References

[1] L. Bar, N. Sochen, and N. Kiryati. Image deblurring in the presence of salt-and-pepper noise. In R. Kimmel, N. Sochen, and J. Weickert, editors, *Scale Space and PDE Methods in Computer Vision*, volume 3459 of *Lecture Notes in Computer Science*, pages 107–118. Springer, Berlin, 2005.

[2] T. F. Chan and C. K. Wong. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7:370–375, 1998.

[3] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. In *Proc. SIGGRAPH 2006*, pages 787–794, New York, NY, July 2006.

[4] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-Laplacian priors. In *Advances in Neural Information Processing Systems*, pages 1033–1041, 2009.

[5] G. Liu, S. Chang, and Y. Ma. Blind image deblurring using spectral properties of convolution operators. *IEEE Transactions on Image Processing*, 23(12):5047–5056, 2014.

[6] L. B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79(6):745–754, June 1974.

[7] P. Moser. Blind image deblurring using the deconvolution operator's spectral properties and non-linear kernel estimation. Master's thesis, UMIT, Hall/Tyrol, Austria, 2015.

[8] N. Persch, A. Elhayek, M. Welk, A. Bruhn, S. Grewenig, K. Böse, A. Kraegeloh, and J. Weickert. Enhancing 3-D cell structures in confocal and STED microscopy: a joint model for interpolation, deblurring and anisotropic smoothing. *Measurement Science and Technology*, 24(12):125703, 2013.

[9] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. T. Flannery. *Numerical Recipes. The Art of Scientific Computing. Third Edition*. Cambridge University Press, Cambridge, 2007.

[10] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

[11] K. Schelten, S. Nowozin, J. Jancsary, C. Rother, and S. Roth. Interleaved regression tree field cascades for blind image deconvolution. In *IEEE Winter Conference on Applications of Computer Vision*, pages 494–501, 2015.

[12] C. J. Schuler, H. C. Burger, S. Harmeling, and B. Schölkopf. A machine learning approach for non-blind image deconvolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1074, 2013.

[13] M. Welk. A robust variational model for positive image deconvolution. *Signal, Image and Video Processing*, 10(2):369–378, 2016.

[14] M. Welk, P. Raudaschl, T. Schwarzbauer, M. Erler, and M. Läuter. Fast and robust linear motion deblurring. *Signal, Image and Video Processing*, 9(5):1221–1234, 2015.

[15] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. MIT Press, Cambridge, MA, 1949.

[16] L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010, Part I*, volume 6311 of *Lecture Notes in Computer Science*, pages 157–170. Springer, Berlin, 2010.

[17] L. Xu, J.SJ. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

[18] Y.-L. You and M. Kaveh. Anisotropic blind image restoration. In *Proc. 1996 IEEE International Conference on Image Processing*, volume 2, pages 461–464, Lausanne, Switzerland, September 1996.

[19] M. E. Zervakis, A. K. Katsaggelos, and T. M. Kwon. A class of robust entropic functionals for image restoration. *IEEE Transactions on Image Processing*, 4(6):752–773, June 1995.

WS 3: Geometry / Sensor Fusion

Graph-Laplacian minimisation for surface smoothing in 3D finite element tetrahedral meshes

Richard Huber, Martin Holler and Kristian Bredies*

University of Graz, Institute for Mathematics and Scientific Computing

Abstract

We propose a new method to improve surface regularity of 3D tetrahedral meshes associated with finite element simulations of the heart. Our approach is to minimise the graph-Laplacian subject to suitable point constraints. These constraints are computed from the whole triangulation and prevent a worsening of mesh quality that would otherwise be caused by the smoothing. The resulting minimisation problem is solved via a primal-dual algorithm, leading to a method that globally updates vertex coordinates in each iteration. Experiments confirm that our method reduces surface oscillations of the mesh while preventing degeneration of the triangulation as indicated by mesh quality metrics.

1. Introduction

In biomedical engineering, the development of a realistic 3D simulation framework for the human heart is currently an active research topic. Such a framework would allow, for example, patient-specific models and more individualised treatment [6]. In order to carry out such simulations, 3D meshes are typically created from segmented magnetic resonance (MR) images, using mesh-generation software such as described in [12]. In view of the subsequent simulations, these procedures ensure a sufficient quality of the triangulation, as indicated by quality metrics, and prevent the creation of degenerate elements. However, due to physical limitations in the image acquisition and, consequently, a low resolution of the image data, such meshes often suffer from artifacts. Those appear in particular in form of oscillations on the otherwise smooth surface (see Section 5.).

It is the goal of this work to provide a method that reduces these oscillations, but maintains high mesh quality. To this aim, we minimise the graph-Laplacian under suitable constraints and adapt the mesh coordinates accordingly. The constraints are computed from the whole initial triangulation and ensure non-degeneracy of the resulting triangulation and maintenance of a high mesh quality, the latter being indicated by quality metrics.

As the computation of meshes from segmented image data and a subsequent reduction of mesh artifacts is a challenge that commonly appears in mesh generation for finite element simulations in many different contexts, a lot of research has already been carried out in that direction. Different to classical mesh improvement dedicated to enhancing the quality of the triangluation, that often focus on a local adaption of nodes [8, 9, 11], our method aims at reducing mesh artifacts and hence is more related to mesh denoising approaches. For the latter, we exemplary refer to [10, 13, 14] and the references therein for recent methods. For a general overview on mesh related topics see [2, 3].

*University of Graz, Institute for Mathematics and Scientific Computing, Heinrichstrasse 36, A-8010 Graz, Austria. Email: richard.huber@edu.uni-graz.at, martin.holler@uni-graz.at, kristian.bredies@uni-graz.at

2. Model problem

The initial setting is as follows: The 3D tetrahedral finite element mesh is given in form of a triangulation. In particular, the coordinates of points of the triangulation, together with edge information, and a masking of surface points is given. The triangulation is assumed to be regular, in particular, all tetrahedra are non-degenerate and disjoint except for their boundaries.

Since the above-described oscillatory artifacts appear on the surface of the mesh, we will only adapt surface points and use the position of interior points only to determine suitable point constraints. This also reduces the computational cost and memory requirements, however, will have some drawbacks as discussed in Section 6.

The triangulation of the surface induces a graph $G = (V, E)$ with vertices $V = \{v_1, \dots, v_N\}$, where N is the number of vertices, and edge set E such that there is an edge between v_i and v_j in G if, and only if, $\{v_i, v_j\} \in E$. We define $U := \mathbb{R}^{3\times N}$ to be the space of point-coordinates of the triangulation, where for $u \in U$, the jth coordinate of the vertex v_i is denoted by $u_i^j \in \mathbb{R}$. Further, we will use the notation $u^j \in \mathbb{R}^N$ for the vector containing all jth coordinates of u and $u_i \in \mathbb{R}^3$ for the coordinates of v_i. With this notation, we define the graph-Laplacian operator as the componentwise matrix-multiplication operator according to

$$\Delta u := \begin{pmatrix} \hat{\Delta} u^1 \\ \hat{\Delta} u^2 \\ \hat{\Delta} u^3 \end{pmatrix}, \text{with the matrix } \hat{\Delta} \in \mathbb{R}^{N\times N}, \text{ given as } \left(\hat{\Delta}\right)_{i,j} := \begin{cases} \mathrm{Deg}(v_i) & \text{if } i = j, \\ -1 & \text{if } \{v_i, v_j\} \in E, \\ 0 & \text{else,} \end{cases} \quad (1)$$

where $\hat{\Delta} u^j$ is a matrix-vector multiplication and $\mathrm{Deg}(v_i)$ denotes the degree of v_i, i.e., the number of neighbours of v_i in G.

In order to smooth the surface, new coordinates of the surface points are computed by minimising the graph-Laplacian under constraints designed to maintain the original mesh structure and to ensure non-degeneracy of the mesh. The minimisation problem is

$$u^+ \in \operatorname*{argmin}_{u\in U} \frac{1}{2}\|\Delta u\|_2^2, \qquad \text{subject to } u \in \Omega, \quad (2)$$

where the feasible set has the form $\Omega = \{u \in U \colon u_i \in \Omega_i \text{ for } i = 1, \dots, N\}$, with pointwise feasible sets Ω_i as defined in the next section. A solution u^+ corresponds to the coordinates of the nodes of the smoothed surface. Note that the topology of the mesh, and in particular the set of edges E, does not change and Δ is linear. A minimisation of $\|\Delta u\|_2^2$ results in the node coordinates adapting to the means of the surrounding ones, and thus, reduces the curvature of the surface. Hence, minimising the graph-Laplacian operator is expected to imply a smoothing of the surface mesh.

Well-posedness. As we will see in the next section, it is reasonable to choose Ω to be non-empty, bounded and closed. Hence, existence of a solution to (2) follows directly from continuity of $u \mapsto \|\Delta u\|_2^2$ and finite dimensionality of U.

3. Suitable constraints

Naturally, the solution of (2) should be close to the original data. Further, the choice of Ω is driven by two requirements, the convexity of Ω and the maintenance of mesh quality:

Mesh quality. An important aim is to keep mesh quality high, since this is needed for the finite element simulation to work. A high mesh quality means that the tedrahedra are non-degenerate, disjoint, and that there is only a small number of very flat tetrahedra. The latter is important since many flat tetrahedra would cause numerical problems in the simulations. Our assumption is that the quality of the original mesh is sufficiently high, therefore we design the constraints such that the movement of the nodes does not significantly worsen mesh quality. In particular, we want to guarantee that no self-intersection of the surfaces of the tetrahedra occurs.

Convexity of Ω. Convexity yields several advantages in optimisation, such as allowing to apply a large range of optimisation methods and ensuring that indeed global optima are approximated. Thus, we aim to define constraints which can be represented as a family of point constraints given in a way that the set of admissible point-coordinates is convex.

In summary, to achieve the best results with our method, we look for a convex set of constraints which allows sufficient movement of the nodes while maintaining a high mesh quality.

Adaptive constraints. We define Ω by fixing an individual radius r_i for each node v_i, and allowing the node only to move within a ball of this radius centered at its original location. Our approach to choose r_i is as follows: Let us fix a surface vertex v in a tetrahedron T. Since the goal is to avoid degenerate tetrahedra, one must in particular prevent self-intersection. Geometrically interpreted, this means that each of the nodes must not pass to the opposite side of T. This motivates the incorporation of the heights on the nodes in T. Indeed, if the other nodes did not change, the distance of v to the opposite side of T would be determined by the corresponding height h of the tetrahedron and one could use h as a limitation on how far the vertex is allowed to move. But since the movement of the other points of the tetrahedron also affects this consideration, and since the node v is not only a node of T, but of several neighbouring tetrahedra, we use all heights h of all tetrahedra containing v to define the constraints. Indeed, for a fixed node v_i, we denote by h_T the minimum of the four heights of a tetrahedron T and $\hat{h}_i = \min\{h_T : v_i \text{ contained in } T\}$. We limit the movements of v_i by $\alpha\hat{h}_i$ with a parameter $0 < \alpha < 1/2$, which is expected to ensure that, even though all nodes move simultaneously, no self-intersections occur. Let u_{0i} denote the original coordinates of the vertex v_i. Thus, the corresponding radii and the resulting feasible sets are given by

$$r_i = \alpha\hat{h}_i \quad \text{with } \hat{h}_i = \min\{h_T \colon v_i \in T\} \text{ and } h_T = \min\{h \colon h \text{ height of } T\}, \tag{3}$$

$$\Omega = \{u \in U \colon \|u_i - u_{0i}\| \leq r_i \text{ for } i = 1, \ldots, N\}. \tag{4}$$

This ensures that no self-intersection occurs and mesh quality is maintained. Figure 1 illustrates such constraints for the case of 2D triangles. In the three-dimensional setting, also the interior vertices adjacent to the surface of the mesh will be incorporated in the computation of constraints.

4. Numerical solution

The aim in this section is to describe an algorithmic framework for the solution of (2) with Ω as in (4). For this purpose, we will use the primal-dual algorithm described in [5], which is an iterative method that allows to solve convex-concave saddle-point problems with non-smooth structure. A non-smooth optimisation method is required to incorporate the proposed point constraint, however, due to differentiability of the graph-Laplacian regularisation and simplicity of the feasible set, also other methods, such as FISTA [1], could be used.

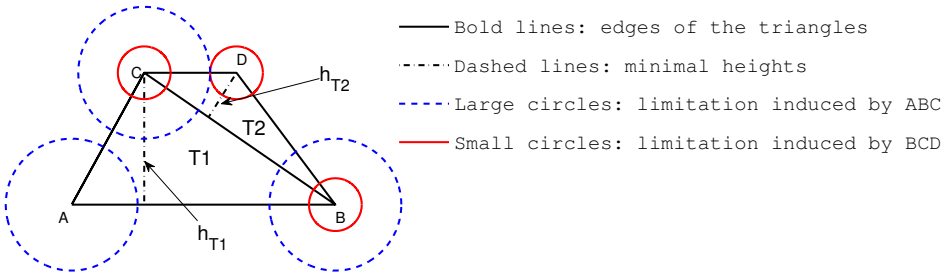


Figure 1: **Triangle $T1 = (A, B, C)$ with minimal height h_{T1} on C and triangle $T2 = (C, D, B)$ with minimal height h_{T2} on D. Limiting the movement of all nodes in a triangle by α times the minimum of the heights induces circles for each node, of which the smallest one is chosen as constraints.**

In order to apply the primal-dual algorithm, Problem (2) is reformulated as a saddle-point problem according to

$$\min_{u\in\Omega} F(\Delta u) \quad \Longleftrightarrow \quad \min_{u\in U} F(\Delta u) + I_\Omega(u) \quad \Longleftrightarrow \quad \min_{u\in U}\sup_{w\in U} \langle w, \Delta u\rangle - F^*(w) + I_\Omega(u), \tag{5}$$

where $F(u) = \frac{1}{2}\|u\|_2^2$, the indicator function of Ω, i.e., $I_\Omega(u) = 0$ for $u \in \Omega$ and ∞ otherwise, and F^* is the convex conjugate of F, defined as $F^*(w) := \sup_{u\in U}\langle w, u\rangle - F(u)$. Explicitly, we get

$$F^*(w) = \sup_{u\in U} \langle w, u\rangle - \frac{1}{2}\|u\|_2^2 = \langle w, w\rangle - \frac{1}{2}\|w\|_2^2 = \frac{1}{2}\|w\|_2^2, \tag{6}$$

where the second equality is due to $u = w$ being the unique critical point of $u \mapsto \langle w, u\rangle - \frac{1}{2}\|u\|_2^2$, which can be confirmed by differentiation, and hence, $u = w$ being the unique global maximiser. Thus, (2) is reformulated as the following saddle point problem

$$\min_{u\in U}\max_{w\in U} L(u, w), \quad \text{where } L(u, w) = \langle w, \Delta u\rangle - \frac{1}{2}\|w\|_2^2 + I_\Omega(u). \tag{7}$$

The following proposition shows that by solving (7), we indeed obtain a solution of the original problem (2).

Proposition. *The saddle point problem* (7) *with feasible set* Ω *defined as in* (4) *admits at least one solution and for any saddle point* (u^+, w^+) *of* (7)*,* u^+ *is a solution of the original minimisation problem* (2)*.*

Proof. Due to [7, VI Prop 2.4, p. 176], it is sufficient to show that for $L\colon U \times U \to \overline{\mathbb{R}}$ defined as in (7), for $u \in U$ fixed, $w \mapsto L(u, w)$ is concave and upper semi-continuous on U, and for $w \in U$ fixed, $u \mapsto L(u, w)$ is convex and lower semi-continuous on U. Further, we need to show that $u \mapsto L(u, w)$ is coercive for fixed w and that

$$\lim_{\substack{\|w\|\to\infty \\ w\in U}} \inf_{u\in U} L(u, w) = -\infty. \tag{8}$$

The convexity/concavity and l.s.c./u.s.c. assumptions are satisfied, in particular due to Ω being convex and closed, and $u \mapsto L(u, w)$ is coercive due to Ω being bounded. Further, for fixed $u \in \Omega$,

$$\lim_{\|w\|\to\infty} \langle w, \Delta u\rangle - \|w\|_2^2 \le \lim_{\|w\|\to\infty} \|w\|\|\Delta u\| - \frac{1}{2}\|w\|_2^2 = \lim_{\|w\|\to\infty} \|w\|\Big(\|\Delta u\| - \frac{1}{2}\|w\|\Big) = -\infty$$

and hence, (8) holds, yielding the existence of a saddle point (u^+, w^+). Due to [7, III Prop 3.1, p. 57], the optimality of u^+ for (2) is a direct consequence of (5). □

The primal-dual algorithm for the solution of (7) will also require knowledge of the operator norm $\|\Delta\|$. An estimate can be found via power iteration [4], which computes λ_{max}, the eigenvalue of Δ with the greatest modulus if it is well separated from other eigenvalues. Note that this eigenvalue λ_{max} equals $\|\Delta\|$ due to Δ being symmetric and positive semidefinite.

The iteration steps of the primal-dual algorithm are given, in the abstract form, as

$$\begin{cases} w_{k+1} = (\mathrm{id} + \sigma \partial F^*)^{-1}(w_k + \sigma \Delta \bar{u}_k) \\ u_{k+1} = (\mathrm{id} + \tau \partial I_\Omega)^{-1}(u_k - \tau \Delta w_{k+1}) \\ \bar{u}_{k+1} = 2u_{k+1} - u_k \end{cases} \tag{9}$$

for suitable parameter $\tau, \sigma \in (0, \infty)$ such that $\|\Delta\|^2 \tau\sigma < 1$. Since $F^*(u) = \frac{1}{2}\|u\|_2^2$ is differentiable, a simple computation shows that $\partial F^*(u) = u$, thus

$$z = (\mathrm{id} + \sigma \partial F^*)^{-1}(u) \iff z + \sigma z = u \iff z = \frac{u}{1+\sigma}.$$

Further,

$$\begin{aligned} z = (\mathrm{id} + \tau \partial I_\Omega)^{-1}(u) &\iff z + \tau \partial I_\Omega(z) \ni u \iff 0 \in \partial\Big(\frac{1}{2}\|u - \cdot\|_2^2 + \tau I_\Omega(\cdot)\Big)(z) \\ &\iff z \in \operatorname*{argmin}_{v \in U} \|u - v\|_2^2 + \tau I_\Omega(v) \iff z \in \operatorname*{argmin}_{v \in \Omega} \|v - u\|_2^2 \\ &\iff z = P_\Omega(u), \end{aligned} \tag{10}$$

where $P_\Omega(u)$ denotes the projection of u onto Ω, i.e., onto the element in Ω with minimal distance to u. Hence, (10) can be solved by projecting onto the closest feasible point. We can compute this projection for each node individually since only point constraints are considered, i.e., whether or not $\|u_i - u_{0i}\| \leq r_i$ does not depend on the other nodes' locations. The projection for each node is simply the projection on the ball of radius r_i centered at the original location u_{0i}, i.e.,

$$P_\Omega(u)_i = p(u_i, u_{0i}, r_i), \quad \text{with} \quad p(x, y, r) = \begin{cases} x & \text{if } \|x - y\| \leq r, \\ \frac{r(x-y)}{\|x-y\|} + y & \text{else.} \end{cases} \tag{11}$$

Note that Ω, and hence, u_{0i} and r_i, do not change during the iteration and r_i is determined according to (3) and (4). By inserting (10) and (11) into (9), the iterations can be computed by simple arithmetic operations resulting in Algorithm 1.

Algorithm 1 Primal-Dual algorithm for minimising graph-Laplacian with adaptive constraints

Input: Original point-coordinates $\tilde{u}_0$ of mesh, edge information $\mathcal{E}$, masking of surface points S.

$u_0 \leftarrow$ extract_surf_coo$(\tilde{u}_0)$, $r \leftarrow$ get_radii$(\tilde{u}_0, \mathcal{E}, S)$, $\Omega \leftarrow$ get_$\Omega(r, u_0)$ ▷ constraints
$\Delta \leftarrow$ get_$\Delta(\mathcal{E}, S)$, $\|\Delta\| \leftarrow$ powiter(Δ) ▷ inititialisation of Laplacian
$u \leftarrow u_0$, $\bar{u} \leftarrow u_0$, $w \leftarrow 0 \in \mathbb{R}^{3\times N}$, $\tau \leftarrow \|\Delta\|^{-1}$, $\sigma \leftarrow \|\Delta\|^{-1}$
repeat
  $w \leftarrow \frac{(w + \sigma\Delta\bar{u})}{(1+\sigma)}$ ▷ update of the dual variable
  $\bar{u} \leftarrow P_\Omega(u - \tau\Delta w)$ ▷ update of the primal variable
  $u \leftarrow 2\bar{u} - u$ ▷ update of the extragradient
  $(u, \bar{u}) \leftarrow (\bar{u}, u)$ ▷ interchange of u and $\bar{u}$
until maximal number of iterations is reached
return u

Output: $u^+ = u$ surface point-coordinates of smoothed mesh.

Note that this is a global method, i.e., it updates the positions of all surface vertices in each iteration, unlike many other surface smoothing algorithms which operate pointwise.

Reiteration. In some situations, the proposed constraints are too restrictive, and hence, the smoothing results are not satisfactory. To overcome that, the point-constraints for each single point would need to be updated iteratively with the position of all other points. This would, however, result in a non-convex problem, preventing the computation of global optima.

A heuristic approach to still achieve some improvement, without re-designing the overall method, is to restart Algorithm 1 after convergence. To this aim, new constraints are computed from the output u^+ and the graph-Laplacian is optimised again subject to these updated constraints. This can be repeated a few times, e.g., 4 times, to allow some more flexibility in the constraint set. In practice, it can be reasonable to reduce the number of iterations performed in Algorithm 1, and do a few outer iterations in order to allow for more movement, while still guaranteeing that no self-intersection occurs and the mesh quality remains high.

Independent of such heuristics, the point-constraints of our method always ensure a non-degenerate triangulation. Also, the inner points of the mesh are not moved by our methods and hence limit the effect of the re-iteration. This, together with the point-constraints, in particular prevents a strong decrease of the volume of the shape, as frequently observed with unconstrained Laplacian smoothing.

5. Experimental results

The proposed method, although rather simple, is quite effective. It allows to smooth the surface and to reduce artifacts significantly while maintaining the original level of mesh quality. Figure 2 illustrates the effects of smoothing, with the original model on the left side, and the smoothed version on the right. The figure shows a mesh of a human heart, where the smoothed version was computed with 3 outer and 1000 inner iterations and with the constraint parameter $\alpha = 2/5$.

The effect of the proposed method on mesh quality can be evaluated quantitatively by measuring ρ, the skewness of a tetrahedron, i.e., the ratio of a tetrahedron's volume to its circumscribed ball's volume. Additionally, we quantify the change of the volume of each tetrahedron and identify changed orientations. This is done for each tetrahedron in the mesh by measuring the ratio of $\det(A)$ in the original and the smoothed mesh, denoted by θ, where A is a parallelepiped induced by a the tetrahedron.

Furthermore, one can observe maximal and minimal angles in the tetrahedra in order to find very flat tetrahedra. Table 1 depicts a quantitative evaluation of the effect of our method on mesh quality by comparing ρ for the original and the smoothed mesh and computing θ. As one can see, the number of flat structures does not increase significantly due to smoothing and for only 1% of the tetrahedra the volume reduced by more than one half. Further, we observed that no sign-flips of the determinant occurred, hence there are no self-intersections.

Percentiles of P	1%	5%	10%
Original mesh	0.0900	0.2350	0.3348
Smoothed mesh	0.0934	0.2047	0.2812

Percentiles of Θ	1%	5%
	0.5473	0.6648

Table 1: Mesh quality corresponding to mesh considered in Figure 2. Percentiles of P and Θ, where P is a vector of ρ for all tetrahedra and Θ is a vector of θ for all tetrahedra.

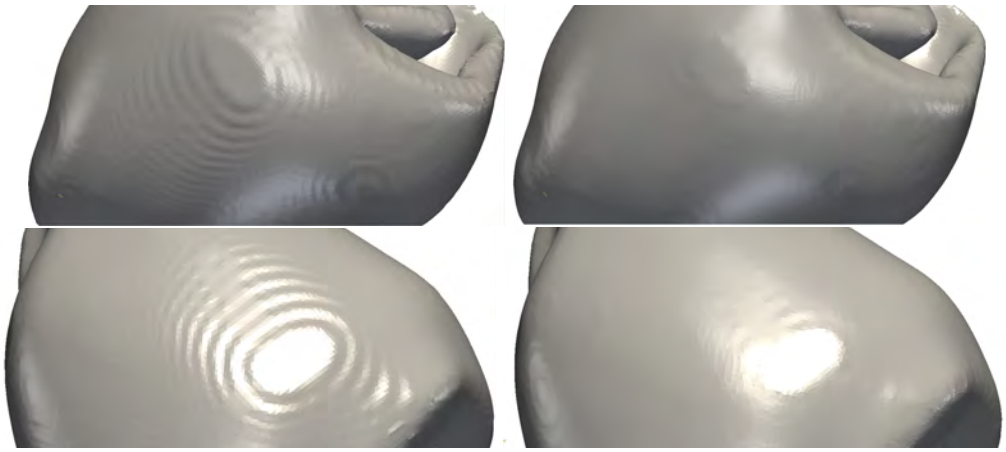


Figure 2: **3D triangulation of a human heart. The left figure shows the surface of the original mesh with artifacts, while the right shows the corresponding smoothed version where the artifacts are reduced.**

6. Discussion and outlook

The proposed method allows for improvement of the visual surface quality in 3D tetrahedral meshes. However, the procedure does not always succeed in removing all artifacts as can be observed particularly when there is an area lying dominantly above its surrounding surface like a plateau. The reason for this might be that only the surface, and thus, the outermost tetrahedra are changed, while the layer below remains unchanged. The constraints that avoid the loss of mesh quality are disadvantageous in this regard, since the second layer prevents the outer layer from sinking. Therefore, the plateau might remain dominant above its surrounding. A possible solution is not only to change the outermost layer, but also a few layers inside as well. However, this would, of course, increase computational costs. Another possibility would be to modify the constraints to avoid such a problem, in particular, also consider non-convex bounds. Indeed, this would allow for more flexibility in choosing the constraints, however, at the cost of losing the advantageous properties gained due to convexity.

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[2] M. Botsch, L. Kobbelt, M. Pauly, L. Alliez, and B. Lévy. *Polygon mesh processing*. CRC press, Taylor and Francis, 2010.

[3] M. Botsch, M. Pauly, L. Kobbelt, P. Alliez, B. Lévy, S. Bischoff, and C. Rossl. Geometric modeling based on polygonal meshes. In *ACM SIGGRAPH Course Notes*, 2007.

[4] S. Börm and C. Mehl. *Numerical Methods for Eigenvalue Problems*. Walter de Gruyter, 2012.

[5] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2010.

[6] A. Crozier, C. M. Augustin, A. Neic, A. J. Prassl, M. Holler, T. E. Fastl, A. Hennenmuth, K. Bredies, T. Kuehne, M. J. Bishop, S. A. Niederer, and G. Plank. Image-based personalization of cardiac anatomy for coupled electromechanical modeling. *Annals of Biomedical Engineering*, 44(1):58–70, 2016.

[7] I. Ekeland. *Convex analysis and variational problems*. SIAM, 1999.

[8] L. Freitag and C. Ollivier-Gooch. Tetrahedral mesh improvement using swapping and smoothing. *International Journal for Numerical Methods in Engineering*, 40(21):3979–4002, 1997.

[9] B. M. Klingne and J. R. Shewchuk. Aggressive tetrahedral mesh improvement. In *Proceedings of the 16th International Meshing Roundtable*, pages 3–23. Springer, 2008.

[10] X. Lu, Z.Deng, and W. Chen. A robust scheme for feature-preserving mesh denoising. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1181– 1194, 2016.

[11] M. Ohlsson, P. Toft, L.K. Hansen, and F.A. Nielsen. Active surface models for brain imaging. *In Proceedings of the Interdisciplinary Inversion Workshop*, 5:68–77, 1997.

[12] A. J. Prassl, F. Kickinger, H. Ahammer, V. Grau, J. E. Schneider, E. Hofer, E. J. Vigmond, N. A. Trayanova, and G. Plank. Automatically generated, anatomically accurate meshes for cardiac electrophysiology problems. *IEEE Transactions on Biomedical Engineering*, 56(5):1318ÔÇô30, 2009.

[13] H. Zhang, C. Wu, J. Zhang, and J. Deng. Variational mesh denoising using total variation and piecewise constant function space. *IEEE Transactions on Visualization and Computer Graphics*, 21(7):873–886, 2015.

[14] Y. Zheng, H. Fu, O.J.C. Au, and C.L. Tai. Bilateral normal filtering for mesh denoising. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1521–1530, 2011.

Depth estimation using light fields and photometric stereo with a multi-line-scan framework

Doris Antensteiner[1], Svorad Štolc[1], and Reinhold Huber-Mörk[1]

Austrian Institute of Technology
Digital Safety and Security Department, Austria
{*doris.antensteiner.fl,svorad.stolc,reinhold.huber-moerk*}*@ait.ac.at*

Abstract
In this paper we deal with a combination of two state-of-the-art computational imaging approaches - (i) light fields and (ii) photometric stereo - in order to improve the quality of 3D reconstructions within a multi-line-scan framework. Computational imaging uses a redundant description of an image scene to reveal information which would not have been available via conventional imaging techniques. In the case of light fields the redundancy is achieved by observing the scene from many different angles, which allows capturing 3D shapes in areas with a prominent surface structure using stereo vision techniques. Contrarily, photometry makes use of multiple illuminations in order to capture local surface deviations without the necessity of any surface structure. As photometric surface reconstruction is very sensitive to fine surface details and light fields excel in capturing global shapes, naturally a more complete description can be achieved through a combination of both techniques. We present a compact hybrid photometric light field setup with relatively low costs and improved accuracy, which is therefore well suited for industrial inspection. A multi-line-scan camera is statically coupled with an illumination source to obtain light field data which is also comprised of photometric information. Novel algorithms have been developed to use this data for an improved 3D reconstruction, which exhibits large-scale accuracy as well as sensitivity to fine surface details.

1. Introduction

Traditional film cameras as well as digital cameras capture light rays and project images of the environment onto a 2D plane. Non-traditional approaches such as multi-camera arrays, plenoptic cameras or coded apertures capture a portion of the so called 4D light field and further subdivide this ray space with respect to position and orientation. Photometric stereo makes explicit use of directional variation of illumination. The presence of different lighting conditions, in cases where the light field is dynamically constructed over time, introduces a photometric variation into the data structure. The combination of light fields, describing the variation of image content over observational directions, and photometric approaches, describing the variation of image content depending on lighting directions, is a promising research direction.

In general, the combination of methods which are (i) locally precise but globally inaccurate with (ii) globally accurate methods, which are lacking local structure, was approached from different perspectives, e.g. combining shading with RGB-D [12], improving a depth map from time-of-flight with polarization cues [3], or using stereo vision with photometric stereo [5]. Methods used for the combination are either based on the fusion of normal vectors provided by different approaches, the fusion of depth measurements, or employing variational methods.

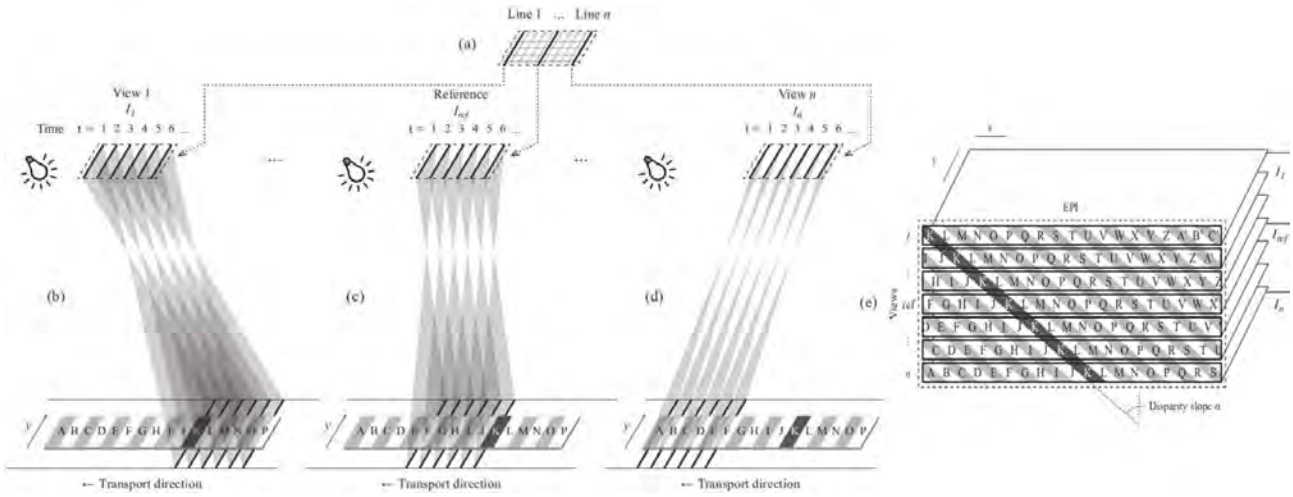


Figure 1: Multi-line-scan setup with directional lighting: a) multi-line sensor, views constructed over time, imaging the object area divided in uppercase letters: b) view 1, c) reference and top-down view, d) view n, e) EPI stack holding views denoting object lines by uppercase letters the disparity slope α.

The depth reconstruction from light field data is usually estimated through the epipolar plane image (EPI) data structure. EPIs were originally introduced for the estimation of structure from motion [1], but they also became a popular tool in light field processing [10],[4]. Kim et al. [4] use an easy criterion for ranking depth hypotheses, namely the best hypothesis is the one, for which as many radiance values as possible along the hypothesized slope in an EPI are similar enough to the radiance in the reference view. Venkataraman et al. [8] use pattern matching between different views, i.e. for a discrete number of hypothesized depths the sum of absolute differences (SAD) of radiances between different views is calculated. Wanner and Goldlücke [10] suggest a statistical approach to estimate the principal orientation of linear structures in EPIs via analysis of the structure tensor constructed locally in small EPI neighborhoods.

This paper is organized as follows. We describe the proposed setup in Sec. 2. In Sec. 3. we describe the fusion framework for light fields and photometric stereo. First results describing the work in progress is given in Sec. 4. In Sec. 5. we draw first conclusions and discuss further work.

2. Multi-Line-Scan Setup

Light fields provide 4-D information, consisting of two spatial and two directional dimensions. They can be captured e.g. by a multiple camera array [11], where each camera has a different viewing perspective of the scene, or by plenoptic cameras [6], which usually make use of a microlens array placed in front of the sensor plane to acquire angular reflectance information.

Our multi-line-scan framework [9] is a light field acquisition setup, where we use an area-scan sensor to observe the object under varying angles, while the object is transported in a defined direction over time. This setup works in real-time and in-line for industrial inspection setups. Fig. 1 illustrates how the light field data is obtained through multiple viewing angles on the moving object over time. Each sensor line observes the conveyor belt in a different viewing angle and captures a certain region. As the object moves under the observed sensor lines, see Fig. 1a, each sensor line captures every object region at distinct time instances, see Figs. 1b,c, and d. We represent the thereby captured light fields as light field image stacks, see Fig. 1e, in which each image is acquired from a slightly different

viewpoint along one direction.

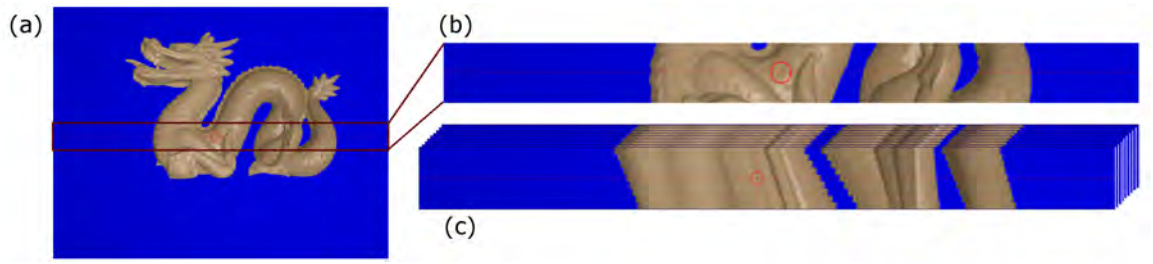


Figure 2: Dragon Stanford [13] scene image (a), zoomed in at the red dotted line (b), where EPI stacks with 9 different viewpoints are shown, each stack formed with a certain light direction (c).

We analyse the captured light fields in the EPI domain [1]. A cut through the light field stack shows linear slope structures, where the angle of the slope corresponds to the disparity and thereby the depth of the scene, as shown in Fig. 2. Each angle of a slope in the EPI stack corresponds to a defined distance between the camera and the object point. Photometric information is obtained by a static light source w.r.t. the sensor while the object is moving. As an object moves on the conveyor belt, the relationship between the illumination and the observation angle changes in a systematic way, so that the surface inclination in the transport direction can be estimated. This photometric information is used to estimate the surface normals of the object.

3. Combination of Light Fields and Photometric Stereo

Photometric stereo describes the surface variation w.r.t. the lighting direction. Reflections of the light on the objects' surface under different lighting orientations provide information about the surface normals at each object point. We combine the depth from light fields with fine surface structures as observed by photometric stereo, to gain an improved depth map of the scene. Figure 3 shows the depth estimation achieved using both light field and photometric stereo independently in a virtual test setup, where we simulated 81 camera viewpoints and 25 illumination angles.

3.1. Light Field Depth Estimation

Depth information of a scene can be retrieved by analyzing the slope angles in EPI stacks. An EPI slice of this stack is shown in Fig. 2, where each angle in the slice refers to a defined depth of a corresponding point in the scene. Using this data we gain a rough absolute depth estimation of the scene.

Analyzing the depth from EPI stacks can be seen as finding such an angle α^* for each point (x, y), where the difference between values at the sheared coordinates $(x(\alpha), y(\beta))$ in the light field L and the reference view I_0 is minimal [7].

$$\alpha^*(x, y) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{i=n} |I_i(x(\alpha), y(\beta)) - I_0(x, y)|$$

The use of a block-matching approach creates a higher robustness to both noise and non-Lambertian lighting conditions. As shown in Fig. 3b, light field data provides quite robust absolute depth estimation, but lacks precision in fine surface details.

(a) dragon (b) LF (c) PS

Figure 3: Light field (LF) and photometric stereo (PS) depth information.

3.2. Photometric Stereo Surface Normals

The surface appearance depends on the shape, reflectance, and illumination of the object. In our virtual test setup we arranged 25 light sources on a sphere around the object in order to capture several images from the same viewpoint under different illumination angles. Contrarily, in our real world setup different illumination angles are achieved by the movement of the object under the light source. We assume a Lambertian reflection model to describe the radiance. The pixel intensity vector $Iv = [i_1, ..., i_{25}]$ for each light source and pixel depends on the illumination vector $L = [L_1, ..., L_{25}]$, as well as on the estimated surface normal unit vector N to the surface and the surface albedo ρ.

$$Iv = \rho \cdot L \cdot N$$

Thereby, we solve the albedo and normal vectors with $\rho \cdot N = L^{-1} \cdot Iv$. The depth map is then integrated using the algorithm of Frankot and Chelappa [2].

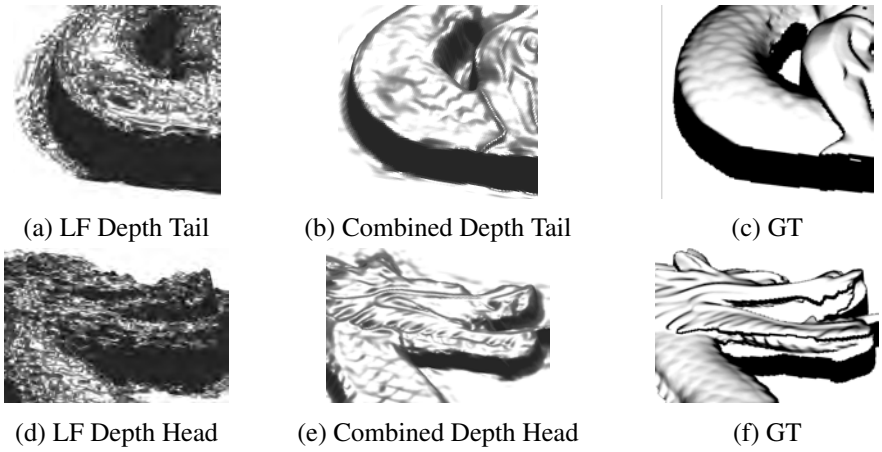
As shown in Fig. 3c, this photometric stereo approach results in fine depth measurements and a strong relative depth accuracy, while the absolute depth accuracy suffers from an accumulative offset. We use the benefits of both the photometric stereo and light field depth estimation to achieve an improved depth estimation result.

3.3. Combination

We refine the light field depth map, as shown in Fig. 3b, using high frequency photometric stereo depth information, as shown in Fig. 3c. Depth from light field yields reliable absolute depth measures, but suffers both from inaccurately estimated details in the structure and high frequency noise. Low frequencies in the light field depth map D_l are extracted using a bilateral smoothing filter f_l. High frequency components are taken from the photometric stereo depth map D_p, using a high-pass image filter f_h. Depth refinement is obtained by replacing high frequency information from the light field depth map by the according high frequencies in the photometric stereo depth map. Our final depth map D is thereby constructed as the linear combination of the low frequency components from the light field depth map and the high frequency components from the photometric stereo depth map, weighted by the factors λ_l and λ_p respectively.

$$D = \lambda_l \cdot D_l * f_l(u, v) + \lambda_p \cdot D_p * f_h(u, v)$$

Results are shown in Fig. 4, where 4a and 4d hold the depth data from light field images from both the head and the tail of the dragon object. The second column, see Figs. 4b and 4e, shows the photometrically refined depth map, using our combinational approach.



(a) LF Depth Tail (b) Combined Depth Tail (c) GT

(d) LF Depth Head (e) Combined Depth Head (f) GT

Figure 4: Combination of light field and photometric stereo in order to improve fine structures in the light field depth estimation. Both rows show a part of the tested object: (a),(d) original light field depth, (b),(e) improved depth by photometric stereo, (c),(f) ground truth depth map.

4. Experimental Results

We performed experiments on several coins, using the multi-line-scan setup, an example is shown in Fig. 5. Transporting the coin, shown in Fig. 5a, along the conveyor belt, we acquired light field and photometric stereo data. The depth information gained from light field data is shown in Fig. 5b. Surface normals are estimated from the same photometric light field data through the detection of a specular lobe in each image location. The specular lobe position corresponds with the local orientation of the surface. The image depth is calculated using the surface normals as described in Subsection 3.2. In Fig. 5c we obtained a refined solution by the combination of photometric stereo depth and light field depth, as described in Sec. 3.

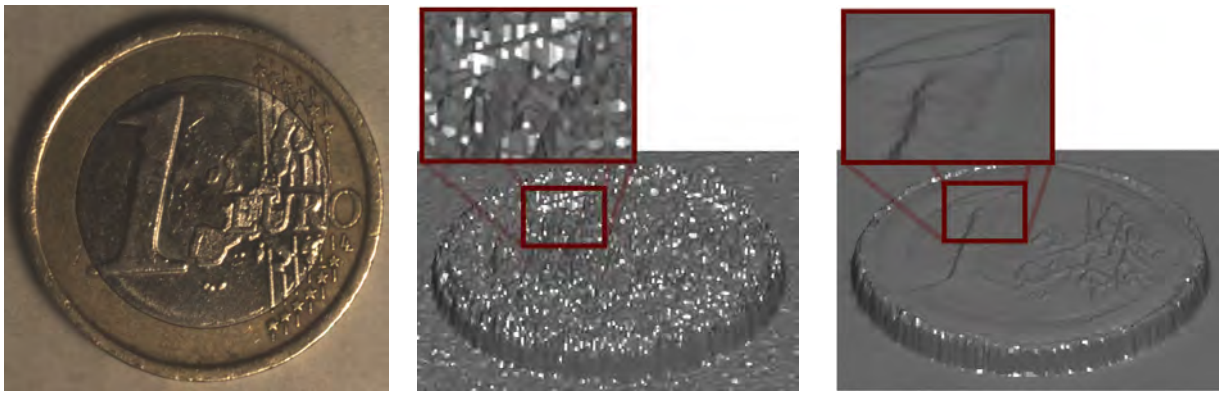
5. Conclusion and Discussion

We discussed the pros and cons of passive (light field) and active (photometric) stereo approaches for depth estimation on virtually generated data. A way to combine both approaches in order to achieve a more precise depth estimation was presented. We also showed initial results on our multi-line-scan setup, which is acquiring light field data and photometrically varying data at the same time, i.e. while observing the object under relative motion. Relative motion between the object and the acquisition device is a typical configuration in industrial vision systems, thus this setup fits well for such applications. Initial results were given and the suggested combination scheme was demonstrated. Further work will cover a more complete evaluation, as well as an improvement of the combination scheme.

References

[1] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolarplane image analysis: an approach to determining structure from motion. *Int. J. Comp. Vis.*, 1(1):7–55, 1987.

[2] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. Pat. Anal. and Mach. Intell.*, 10:439–451, 1988.



(a) Coin image (b) LF depth (c) Combined LF and PS

Figure 5: Coin image (a), acquired with our Multi-Line-Scan Setup, with an estimated depth both from light field and our improved result through combination with photometric stereo data.

[3] A. Kadambi, V. Taamazyan, B. Shi, and R. Raskar. Polarized 3D: High-quality depth sensing with polarization cues. In *Proc. Intl. Conf. Comp. Vis. (ICCV)*, 2015.

[4] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and Markus Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graphics*, 32(4):73:1–73:12, 2013.

[5] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Trans. Graph.*, 24(3), August 2005.

[6] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University, April 2005.

[7] M. Tao, P. Srinivasa, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi. Depth from shading, defocus, and correspondence using light-field angular coherence. In *Proc. Comp. Vis. and Pat. Rec. (CVPR)*, June 2015.

[8] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. PiCam: an ultra-thin high performance monolithic camera array. *ACM Trans. Graph.*, 32(5), 2013.

[9] S. Štolc, D. Soukup, B. Holländer, and R. Huber-Mörk. Depth and all-in-focus imaging by a multi-line-scan light-field camera. *J. of Electronic Imaging*, 23(5):053020, 2014.

[10] S. Wanner and B. Goldlücke. Globally consistent depth labeling of 4D light fields. In *Proc. of Comp. Vis. Pat. Rec. (CVPR)*, pages 41–48, 2012.

[11] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, July 2005.

[12] L. F. Yu, S. K. Yeung, Y. W. Tai, and S. Lin. Shading-based shape refinement of RGB-D images. In *Proc. of Comp. Vis. and Pat. Rec. (CVPR)*, pages 1415–1422, 2013.

[13] The Stanford 3D Scanning Repository, http://graphics.stanford.edu/data/3Dscanrep/.

Guided Sparse Camera Pose Estimation

Fabian Schenk[1], Ludwig Mohr[1], Matthias Rüther[1], Friedrich Fraundorfer[1], and Horst Bischof[1]

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*schenk,mohr1,ruether,fraundorfer,bischof*}*@icg.tu-graz.ac.at*

Abstract

In this paper, we present an idea for a sparse approach to calculate camera poses from RGB images and laser distance measurements to perform subsequent facade reconstruction. The core idea is to guide the image recording process by choosing distinctive features with the laser range finder, e.g. building or window corners. From these distinctive features, we can establish correspondences between views to compute metrically accurate camera poses from just a few precise measurements. In our experiments, we achieve reasonable results in building facade reconstruction with only a fraction of features compared to standard structure from motion.

1. Introduction

Structure from motion (SfM) has been an active research area in computer vision for decades as it is of interest in a wide range of practical applications such as robotic navigation and augmented reality. Common SfM approaches exploit a huge number of feature correspondences and finding suitable starting views poses a challenge, which is not necessarily simplified by the abundance of features. Often, this is tackled by assuming a set of ordered images or incorporating additional measurements for camera pose initialization. In a subsequent step, all the views are merged into a common global coordinate system, where the whole scene structure in 3D is calculated and refined together with the camera poses. The optimization of the final scene structure is computationally demanding due to the large amount of correspondences over multiple camera views.

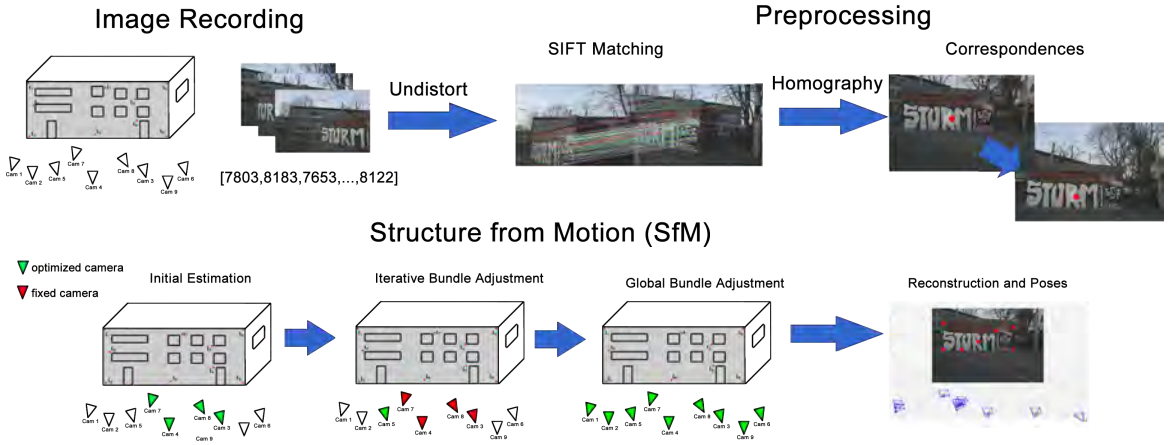


Figure 1: Our proposed sparse SfM approach, where we take RGB images and laser distance measures to estimate camera poses and a sparse point cloud.

In this work, we present an idea to estimate metrically correct camera poses with just a small number of features (see Fig. 1). Our hardware setup consists of an RGB camera and a laser range finder (LRF) (see Fig. 2 (a)). The LRF allows us to select highly distinctive features for pose estimation while at the same time obtaining their accurate distance. We focus on the reconstruction of facades, enabling us to utilize homographies instead of fundamental matrices for correspondence computation. For pose estimation, we use laser points with known distance from the camera and their respective matches in other views.

Finally, we compare our approach to the freely available SfM framework OpenMVG [10] and show that we achieve reasonable results for the camera poses with just a fraction of correspondences. This is of special interest for metric reconstruction on devices with constraints on computational power, e.g. mobile devices or UAVs.

2. Related Work

Most of the work related to the task of calibrating the extrinsic relationship of an LRF to a projective camera consider a setup with either a 2D [18, 7] or 3D LRF [14, 2]. Further, they rely on user input to establish correspondences between the laser measurements and the images taken by the camera.

We on the other hand want to solve the task of extrinsic calibration of a 1D LRF to a camera without user interaction. We require the 3D world position of the plane whose distance is measured to be inferable from the images as well as the laser point produced by the LRF to be visible within the image. In contrast to [13], where they jointly perform geometric camera and LRF calibration, we do not refine the intrinsic calibration of the camera using the LRF measurements but expect the intrinsic calibration to be done beforehand and to be of sufficient quality.

SfM algorithms for 3D reconstruction and camera pose estimation from unstructured data usually only capture the scene up to scale. In [1, 15] the authors perform large scale 3D scene reconstruction from Internet photos. Their work examines 3D modeling from unstructured data, yet the reconstruction can be only performed up to scale due to inherent lack of metric information. In [3], the authors first solve the relative motion on a local scale among just a few images, and then use these local relations as initialization for the global solution.

Methods solving the metric reconstruction problem with the SfM paradigm often rely on either an underlying structure of data (sequential image capturing, constant acquisition frame rate) in connection with registered motion estimations using GPS or inertial measurements as in [16, 3] or rely on direct geometry measurements with 3D LRFs and subsequent registration of the resulting point clouds [6, 7].

The approach presented in [12] is the one most similar to ours. However, in addition to 1D laser measurements corresponding to images of the scene, they leverage motion estimations between images through IMU data as well as interactive gestures for semantic cues aiding in reconstruction.

We propose an approach for metric camera pose estimation from unstructured images. Instead of searching for dense point correspondences among all images, we restrict ourselves to a sparse wireframe model with each image contributing just a single point (the location of the laser distance measurement). This allows us to ensure robust reconstruction by choosing distinct and easily matchable locations on the facade during data acquisition.

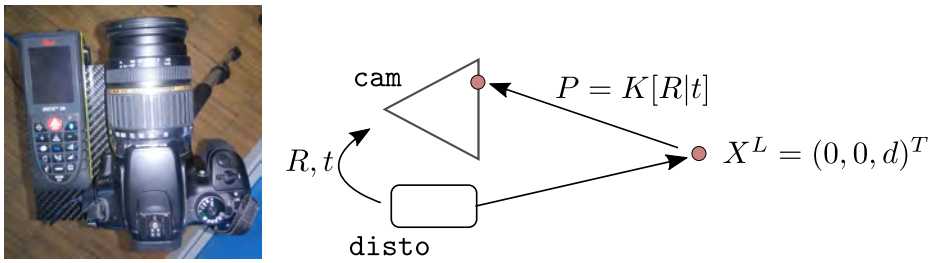


Figure 2: Left: The setup of the DSLR and laser range finder on the carbon panel. Right: The schematics of the camera and laser range finder setup and the idea behind calibration.

3. Camera Setup and Calibration

Our hardware setup consists of a standard DSLR mounted onto a rigid carbon panel next to a 1D LRF (see Fig. 2), which we control remotly via USB and Bluetooth for easy acquisition of images and distance measurements. We perform intrinsic calibration with a modified version of the Bouguet toolbox and a custom target as proposed in [4]. In the remaining part of the paper we expect the camera to be calibrated and the geometric distortions introduced by the camera and lens assembly to be removed from the images. This enables us to infer the real world line of sight $\mathbf{l}_{los}$ with respect to the center of projection of the camera of any given pixel in an image I_i as:

$$\mathbf{l}_{los,i} = \mathbf{K}^{-1} \cdot \mathbf{l}_{2D,i,i}, \tag{1}$$

where $\mathbf{K}$ denotes the camera matrix and $\mathbf{l}_{2D,i,i} = [x, y, 1]^T$ is the 2D position of the laser point i in I_i in homogeneous coordinates.

In theory, provided an intrinsic camera calibration $\mathbf{K}$ and a known plane in 3D, the rotation $\mathbf{R}_{LRF}$ and translation $\mathbf{t}_{LRF}$ of the laser range finder relative to the camera can be estimated using two measurements only. Yet, in order to obtain a more robust estimation, we take several measurements $M = \{d_i, I_i\}_1^N$ of distances d_i with corresponding images I_i. Since the application is 3D facade reconstruction and we expect the facade measurements to be taken nearly fronto-parallel to the image sensor, we restrict the extrinsic calibration sequence to a fronto-parallel movement of the target relative to the camera, ensuring that the position at which the LRF takes its measurement is well within the calibration target.

In a first step, we detect the target position and orientation in 3D relative to the camera's center of projection, as well as the target position in 2D within the image. We detect the laser point as brightest object on the calibration target using adaptive thresholding and then take its center of mass as the 2D position $\mathbf{l}_{2D,i,i} = [x_i, y_i, 1]^T$ of the laser point i in I_i. We then calculate the position $\mathbf{l}_{3D,i}$ in 3D of a laser point by intersecting the line of sight $\mathbf{l}_{los,i}$, on which the laser point lies, with the target plane.

When holding the camera positions fixed and moving the calibration target plane relative to it, all points $\mathbf{l}_{3D,i}$, $i \in \{1, \ldots, N\}$ lie on a straight line. This line corresponds to the viewing direction $\mathbf{l}_{LRF}$ of the LRF, which we calculate by fitting a line to the measurements using singular value decomposition on the 3D points stacked to a matrix $\mathbf{L}_{3D} = [\mathbf{l}_{3D,1}, \cdots, \mathbf{l}_{3D,N}]$. The right-singular vectors obtained by SVD correspond to the orthogonal directions of maximum variance within the data. Thus, the right-singular vector corresponding to the largest singular value of $\mathbf{L}_{3D}$ coincides with the viewing direction of the LRF, provided that the uncertainty of the estimation of the LRF's origin is sufficiently smaller than the relative movement of the calibration target.

We obtain several noisy estimates for the position $\mathbf{t}_{LRF}$ of the LRF through the correspondence

$$\mathbf{t}_{LRF,i} = \mathbf{l}_{3D,i} - d_i \cdot \mathbf{l}_{LRF}, \tag{2}$$

where $\mathbf{l}_{LRF}$ has been normalized to unit length. We obtain the final estimate for the position $\mathbf{t}_{LRF}$ of the LRF by taking the median of all noisy estimates. The rotation $\mathbf{R}_{LRF}$ is given by the angle between the viewing direction $\mathbf{l}_{LRF}$ of the laser range finder and the cameras optical axis in the plane spanned by the optical axis of the camera and $\mathbf{l}_{LRF}$.

4. Sparse Pose Estimation and 3D Scene Reconstruction

The proposed approach is structured in steps typical to SfM pipelines: image recording, preprocessing, relative pose and motion estimation between views and ultimately 3D reconstruction. Since it is aimed at the reconstruction of building facades, which can to a large extent be modeled as a set of flat surfaces, it is sufficient to reconstruct the building as a wire-frame model using surface vertices together with a few supporting points on the walls. We compute SIFT matches to estimate homographies between image pairs (I_i, I_j), which can be used to establish correspondences based on the known laser point $\mathbf{l}_{3D,i}$ in I_i and its respective 2D position $\mathbf{l}_{2D,i,j}$ in I_j.

Using an initial set of 4 images with full correspondences and the laser measurements, we are able to initialize and calculate an early estimate for our model and the relative camera poses. Then we iteratively add the remaining cameras and distance measurements and finally refine the poses with a global bundle adjustment. Since we know the respective distance information to each camera pose, this estimation is accurate in its scale.

4.1. Image Recording

Since we perform sparse camera pose estimation and reconstruction, the accuracy of the solution depends upon a few, yet highly significant features which are easily found in images taken from different perspectives. For a good reconstruction, the significant features should be chosen in a way such that they lie on the facade and are well-distributed over its surface including the corner points, e.g. vertices of walls and corners of windows. Figure 1 depicts the data recording process, where we take RGB images from various view points while measuring the distance of a single point in the respective image with the LRF.

4.2. Preprocessing and Feature Extraction

To keep our approach as flexible as possible and to reduce the complexity during manual data acquisition, we assume no particular order of the images. Initially all possible image pairs are added to a working set. We extract SIFT features [9] from gray-scale versions of the images and establish point correspondences using a FLANN-based matcher [11] followed by Lowe's ratio test to filter outliers. With these correspondences, we robustly estimate a homography between each image pair using RANSAC [5] with a threshold of 1 px. We only want to keep image pairs with a certain overlap in the working set, thus we filter out all with less than $n = 10$ inliers according to the RANSAC estimate and a ratio of inliers to number of matches of $< 50\%$. As a measure for the quality of the remaining image correspondences, we define an error $E_{i,j}$ for an image pair (I_i, I_j) using the 2D positions $P_{SIFT,i}$ and $P_{SIFT,j}$ of their matched features as follows:

$$E_{i,j} = mean(||P_{SIFT,i} - P_{SIFT,j}||_2), \forall i, j \in N, i \neq j. \tag{3}$$

The idea is that the Euclidean distance between SIFT matches of an image pair taken spatially closer together is lower than when taken from positions farther apart. E is used to find the images to initialize the algorithm and to find subsequent images to iteratively extend the model.

4.3. Laser Point Correspondence Computation

With the image pairs (I_i, I_j) remaining in the working set (see Sec 4.2.), we establish correspondences for the measurements of the LRF. We then compute the 2D laser point $\mathbf{l}_{2D,i,i}$ by projecting $\mathbf{l}_{3D,i}$ into its respective image using the extrinsic calibration (see Sec. 3.). As we typically deal with planar structures like facades, we estimate a homography and transform $\mathbf{l}_{2D,i,i}$ into image I_j to get $\mathbf{l}_{2D,i,j}$. This approach proves to be fairly robust in our experiments, however due to the highly repetitive nature of many facades, false positives still pose challenge.

4.4. Structure from Motion (SfM)

Our structure from motion (SfM) approach consists of three successive steps: finding an initial set for model initialization (i), iteratively adding one image at a time (ii) and one final bundle adjustment over all pairs (iii).

Model Initialization
As for all iterative SfM systems, finding a good set of starting images is challenging. Due to the many available features in common approaches, only one image pair is necessary to initialize camera pose estimations. In contrast, our approach needs a larger initial set to account for its sparse nature. Each camera has 6 degrees of freedom (3 for rotation, 3 for translation), hence we need at least 6 equations to estimate its pose. In the previous step we obtained for each image pair (I_i, I_j) two 3D-2D correspondences $\mathbf{l}_{3D,i} \Leftrightarrow \mathbf{l}_{2D,i,j}$ and $\mathbf{l}_{3D,j} \Leftrightarrow \mathbf{l}_{2D,j,i}$, i.e. 4 equations for each given image pair. For a set of at least 4 images, the resulting equation system is solvable with 6 different image pairs resulting in 4 equations each. The initial set I_{init} is chosen as the set of 4 images with the smallest sum of mutual errors $E_{i,j}$.

We solve the task of finding relative rotations $\mathbf{R_i}$ and translations $\mathbf{t_i}$ in 3D for each camera by minimizing the reprojection error $C(\cdot)$ of a laser measurement $\mathbf{l}_{3D,i}$ and its 2D correspondences $\mathbf{l}_{2D,i,j}$. with bundle adjustment. The reprojection error is defined as:

$$C(I_{init}) = min||\pi(\mathbf{R}_j(\mathbf{R}_i^{-1}\mathbf{l}_{3D,i} - \mathbf{t_i}) + \mathbf{t_j}) - \mathbf{l}_{2D,i,j}||_2^2, \forall i, j \in I_{init}, i \neq j, \tag{4}$$

with $\mathbf{R}$ and $\mathbf{t}$ the rotation and translation of each respective view and $\pi(\cdot)$ the projection. This formulation first projects a laser measurement $\mathbf{l}_{3D,i}$ in 3D from its respective camera coordinate system i into a common world coordinate system and subsequently reprojects it to the camera coordinate system j. We solve the minimization problem of bundle adjustment with a Levenberg-Marquardt [17] least-squares solver. We denote the resulting set of rotations and translations of all camera views currently involved as our current model M_{curr}.

Iterative Bundle Adjustment
In the next step, we extend our model M_{curr} by adding a new image I_k from the pool of candidates. We find I_k by summing up the error $E_{curr,k}$ of all possible image pairs (I_{curr}, I_k) and take the one with the most correspondences and the lowest overall error. We also set the initial rotation $\mathbf{R}_k$ and translation $\mathbf{t}_k$ of the newly added camera equal to the parameters of the closest camera, i.e. the one with the lowest $E_{i,k}, i \in M_{curr}$. For each image pair, i. e. 3D-2D correspondence, we get two

independent equations for the x- and y-position, thus we need at least 3 correspondences to solve the 6 degrees of freedom given by $\mathbf{R}_k$ and $\mathbf{t}_k$. As mentioned in Section 4.3., outliers (wrong matches) are possible, thus in practice we use at least 4 correspondences.

In a first step, we compute the reprojection error of all correspondences C (4) using the initialized camera pose and the mean reprojection error $\bar{C}$. Then we take all correspondences with an error smaller than $1.5\bar{C}$ or a threshold ϵ_C. We then optimize with the same cost function (4) as for the initial bundle adjustment with the major difference that only the pose of the newly added camera k is optimized, while the rest of the system M_{curr} is fixed. After optimization, we again filter bad correspondences with the same approach as described above and perform a second optimization, which is usually very fast due to the already good estimation. We iterate through all images until no more can be added, i.e. do not fulfill any of the conditions.

Global Bundle Adjustment
While keeping the camera system M_{curr} fixed and only optimizing the new camera k is very fast and gives an estimate of the model structure, it does not replace a global optimization approach. We perform a final global bundle adjustment step, where all the camera poses are optimized. In this case, we take all the correspondences used during the iterative bundle adjustment step and initialize the camera poses with the previously computed rotation and translation. Here, we also use the cost function presented in (4). Similar to the iterative bundle adjustment, we again filter outliers with the reprojection error after optimization, but instead impose that the error must be smaller than $< 1.2\bar{C}$. After this final filtering step, we perform one last global bundle adjustment.

5. Results and Discussion

In this section, we present early results of our guided sparse camera pose estimation. We evaluate our approach on two datasets from different buildings, one with a well-textured facade and one with redundant structures. As reference, we use the open-source SfM framework OpenMVG, which can achieve an accuracy of around 1 cm in ideal cases [10]. It utilizes SIFT features and many correspondences to estimate camera poses and a point cloud. OpenMVG chooses the starting views randomly and in our evaluation we had to start SfM multiple times to get a reconstruction. We evaluate the distance between the camera centers generated by the two approaches. OpenMVG estimates the reconstruction up to scale, thus to metrically measure the distance between cameras, we transform its world coordinate system to ours using a robust similarity transform.

Figure 3 shows a histogram, where the bins show the number of distances between camera centers in the respective range in cm. Cameras in the first few bins are closer to OpenMVG, while the cameras in the last bin are farther away. Especially in the first experiment we achieve reasonable results and that with only 90 correspondences compared to OpenMVG's 2969, which is a reduction by a factor of 30. In the second experiment, we only use 56 correspondences compared to 1880 in the reference. The histograms show that we are centimeters away from OpenMVGs reconstruction even though we still achieve visually appealing results when reprojecting the laser points into the images (see Fig. 4). Due to the sparse correspondences, even one unfiltered outlier can decrease the final result significantly.

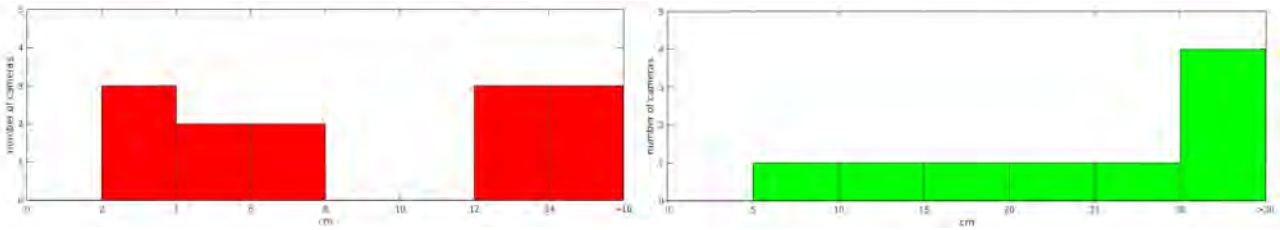


Figure 3: The two experiments and the distances between our camera positions and OpenMVG's. The bins of the histogram show the number of distances between camera centers in the respective range [cm].

Figure 4: The reprojected laser point from the computed camera poses.

6. Conclusion and Outlook

In this paper, we presented a first evaluation of our idea to utilize a combination of RGB camera and LRF for sparse camera pose estimation, where we can select significant features during the image recording process with the LRF. We showed that metrically accurate camera pose estimation with just a few correspondences is possible. In future work, we plan a more sophisticated way to redetect significant features in the other views without the use of SIFT matches and homography estimation, which would also enable the reconstruction of more complex 3D structures. Additionally, we plan to improve the accuracy by robustifying our approach against outliers during bundle adjustment.

An interesting direction for future work is inspired by Li et al. [8], where they address the SfM problem by estimating up-to-scale edge lengths of a rigid graph constructed from 3D features and their respective image rays. We would like to investigate whether the laser range measurements can be directly used as edge lengths for this approach.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, pages 72–79, Sept 2009.

[2] H. Aliakbarpour, P. Nunez, J. Prado, K. Khoshhal, and J. Dias. An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance. In *ICAR*, pages 1–6, June 2009.

[3] H. Aliakbarpour, K. Palaniappan, and G. Seetharaman. Fast structure from motion for sequential and wide area motion imagery. In *Computer Vision Workshop (ICCVW)*, pages 1086–1093, Dec 2015.

[4] David Ferstl, Christian Reinbacher, Gernot Riegler, Matthias Rüther, and Horst Bischof. Learning depth calibration of time-of-flight cameras. In *BMVC*, pages 102.1–102.12. BMVA Press, September 2015.

[5] 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*, 24(6):381–395, 1981.

[6] Christian Früh and Avideh Zakhor. An automated method for large-scale, ground-based city model acquisition. *IJCV*, 60(1):5–24, 2004.

[7] Ji Hoon Joung, Kwang Ho An, Jung Won Kang, Myung Jin Chung, and Wonpil Yu. 3d environment reconstruction using modified color icp algorithm by fusion of a camera and a 3d laser range finder. In *IROS*, pages 3082–3088. IEEE, 2009.

[8] Hongdong Li. Multi-view structure computation without explicitly estimating motion. In *CVPR*, pages 2777–2784. IEEE, 2010.

[9] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[10] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*, pages 3248–3255, Dec 2013.

[11] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2:331–340, 2009.

[12] Thanh Nguyen, Raphael Grasset, Dieter Schmalstieg, and Gerhard Reitmayr. Interactive syntactic modeling with a single-point laser range finder and camera. In *ISMAR*, pages 107–116. IEEE, 2013.

[13] Thanh Nguyen and Gerhard Reitmayr. Calibrating setups with a single-point laser range finder and a camera. In *IROS*, pages 1801–1806. IEEE, 2013.

[14] D. Scaramuzza, A. Harati, and R. Siegwart. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. In *IROS*, pages 4164–4169, Oct 2007.

[15] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008.

[16] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *ICCV*, pages 65–72, Dec 2013.

[17] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment-a modern synthesis. In *Vision algorithms: theory and practice*, pages 298–372. Springer, 1999.

[18] Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *IROS*, volume 3, pages 2301–2306. IEEE, 2004.

WS 4: Tracking / Detection

Explaining Point Cloud Segments in Terms of Object Models

Manuel Lang[1] and Justus Piater[1]

Intelligent and Interactive Systems
Institute of Computer Science
University of Innsbruck, Austria
{*csae6836,justus.piater*}*@uibk.ac.at*

Abstract
Segmenting the signal of a 3D-sensor represents a core problem in computer vision. Describing segments at the object level is a common requirement for higher-level tasks like action recognition. Non-parametric techniques can provide segmentation without prior model information. However, they are also prone to over- and under-segmentation, especially in case of high occluded scenes. In this paper we propose an approach to segmenting a 3D scene based on a set of known object models. Six-degree-of-freedom (6DOF) model poses result from recognition and pose estimation by exploiting distinct object shapes acquired from a non-parametric segmentation stream. The aligned object models are used in order to resolve over- and under-segmentation by following a bottom-up strategy. Segmentation refinement results from contracting and subdividing input segments in accordance to aligned object models. The proposed algorithm is compared to a trivial model-based segmentation approach that neglects the segmentation stream. Both approaches are evaluated on a set of 24 scenes which are divided into four different complexity categories. The complexity of the scenes ranges from simple to advanced, objects are placed in sparse configurations as well as highly occluded compositions.

1. Introduction

Describing point cloud segments at the object level is of significant importance in the area of computer vision. Having a mechanism that allows to discriminate between individual objects in a captured scene can be useful for higher-level tasks like action recognition, planning and execution [2, 18]. Depth information can provide valuable cues for tasks like segmentation, recognition, pose estimation and tracking [1, 3, 6, 16]. A major challenge is to apply recognition and pose estimation in occluded environments, where scenes are captured by low-resolution RGBD-sensors. This work concentrates on recognition, pose estimation and segmentation of known objects which are part of an assembling task. The objects are placed in table-top scenes that are captured by a Kinect sensor.

The main contribution of this paper can be summarized as follows. Starting from a given model-free[1] segmentation input stream, we propose to execute segment-based object recognition and pose estimation by following a bottom-up strategy. We present a combined recognition, pose estimation and segmentation workflow that exploits geometrical cues delivered by the segments that are computed

[1]In the context of this paper the term *model-free* means that the underlying process does not rely on object models that have to be specified by a supervisor.

by a model-free segmentation process. The complexity of the recognition task is reduced stepwise by handling large segments before small segments. The input segmentation is refined iteratively by exploiting collected 6DOF model pose information. Recognition and pose estimation rely on object models that are specified by 3D meshes as shown in figure 1. Object recognition is bound to certain time constraints, therefore the proposed algorithm does not execute in real-time. The utilized segmentatic
rest
by t
essa
inpu

Figure 1: The set of object models that are used for recognition and pose estimation.

2. Related Work

Exploiting low-level processing outcomes in higher-level tasks is a fundamental paradigm in computer vision [4, 8, 19]. At present, there exist many segmentation methods that apply to RGBD data [1, 7, 13, 9]. Global surface descriptors are commonly applied to pre-segmented scenes [4]. In this paper we concentrate on local descriptors [5]. The latter type is more suitable for our dataset, since it is more robust against clutter and occlusion. Model information is frequently used for object tracking in videos. The method proposed in [14] uses model information to track 6DOF poses. A RGBD-based segmentation and tracking approach that uses adaptive surface models is proposed in [10]. Our approach concentrates on a combination of object recognition, pose estimation and segmentation in RGBD-images.

3. Background

The following sections provide information about the methods that have been utilized in this paper. Recognition and pose estimation is addressed in the subsequent section 3.1.. Section 3.2. introduces a method that delivers model-free segmentation. The model-based point cloud segmentation that is described in section 3.3. acts as a baseline for the bottom-up segmentation approach proposed in section 4.2..

3.1. Point-based Object Recognition and Pose Estimation

At present, there exists a large variety of different object recognition and pose estimation approaches. An appropriate method should be robust against noise which is introduced by the sensor and it should provide reliable results even in the case of occluded scenes. Scenes are captured by a depth-sensor and object models are represented as point clouds that are sampled from 3D meshes. The method used in this paper estimates 6DOF poses by applying a point-based recognition pipeline [3]. The pipeline is publicly available as part of the Point Cloud Library (PCL) [16]. Figure 2 shows the single steps that are executed in order to recognize the objects in the scene. The first stage extracts keypoints from

model and scene point clouds. In general, keypoints are defined by detecting characteristic surface points. A simple and efficient alternative is to sample keypoints uniformly from the surface. The local geometry of each keypoint is described by the *Signature of Histograms of Orientation* (SHOT) descriptor [17], which delivers favorable results for the evaluated dataset. PCL provides a variety of different descriptor implementations. A comprehensive comparison can be found in [5]. Correspondences are generated by matching scene descriptors against a database of offline computed model descriptors. The next step clusters geometrically consistent correspondences into groups. Starting from a seed correspondence $c_i = \{p_i^m, p_i^s\}$ (p_i^m and p_i^s denote corresponding key points of model and scene), geometrical consistency follows from the following relation

$$|||p_i^m - p_j^m||_2 - ||p_i^s - p_j^s||_2| < \varepsilon \tag{1}$$

where ε defines a distance threshold between the keypoints. A minimum of three correspondences is required to estimate a 6DOF pose. The absolute orientation step eliminates correspondences that are not consistent with a unique 6DOF pose. The utilized recognition pipeline provides an optional iterative closest point (ICP) refinement step, which can be applied on the recognized hypotheses. The number of ICP iterations has been set to a low value. Running more than 5 ICP iterations on the given dataset does not result in significant recognition improvements. The final hypothesis verification step determines a set of non-conflicting model hypothesis that are in accordance with the scene point cloud. Hypothesis that result from unexpected objects within the scene have to withstand the following quality measurement. An acceptance function evaluates the number of supported model points that are close to scene points, as well as the number of unsupported model points (visible model points that have no counterpart in the scene). A detailed description of the hypothesis verification algorithm that has been utilized in this paper is given in [12].

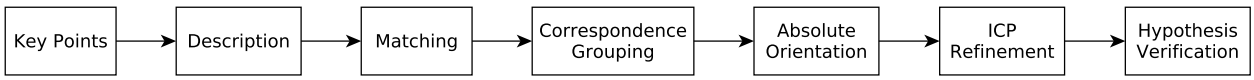


Figure 2: Recognition pipeline used in this paper.

3.2. Model-Free Point Cloud Segmentation

Segmentation results from summarizing interesting and distinguishable image properties. Higher-level visual tasks like object recognition and pose estimation can benefit from such condensed image representations. A method that segments the signal of a RGBD-sensor, without explicit object model information has been presented in [1]. Homogeneous regions (segments) are generated by using color information. In addition, the method exploits depth information in order to support the segmentation and tracking process. Figure 3 shows two example scenes that have been segmented by this method. The segmentation result depends on several factors like scene density, degree of occlusion, object geometry, light conditions, etc.

Figure 3: Point cloud segmentation generated by a color-based model-free method.

3.3. Model-based Point Cloud Segmentation

A trivial model-based point cloud segmentation results from evaluating the point vicinity of recognized object models. Object models are aligned with the scene point cloud by applying point-based methods, as described in section 3.1.. It is reasonable to assume that a model point that is close to a scene point indicates a *model-explained* segment membership of this point. Spatial decomposition techniques such as kd-trees provide an efficient structure to determine the k closest points of a query point [15]. The set of scene points that are explained by an aligned object model results as follows. Each point that has been sampled from the model point cloud defines k nearest neighbors (kNN) in the scene point cloud. The nearest neighbor search is carried out in a kd-tree, which represents the scene point cloud. Choosing the value for k results in a trade-off between segment density and sharpness of the segment edges.

4. Segment-based Object Recognition and Pose Estimation

Rising the degree of occlusion in a scene inevitably complicates the segmentation process. Nevertheless, the set of regions that result from the model-free segmentation method described in section 3.2. can preserve a certain amount of object characteristics, even in the occluded case. This motivates a segment-based recognition and pose estimation approach where the model-free segmentation acts as main input. Single segments like the one shown in figure 3 are often not expressive enough to apply recognition and pose estimation on them. Many of them show less variation in surface-normal orientation. We propose to generate larger surface patches in order to increase the recognition output. Surface patches are created by clustering a set of adjacent segments together. Figure 4 provides an overview of how segment-based model poses are generated iteratively in order to refine model-free segmentation in a bottom-up way. In the rest of this paper, the terms surface patch and segment are interchangeable, since single segments can also act as simple surface patches.

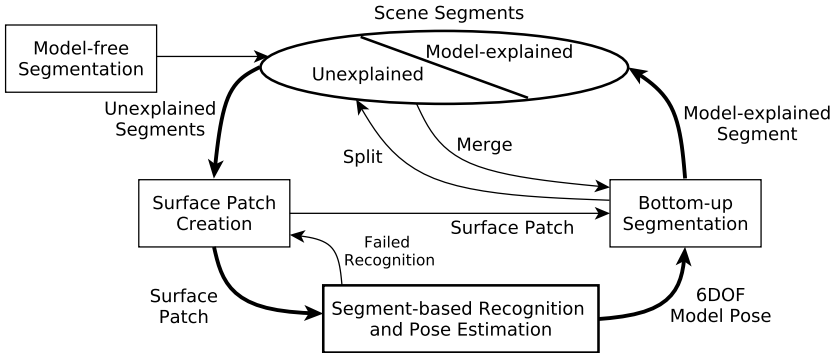


Figure 4: Iterative application of segment-based object recognition and pose estimation.

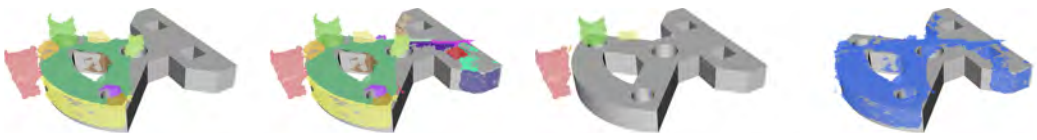
4.1. Adaptive Correspondence Grouping

The correspondence-based recognition and pose estimation method that has been introduced in section 3.1. searches for a set of non-conflicting hypotheses that describe the whole scene at once. In contrast, we propose a segment-based bottom-up strategy. This approach is motivated by two considerations. Firstly, restricting recognition and pose estimation to a surface patch, that preserves certain object characteristics, could reduce the number of wrong hypotheses. Secondly, following a bottom-up strategy that handles large surface patches early, reduces the complexity of the recognition task for smaller segments. The latter consideration is gaining relevance if the scene is a composition of

large and small objects. The proposed approach utilizes the recognition pipeline shown in figure 2. Model hypothesis are computed from consistent correspondence groups, as described in section 3.1.. However, in this case the proposed algorithm adapts the number of correspondences that are required to form a consistent group. According to [3] the correspondence grouping threshold trade-offs the number of correct recognition for the number of wrong recognitions. In general the size of the group can range between three (the minimum required to compute a 6DOF pose) and the number of correspondences that are found in total. A high threshold generates few hypotheses whereas a low threshold leads to many hypotheses. An optimal threshold is influenced by many factors like surface patch size, level of over- and under-segmentation, object similarity, object geometry and also the noise-level of the 3D-sensor. We propose to adapt the correspondence grouping threshold in accordance to the hypothesis verification process which is the last stage in the recognition pipeline shown in figure 2. Starting from a large value the correspondence grouping threshold is reduced stepwise until at least one hypothesis survives the verification process. If the threshold falls below the absolute minimum of three, recognition fails. The acceptance function of the hypothesis verification process also offers opportunities for a segment-based parameter tuning. The thresholds for the number of supported and unsupported scene points, as described in section 3.1., can be weaken if the surface patch size exceeds a certain threshold. This adjustment is justifiable since large surface patches commonly generate fewer hypothesis that are more discriminable.

4.2. Bottom-up Segmentation

The basis for the bottom-up segmentation process is a 6DOF model pose that results from segment-based object recognition and pose estimation. In contrast to the trivial model-based segmentation process that has been described in section 3.3., we propose a recycling of the model-free segmentation stream. According to figure 4, model-free (unexplained) segments are merged and splitted in accordance to the recognized object model that has been placed at the estimated pose. The segmentation process can be described as follows: If recognition fails surface patch creation restarts with the next largest segment. In case of successful recognition, the initial surface patch is extended with parts of unexplained segments that are covered by the aligned object model. Covered segment parts are determined by applying a segment-based radius search in a kd-tree, similar to the approach described in section 3.3.. The search radius is set to a fraction of the object model size. Surface patch parts that are not covered by the recognized object model are separated from the current surface patch and fed back into the recognition process. The process restarts until each unexplained segment becomes part of a model-explained segment or gets labeled as unrecognizable. The single steps of the segmentation process are shown in figure 5. Figure 5a shows the recognized object model that has been aligned with the initial surface patch. Figure 5b shows the extension of the initial surface patch. The separation of non-covered segment parts is shown in 5c. The final result of the model-explained segment can be seen in figure 5d.



(a) (b) (c) (d)

Figure 5: Bottom-up segmentation. (a) Object model aligned with surface patch. (b) Merging of covered segments. (c) Splitting of non-covered (unexplained) segments. (d) Final segmentation result.

5. Results

The proposed algorithms have been evaluated on 24 different scenes which are divided into four complexity categories. The first category contains simple object compositions where objects are widely spread over the field of view. Category two consists of dense scenes that are commonly under-segmented. The third category contains disordered scenes, showing a high degree of clutter. The last and most challenging category contains objects that are assembled together, which results in a high degree of occlusion. Figure 7 shows an instance of each category. The proposed segment-based bottom-up approach is compared to the point-based method that has been described in section 3.3.. The algorithms have been tested on an Intel(R) core(TM) i5 2.53GHz CPU (multiple cores) with 7.7GB RAM. The average scene execution time[2] of the segment-based approach is 154.38 seconds. The point-based approach executes in 129.37 seconds.

5.1. Recognition Rate

Table 1 summarizes the recognition results of the object models shown in figure 1. As shown in the table, segment-based adaptive correspondence grouping (CG) outperforms the point-based method for almost all models. Figure 6 shows a more detailed comparison between all four evaluated complexity categories. The low *bolt sensor* rating is caused by the segment-based parameter tuning of the hypothesis verification process that has been discussed in section 4.1.. In this case, the verification process eliminates too many reasonable hypotheses, which finally leads to confusion with similar looking *bolt angular* and *shaft* objects.

method / **model**	point-based CG	segment-based CG
faceplate	91.67	97.92
separator	83.33	100.00
pendulum	75.00	87.50
shaft	95.83	100.00
bolt angular	79.17	85.42
bolt sensor	75.00	60.42
pendulum head	83.33	95.83
average	82.92	**87.08**

Table 1: Recognition rate comparison of the point-based baseline method and the proposed segment-based method. CG - Correspondence Grouping

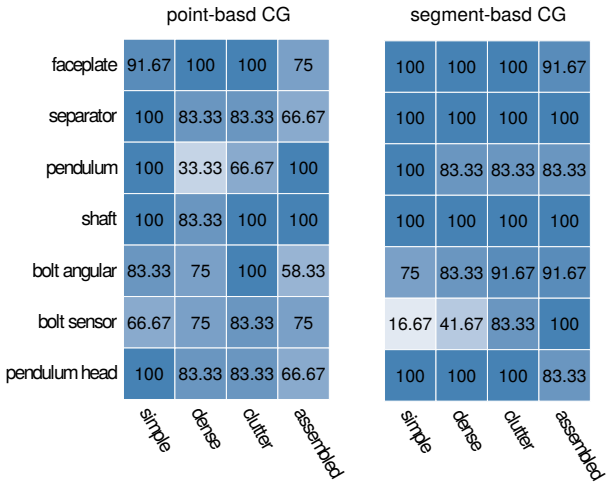


Figure 6: Recognition rate comparison between four evaluted scene complexities.

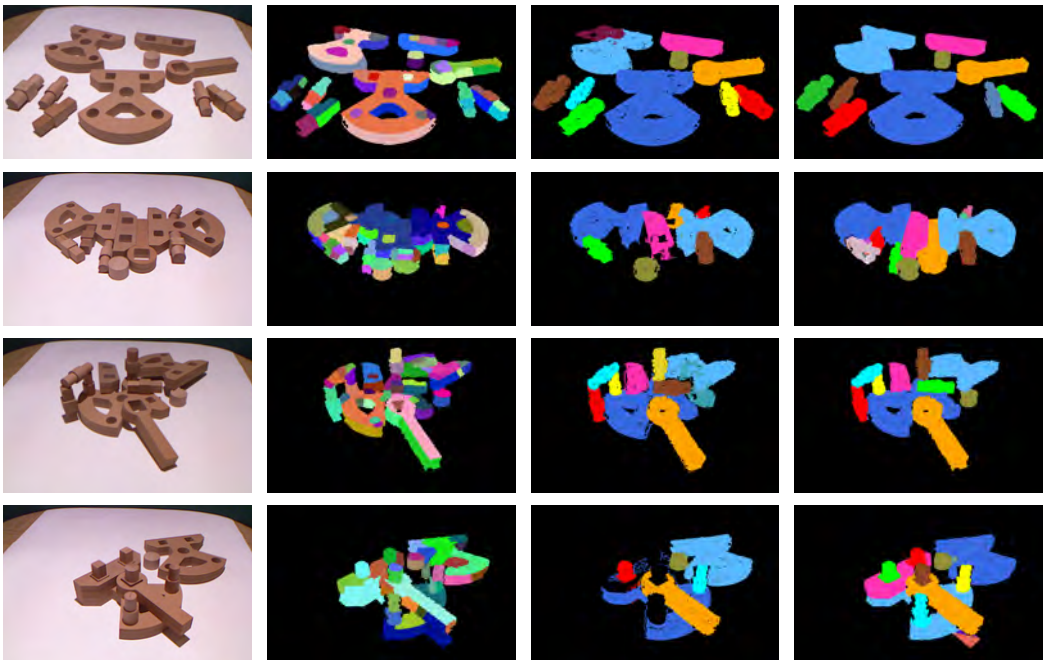
5.2. Segmentation

Figure 7 shows a segmentation comparison of four selected scenes. As shown in the image, the segmentation quality strongly depends on the accuracy of the estimated model poses. Object confusion impairs the segmentation result. The bottom-up segmentation benefits from the recycling of model-free segments. The segment recycling results in sharper edges when compared to the trivial model-based segmentation method. The destructive characteristic of the model-based segmentation results from an inherently trade-off between sharp segment margins and segment density.

[2]The real-time model-free segmentation process, which is not part of this evaluation, relies on a GPU-based system.

6. Conclusion

We have presented a segment-based object recognition and pose estimation approach. The proposed bottom-up segmentation strategy reduces the complexity of the recognition task in an iterative way. The geometrical cues of the model-free input segmentation can successfully be exploited in order to improve recognition rates in occluded scenes. The proposed segment-based recognition and pose estimation approach relies on correspondence-based recognition. False hypothesis are suppressed by adapting the cardinality of consistent correspondence groups. The estimated 6DOF pose information can effectively be used in order to resolve over- and under-segmentation of the model-free input stream. The suitability of our approach was demonstrated on 24 scenes. The complexity of the evaluated dataset reaches its maximum in assembled object compositions. The efficiency of the segment-based object recognition and pose estimation is bound to the amount of under-segmentation in the surface patch.



(a) (b) (c) (d)

Figure 7: Segmentation comparison. (a) RGB input. (b) Model-free segmentation. (c) Model-based segmentation (baseline). Unrecognizable objects are colored black. (d) Bottom-up segmentation.

References

[1] A. Abramov, J. Papon, K. Pauwels, F. Wörgötter, and B. Dellen. Depth-supported real-time video segmentation with the kinect. In *IEEE workshop on the Applications of Computer Vision WACV*, 2012.

[2] Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, KeJun Ning, Babette Dellen, and Florentin Wörgötter. Learning the semantics of object-action relations by observation. *I. J. Robotic Res.*, 30:1229–1249, 2011.

[3] Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze. Tutorial: Point cloud

library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robot. Automat. Mag.*, 19:80–91, 2012.

[4] Aitor Aldoma, Markus Vincze, Nico Blodow, David Gossow, Suat Gedikli, Radu Bogdan Rusu, and Gary R. Bradski. Cad-model recognition and 6dof pose estimation using 3d cues. In *ICCVW*, 2011.

[5] Luís A. Alexandre. 3d descriptors for object and category recognition: a comparative evaluation. In *in Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[6] Renaud Detry and Justus H. Piater. Continuous surface-point distributions for 3d object pose estimation and recognition. In *ACCV*, 2010.

[7] Aleksey Golovinskiy, Thomas Funkhouser, and Traac Light Car. Min-cut based segmentation of point clouds. 2009.

[8] Aleksey Golovinskiy, Vladimir G. Kim, and Thomas A. Funkhouser. Shape-based recognition of 3d point clouds in urban environments. In *ICCV*, 2009.

[9] Steven Hickson, Stan Birchfield, Irfan A. Essa, and Henrik I. Christensen. Efficient hierarchical graph-based segmentation of rgbd videos. In *CVPR*, 2014.

[10] Farzad Husain, Babette Dellen, and Carme Torras. Consistent depth video segmentation using adaptive surface models. *IEEE T. Cybernetics*, 45:266–278, 2015.

[11] Ajmal S. Mian, Mohammed Bennamoun, and Robyn A. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:1584–1601, 2006.

[12] Chavdar Papazov and Darius Burschka. An efficient ransac for 3d object recognition in noisy and occluded scenes. In *ACCV*, 2010.

[13] Jeremie Papon, Tomas Kulvicius, Eren Erdal Aksoy, and Florentin Wörgötter. Point cloud video object segmentation using a persistent supervoxel world-model. In *IROS*, 2013.

[14] Karl Pauwels, Leonardo Rubio, Javier Díaz, and Eduardo Ros. Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues. In *CVPR*, 2013.

[15] Radu Bogdan Rusu. Semantic 3d object maps for everyday manipulation in human living environments. In *DE*, 2009.

[16] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *ICRA*, 2011.

[17] Samuele Salti, Federico Tombari, and Luigi di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014.

[18] Emre Ugur and Justus H. Piater. Refining discovered symbols with multi-step interaction experience. In *HUMANOIDS*, 2015.

[19] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

WS 5: Vision for Robotics I

Real-time tracking of rigid objects using depth data

Sharath Chandra Akkaladevi[1, 2], Martin Ankerl[1], Gerald Fritz[1], Andreas Pichler[1]

[1]Department of Robotics and Assistive Systems
Profactor GmbH, Im Stadtgut A2, Steyr-Gleink, 4407, Austria
{firstname.lastname}@profactor.at

[2]Institute of Networked and Embedded Systems
Alpen-Adria-Universität Klagenfurt, Austria

Abstract
In this paper, a robust, real-time object tracking approach is presented. The approach relies only on depth data to track objects in a dynamic environment and uses random-forest based learning to deal with problems like object occlusion and clutter. We show that the relation between object motion and the corresponding change in its 3D point cloud data can be learned using only 6 random forests. A framework that unites object pose estimation and object pose tracking to efficiently track objects in 3D space is presented. The approach is robust against occlusions in tracking objects and is capable of real-time performance with 1.7ms per frame. The experimental evaluations demonstrate the performance of the approach against robustness, accuracy and speed and compare the approach quantitatively with the state of the art.

1. Introduction

Object tracking has been widely researched in the vision community over the recent past and many methods are proposed in literature to track objects [6]. Until the last decade the methods mainly considered 2D image data as input and in some cases stereo vision and served applications like surveillance, military use, security and industrial automation. However, 2D image data only captures the 3D projection into two dimensions and is sensitive to illumination changes. With recent development of RGB-D devices like Kinect, researchers all over the world are exploiting depth data for object recognition and tracking [7]. Tracking can be defined as the problem of estimating the trajectory (6 DOF – 3 translations, 3 rotation parameters) of an object in the 3D image plane as it moves around a scene. Though there has been a lot of work in tracking humans using RGB-D devices [8], not much work is done in the field of tracking objects that could be used in industrial settings which often have real-time requirements.
Object tracking in general is a challenging problem. Tracking objects becomes difficult due to abrupt object motions, object to object occlusions, clutter, camera motion and noisy sensor data. When considering its application in industrial settings the problem of designing a successful tracking algorithm becomes even more difficult. This is due to the requirement of higher levels of robustness, accuracy and speed. Also, industrial objects tend to have little texture. In this paper, we describe an approach for real-time tracking of objects [12] that aims to answer these challenges. The main contribution of this paper is the extended evaluation of the work in [12] and its comparison with the state of the art approaches.

Inspired by [5] we describe a fast and accurate 3D object tracking algorithm for rigid objects. The proposed approach is model-based, uses only depth data and achieves very good accuracy utilizing a framework that combines object localization and object tracking.

2. State of the Art

With the introduction of RGB-D sensors like Kinect, various approaches for object tracking in 3D were proposed, which ranged from tracking humans [10], hand tracking [8], and tracking rigid and non-rigid objects. For a better comparison of our approach with the state of the art, the scope of this section is limited to approaches that focus on frame-to-frame tracking of rigid objects using RGB-D data. The proposed approaches can be broadly classified into two categories: a) approaches that are based on 3D models and b) approaches that do not assume pre-defined object models.

For example, an approach that does not rely on prior knowledge of the target object representation is described in [14]. The approach uses adaptive Gaussian Mixture Models (GMM) to represent multiple objects that move independently. The object model is updated incrementally at each time instant with the help of the feedback results from the robust tracking process. To correct falsely detected objects in presence of occlusions and various types of interactions among multiple objects, an approach that exploits component-level spatiotemporal association is proposed in [10]. However, the approaches of "individuation-by-feature" [14] and "individuation-by-location" [10] require high computation time to learn each object model at every time instant and would exponentially increase with the number of objects and their spatial relations. Moreover, in an industrial environment which involves human actions, situations keep changing every time instant. To achieve robustness and computational efficiency in such scenarios, applying one individuation method is not sufficient. To alleviate this problem, an approach that determines individuation strategy (by location and/or by feature) depending on the object situation is proposed in [15]. The main assumption of the approach is that falsely segmented objects can be detected and rectified using both location and position information. It also assumes that objects do not change substantially in terms of shape or position from one frame to the other. A probabilistic framework for simultaneous tracking and reconstruction of rigid 3D objects using RGB-D sensor is proposed by [7], where the probabilistic method is used to statistically determine occlusions. Intensity images are used to model appearance of an object while modeling occlusions.

With the availability of reliable, fast and simple object reconstruction solutions like ReconstructMe[1], 3D object models can be obtained in real-time. A popular approach for model-based object tracking is based on the particle filters [10][18]. For example, the authors in [4] propose a 3D model-based visual tracking approach using edge and keypoint features in a particle filtering framework. This approach does not assume the initial pose of the object. It uses given 2D-3D keypoint correspondences to calculate a set of possible pose hypothesis of the object. Once the intial pose is estimated, edge points are used to track movement of the object from frame-to-frame. This approach is extended by [5] where an RGB-D object tracking method using a particle filter on GPU is proposed.

Another popular method for 3D object tracking is the Iterative Closest Point (ICP) approach which has many variants [16]. The algorithm uses a set of initial parameters and refines them iteratively to reach a set of optimal parameters by minimizing the object function. This approach has problems in dealing with occlusions and object clutter, which result in a local-minimum. To overcome this problem, a model-based learning approach is proposed in [18]. This approach learns the relation between the parameters that induce object' motion and the change they induce on the 3D point cloud using random forests. In order to track the object in motion, the change in the 3D depth data

[1] ReconstructMe http://reconstructme.net

is used to predict the parameters of this motion. The advantage of using random forests is that it is a collection of trees that learn and predict independently, even when some input data is affected due to occlusions, other trees can still provide good predictions. In order to track objects in different views, [18] trains a random forest for multiple views of the object that leads to a high computational effort. Moreover, the approach is not suitable for tracking symmetrical objects as the multiple-pose hypotheses are averaged and this leads to erroneous tracking of symmetrical objects. An offline learning based approach with known 3D object models based on particle filters is proposed in [9]. In [20], the authors propose a learning based approach inspired by [18] with reduced computational cost and improved occlusion handling capability.
In the proposed approach, we make the following contributions: a) we argue that it is sufficient to train only 6 random forests, to learn the relation between object motion and its corresponding change in 3D point cloud data, which in turn reduces the computational complexity b) dealing with symmetrical and non-symmetrical objects and c) a framework that is capable of tracking objects in presence of partial occlusions. A quantitative comparison is also carried out in this paper that uses synthetic data (that includes ground truth) provided by [5] to compare our approach against the state of the art.

3. Method

This section illustrates the proposed approach for localizing and tracking 3D objects with high performance and accuracy. First we describe the global localization algorithm RANGO, followed by the local tracking algorithm. Then, we illustrate how both components are combined into the full tracking framework

3.1. RANGO – RANdomized Global Object localization

RANGO is an algorithm for 3D object localization. It is based on a random sampling algorithm (RANSAC) described in [1][3] with several performance and robustness improvements, allowing a very fast detection rate when compared to the registration approach proposed in [7]. Its main contribution is the replacement of K-nearest neighborhood search for inlier detection with a probabilistic grid based approach. Thus the time complexity for the evaluation of a hypothesis (acceptance function) is reduced from $O(n * log(m))$ where n is the number of model points, m denotes the number of points in the scene, to $O(n)$. Additionally, the evaluation of the number of model points that fit the hypothesis is stopped early when the probability of finding a good match is too low.
Sparse 3D Voxel Grid. Each 3D point of a scene is approximated into a sparse axis aligned 3D grid. Each voxel of this grid is defined by a (x, y, z) tuple where x, y, z are (integer) coordinates for the voxel location. In RANGO this (x, y, z) position is hashed into a single 32bit number which is used as an index in a hash table. Due to hashing collisions it is possible that two different points hash to the same voxel even though their position is unrelated, but the probability is low enough that it is not a problem for our use case. This 3D voxel grid is then used for fast verification of candidate transformations.
To evaluate a transform matrix, we iterate over a set of sample points of the model and transform them into the scene. Each sample point is hashed into the 3D voxel grid containing scene points. If the hashed voxel is filled with a point and has a similar normal vector orientation as the model point, we count that as an inlier. This verification method has a complexity of $O(m)$ where m is the number of sample points. This verification is only approximate as it is possible to miss a neighboring sampling point because we only lookup the voxel the sample point hashes to, ignoring neighboring

voxels. A kd-tree would allow for exact nearest neighbor queries, but would have $O(m * log(n))$ complexity. The speedup achieved by the $O(m)$ verification allows us to evaluate more candidate transformations at the same time to boost accuracy.

Filtering Candidate Solutions. After the random sampling and matching process is over, the candidate solutions are filtered, since it is likely that multiple similar solutions have been found. In [3] a pose clustering approach is used. The pose clustering combines multiple similar poses to find an average position from these candidate solutions. This approach falls short for symmetric objects. E.g. a sphere where the reference frame is off center will result in many different potential poses, but each pose will have a completely different translation and rotation. In RANGO, we have replaced the clustering approach with a filtering approach. All candidate solutions are sorted by the number of inliers, highest number first. Iteratively, each solution is re-checked if it meets a given inlier threshold, and if it does, all scene voxels that were used in this inlier check are removed from the 3D voxel grid. This way only the best fitting candidate solution for a potential pose is used, while it is still possible to find multiple instances of the same object in the scene data. This approach works well for both complex and symmetrical objects.

3.2. Multi-Forest Tracking

Our multi-forest tracking approach is a variation of the multi-forest tracking algorithm described in [18]. Our modification to this algorithms retains the performance characteristic of [18] while having a significantly lower memory and training overhead, which allows the use of this algorithm on devices with limited computational power such as a tablet pc. It is noteworthy that we only use depth data for both tracking and object localization. The reason is that our main goal is to be able to robustly track industrial parts, and these usually do not carry much color information.

Single-view-Tracking. The multi-forest tracker described in [18] uses $6 * n_c * n_t$ random forests, where the number of dimensions to represent a pose is 6, n_c is the number of camera positions and n_t the number of trees in each forest. For each camera position sample points of the objects are extracted and used to train 6 random regression forests for tracking of this camera view. An algorithm switches between the camera views that are currently best visible. They parameterize this as $n_c = 42$ and $n_t = 100$ resulting in 25200 random trees. Each tree is generated from a test set of 50000 samples, resulting in a significant training effort. In our approach we have reduced this effort significantly to only $6 * n_t$ trees. After the samples have been generated, each random forest is trained with all samples for a single dimension of the pose vector. That is, random forests 1 to 3 are trained for changes in translation (*x*, *y*, and *z)*, and random forest 4 to 6 are trained on the changes in rotation (roll, pitch and yaw) parameters respectively. During tracking, the depth changes are used to predict the changes in pose vector by simply combining the predictions of each random forest.

In practice, we set $n_t = 70$, resulting in only 420 random trees which in turn leads to a 60 times faster training time and 60 times less memory requirements. It typically takes about 3 minutes on an Intel(R) Core(TM) i5-3570 CPU (the proposed approach is implemented and tested on such a workstation) to train a new object for tracking. This low memory requirement also allows the tracking to run in real time.

We initially sample a set of approximately 400 points from the surface of the object. Sampling is done by raytracing points onto the objects surface, creating a 3D grid around that object and then sampling a single point per grid. This sampling approach leads to evenly spaced sample points all over the visible surface of the object. The depth distance of these sample points to the visible depth map is then used in the training data. Since we have sampled points from all around the objects, many points will not lie on the visible surface but will be behind it. We rely on the random

regression forest to figure out what this means in term of object movement. In our experiments, the use of these single set of points has proven to lead to a highly stable tracking performance.

Pose Forecast. Tracking is performed on a frame by frame basis, by checking how depth values from sample points of the current positions vary when compared to the depth values of the current depth map. This means when an object is moving fast between two frames so that no or only few sample points of the previous position overlaps with the new position, the object cannot be tracked. In our algorithm we use the previously estimated movement prediction as a starting guess for the new movement prediction. This allows for objects to move further between two frames, and also provides a more accurate initial guess for the pose estimation. Initially when performing object localization, a movement of zero is assumed.

3.3. Combined Object Localization with Tracking

The full object tracking framework as shown in Fig. 1 combines both global object localization and tracking. The goal of this framework is to produce a continuous, low latency stream of the current object position. Whenever possible the system uses the fast tracking described above, and if tracking was not successful it switches back to the slower but global object localization. To track multiple objects simultaneously this tracking framework is run in parallel for each object.

From Object Localization to Tracking. Depending on the configuration of the object localization, it is possible to find multiple instances of the same object in a single frame. For our use case we assume that only a single instance is the correct one. To determine which of these instances the correct one is, we perform two checks. First, the instances are ranked by the number of *inliers (r)*, and only if it passes a certain threshold it is considered as a potential correct pose. For each pose the multi forest tracker is evaluated and tracking is performed. When running the tracking algorithm on a correct pose, the estimated tracking transformation should be a minimal *movement (m)*. When the tracking algorithm would estimate a large movement, this means that either this current position is wrong or cannot be tracked successfully. We use this tracking movement as an additional filtering criterion. In practice, we use both parameters to form a single sorting criterion *quality (q)* in (1):

$$q = \frac{r}{m^2} \qquad (1)$$

We rank all candidate poses by this criterion, which leads to more accurate results than the use of either one of the criterion separately.

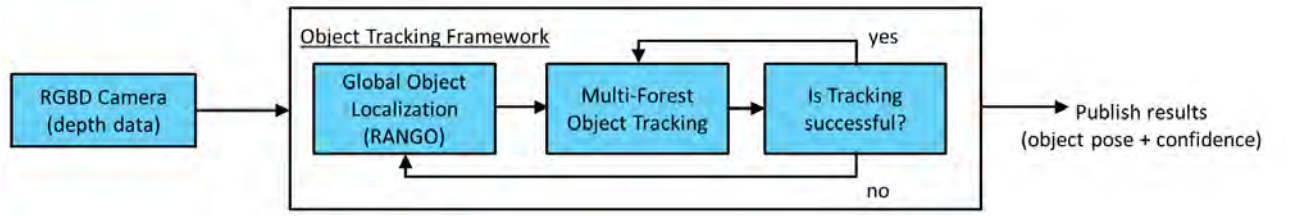


Figure 1. Object tracking framework. The framework first performs global object localization on the input depth data. After the detection results are filtered they are passed to the Object Tracking module. If tracking was successful, the framework will directly use the tracking module to track objects for the next sensor input. If not, it will revert back to global object localization for the next sensor input frame. The tracking results are published to a higher level system.

Tracking Verification. After tracking has been performed, we calculate the movement of the object with respect to the camera position. To determine if tracking was successful, we calculate if the current movement is reasonably realistic. We do this by calculating the acceleration of the object within the last 3 frames. If the acceleration is above a threshold, we consider the tracking as

not successful. This happens when the tracking has "lost" the object and consecutive tracking iteration move the object around in the sensor data. In practice we have set this velocity to 80mm per frame. With 30 fps this means a velocity of about 2.4m per second. This is enough to accurately track quickly moving objects, while being robust to detect the random movements that typically occur when tracking of the object fails.

4. Experiments and Evaluation

In order to compare the performance of our approach against the state of the art we use the synthetic dataset provided by Choi and Christensen [5]. The approach is also evaluated on real-world objects and the results are presented in [12]. Interested readers can find more information here[2]. The dataset in [5] consists of four object models and a synthetic test sequence (1000 RGB-D frames) for each object. The test sequence is obtained by placing each object in a virtual kitchen model and moving a virtual camera around the model. The object trajectories w.r.t the virtual camera coordinate frame serves as the ground truth pose (error free since it is generated via rendering) of the object. Fig. 2 shows one such frame of each object sequence. The performance of the proposed tracking approach on the synthetic data set is as shown in Figure 3.

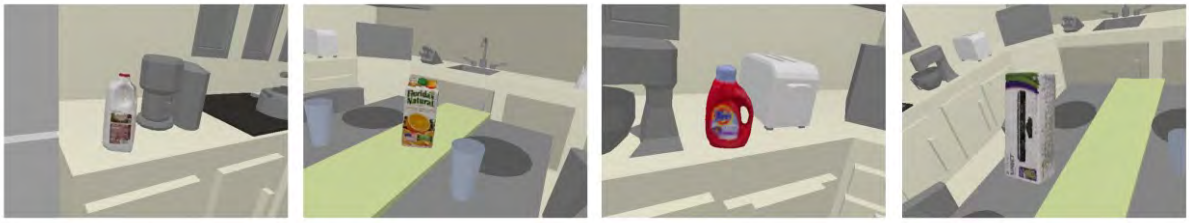


Figure 2. Example images from the synthetic test data set provided by [5], a) Milk b) Orange Juice c) Tide d) Kinect Box

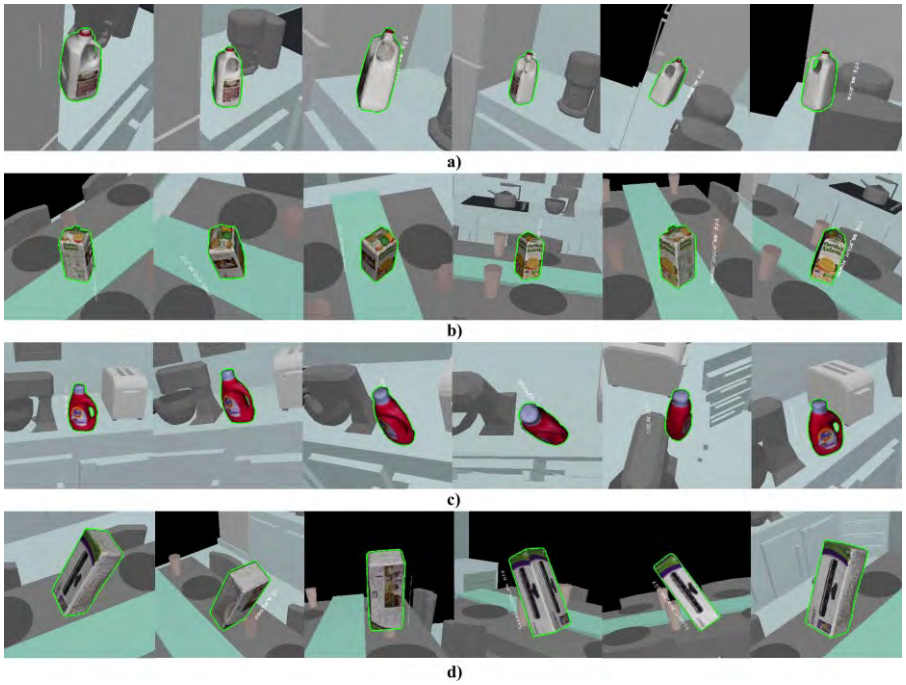


Figure 3: Results of the tracking approach on the synthetic data set a) Milk b) Orange Juice c) Tide and d) Kinect box

[2] http://tracking.profactor.at

Evaluation. The evaluation aims at comparing our approach against the particle filter based approaches [5][18][9] and online learning based approach [20] in estimating the translation (in x, y, and z axis) and rotation (roll, pitch and yaw) parameters. We compute the root mean square (RMS) errors (translation, rotation) and average time per frame. Table I shows our approach outperforms [5] and [18] over all sequences. Unlike [5] our approach only uses depth data for 3D tracking. It also performs better than [9] on average of about 0.31 mm and 1.16 deg in estimating the translation and rotation parameters respectively. Our approach requires much less computational time (1.7 ms per frame) when compared with [9] (131 ms). Though our approach performs on par with [20] in terms of run-time, it performs better in estimating the translation (by 0.31 mm) and rotation parameters (by 0.15 deg) on average.

TABLE I. COMPARISION OF OUR APPROACH WITH THE STATE OF ART AGAINST THE RMS ERRORS IN TRANSLATION (IN MM), ROTATION (DEGREES) AND THE RUNTIME (MS)

			PCL [18][1]	Choi [5][2]	Krull [9][3]	Tan [20][4]	Ours[5]
a) Milk	***RMS Error***	*Transl. (x)*	13.38	0.93	**0.51**	1.23	0.63
		Transl. (y)	31.45	1.94	1.27	**0.74**	1.19
		Transl. (z)	26.09	1.09	0.62	**0.24**	0.48
		Roll	59.37	3.83	2.19	0.50	**0.19**
		Pitch	19.58	1.41	1.44	**0.28**	**0.28**
		Yaw	75.03	3.26	1.90	0.46	**0.27**
		Time	2205	134	135	**1.5**	1.7
b) Orange juice	***RMS Error***	*Transl. (x)*	2.53	0.96	0.52	1.10	**0.39**
		Transl. (y)	2.20	1.44	0.74	0.94	**0.37**
		Transl. (z)	1.91	1.17	0.63	**0.18**	0.37
		Roll	85.81	1.32	1.28	0.35	**0.12**
		Pitch	42.12	0.75	1.08	0.24	**0.17**
		Yaw	46.37	1.39	1.20	0.37	**0.15**
		Time	1637	117	129	**1.5**	1.69
c) Tide	***RMS Error***	*Transl. (x)*	1.46	0.83	0.69	0.73	**0.42**
		Transl. (y)	2.25	1.37	0.81	0.56	**0.51**
		Transl. (z)	0.92	1.20	0.81	**0.24**	0.64
		Roll	5.15	1.78	2.10	0.31	**0.22**
		Pitch	2.13	1.09	1.38	**0.25**	0.29
		Yaw	2.98	1.13	1.27	0.34	**0.30**
		Time	2762	111	116	**1.5**	1.7
d) Kinect Box	***RMS Error***	*Transl. (x)*	43.99	1.84	0.83	1.54	**0.30**
		Transl. (y)	42.51	2.23	1.67	1.90	**0.49**
		Transl. (z)	55.89	1.36	0.79	0.34	**0.31**
		Roll	7.62	6.41	1.11	0.42	**0.21**
		Pitch	1.87	0.76	0.55	**0.22**	0.27
		Yaw	8.31	6.32	1.04	0.68	**0.23**
		Time	4539	166	143	**1.5**	1.71
Mean		***Transl.***	18.72	1.36	0.82	0.81	**0.50**
		Rot.	29.70	2.45	1.38	0.37	**0.22**
		Time	2786	132	131	**1.5**	1.7

1,2 Intel Core2 Quad CPU Q9300, 8G RAM with Nvidia GTX 590 GPU; **3** Intel(R) Core(TM) i7 CPU with a Nvidia GTX 550 TI GPU; **4** Intel(R) Core(TM) i7 CPU; **5** Intel(R) Core(TM) i5 CPU

6. Conclusion

We have presented a framework for combining object tracking and object localization to provide robust tracking performance in a challenging scenario. A quantitative analysis of the evaluation on popular test data set is also presented. The evaluation shows that our approach performs better than the state of art in terms of estimating the translation and rotation parameters. The approach is

capable of real-time computation at 1.7 ms per frame on average. The next steps would be to combine the real-time object tracking approach with human tracking and extend the framework towards human activity recognition in industrial settings.

7. Acknowledgment

This research is funded by the projects KoMoProd (Austrian Bundesministerium für Verkehr, Innovation und Technologie), SIAM (FFG, 849971) and CompleteMe (FFG, 849441).

8. References

[1] Armando Pesenti Gritti *et al.*, "Kinect-based people detection and tracking from small-footprint ground robots", *in Proc. IEEE Int. Conf. on Intelligent Robots and Systems*, 2014.

[2] A. Yilmaz *et. al.*, "Object tracking: A survey," *ACM Computer Surveys (CSUR)*, vol. 38, no. 4, pp. 1–45, 2006.

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

[4] C. Choi *et al*., "Robust 3D visual tracking using particle filtering on the special Euclidean group: A combined approach of keypoint and edge features", *Int. Journal of Robotics Research,* vol. 31, no.4, pp. 498–519, 2012.

[5] C.Choi and HI. Christensen, "RGB-D Object Tracking: A Particle Filter Approach on GPU", *in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 1084-1091, 2013.

[6] C. Papazov and D. Burschka, "An efficient RANSAC for 3-D object recognition in noisy and occluded scenes", *in Proc. 10th Asian Conf. Computer Vision (ACCV),* 2010, pp. 135–148.

[7] C. Ren, V. Prisacariu, D. Murray, and I. Reid, "Star3d: Simultaneous tracking and reconstruction of 3d objects using rgb-d data," *in Proc. Int. Conf. on Computer Vision,* 2013.

[8] J. Suarez and R.R. Murphy, "Hand gesture recognition with depth images: A review," *in IEEE Int. Symposium on Robot and Human Interactive Communication*, pp. 411–417, 2012.

[9] Krull Alexander *et al.,* "6-dof model based tracking via object coordinate regression", *Computer Vision—ACCV, 2014*, Springer International Publishing, pp 384-399, 2015.

[10] M. Isard and A. Blake, "Condensation - Conditional density propagation for visual tracking," *Int. Journal of Computer Vision*, vol. 29, no. 1, 1998.

[11] Mao Ye *et al.*, "A survey on human motion analysis from depth data", *in Timeof- Flight and Depth Imaging. Sensors, Algorithms, and Applications*, pp. 149–187, Springer, 2013.

[12] S. Akkaladevi, M. Ankerl *et al.,* "Tracking multiple rigid symmetric and non-symmetric objects in real-time using depth data," [to appear] *in Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[13] S. Koo, D. Lee, and D. Kwon, "Multiple object tracking using an rgb-d camera by hierarchical spatiotemporal data association", *in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 1113–1118, 2013.

[14] S. Koo, D. Lee, and D. Kwon, "Incremental object learning and robust tracking of multiple objects from rgb-d point set data," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 108–121, 2014.

[15] S. Koo, D. Lee, and D. Kwon, "Unsupervised object individuation from rgb-d image sequences", *in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 4450– 4457, 2014.

[16] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm", *in Proc. IEEE Int. Conf. on 3-D Digital Imaging and Modeling*, pp. 145–152, 2001.

[17] S. Winkelbach, S. Molkenstruck, F. M. Wahl, "Low-Cost Laser Range Scanner and Fast Surface Registration Approach", *Pattern Recognition (DAGM 2006) LNCS 4174*, pp. 718-728, Springer 2006.

[18] Rusu, R. B. Rusu and S. Cousins. "3d is here: Point cloud library (pcl)", *in Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1- 4, 2011.

[19] Tan David Joseph, and Slobodan Ilic, "Multi-forest Tracker: A Chameleon in Tracking", *in Proc. Of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2014.

[20] Tan David Joseph *et al.*, "A Versatile Learning-based 3D Temporal Tracker: Scalable, Robust, Online" *in Proc. IEEE Int. Conf. on Computer Vision,* vol. 1, No. 4, 2015.

[21] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, A. Hengel, "A survey of appearance models in visual object tracking", *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 1-48, 2013.

On a Fast Implementation of a 2D-Variant of Weyl's Discrepancy Measure

Christian Motz[1] , Bernhard A. Moser[1]

Knowledge-Based Vision Systems
Software Competence Center Hagenberg, Austria
christian.motz@scch.at, *bernhard.moser@scch.at*

Abstract
Applying the concept of Hermann Weyl's discrepancy as image similarity measure leads to outstanding robustness properties for template matching. However, in comparison with standard measures this approach is computationally more involving. This paper analyzes this measure from the point of view of efficient implementation for embedded vision settings. A fast implementation is proposed based on vectorization of summed-area tables, resulting in a speed-up factor 16 compared to a standard integral image based computation.

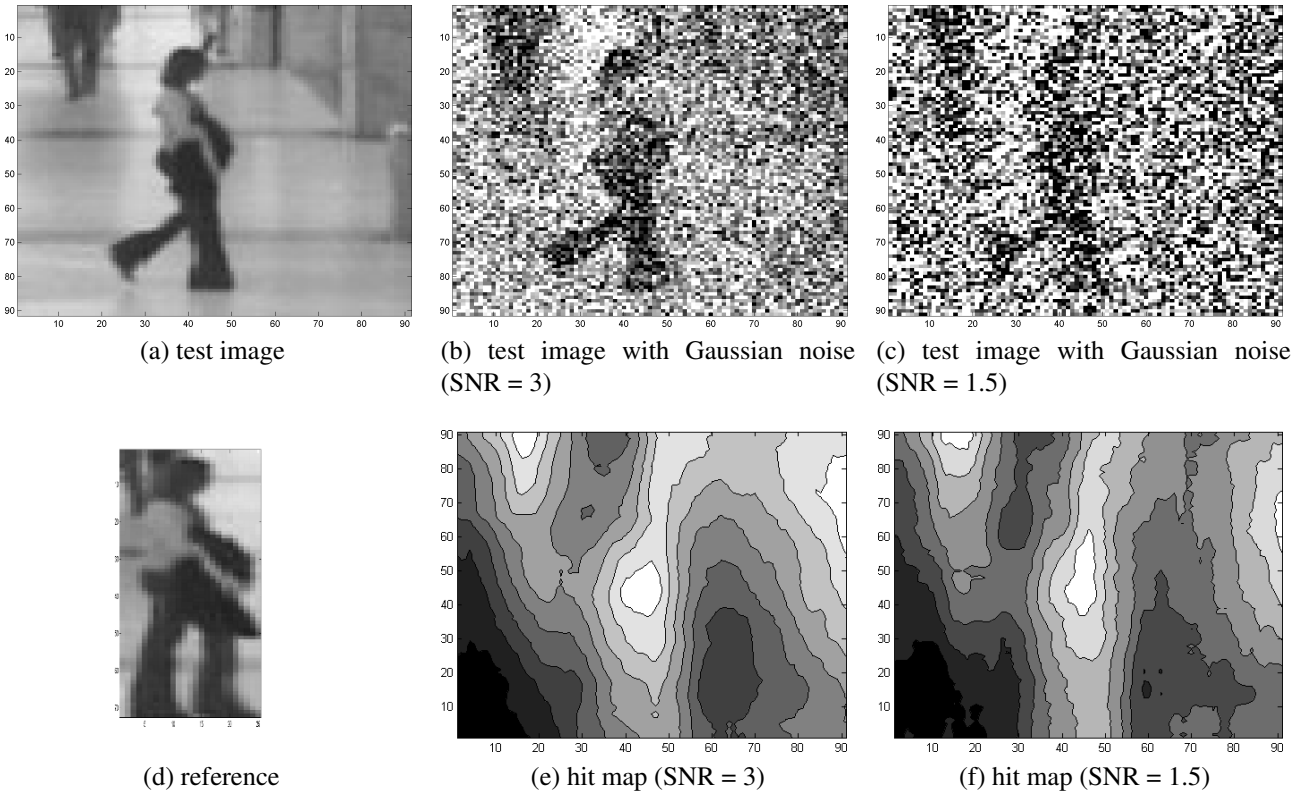
1. Introduction

In this paper we take up a novel concept of similarity measure due to [1] and investigate its applicability for the requirements of embedded vision. The core idea of this measure is its design principle based on a family of subsets rather than evaluating the aggregation of point-wise comparisons on a pixel-by-pixel level. In contrast to pixel-by-pixel based approaches with subsequent commutative aggregation such as mutual information of normalized cross correlation the subset-based approach also takes spatial arrangements into account which makes this approach interesting for pattern analysis and matching purposes [2].

This measure goes back to H. Weyl already 100 years ago and was studied in the context of evaluating the quality of pseudo-random numbers and measuring irregularities of probability distributions [3]. For one-dimensional signals (vectors) it is defined as

$$\|(x_1,\ldots,x_n)\|_D = \max_{1\le a,b\le n} |\sum_{i=a}^{b} x_i| = \max_r\{0,\sum_{i=1}^{r} x_i\} - \min_s\{0,\sum_{i=1}^{s} x_i\}$$

Interestingly, this measure not only plays a central role in discrepancy theory which is related to low complexity algorithmic design by means of low discrepancy sequences [4], but as found out recently, also in other fields of applications, e.g. in event-based signal processing [5, 6], random walk analysis [7] and image and volumetric data analysis by extending it to higher dimensions by means of integral images [1]. As pointed out in [1] the extension is not unique. A possible extension is given by Equation (1).



(a) test image (b) test image with Gaussian noise (SNR = 3) (c) test image with Gaussian noise (SNR = 1.5)

(d) reference (e) hit map (SNR = 3) (f) hit map (SNR = 1.5)

Figure 1. Illustration of a pattern matching problem: find reference image in the test image. The images are taken from frame 697 **and frame** 705 **of the EC Funded CAVIAR project/IST 2001 37540 (”Shopping Center in Portugal”, ”OneLeaveShop2cor”). The hit map is computed using the measure (1)**

$$\begin{aligned}\|f\|_D := \max \Bigg\{ & \max_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(i,j)}\} - \min_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(i,j)}\}\,, \\ & \max_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(N-i,j)}\} - \min_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(N-i,j)}\}, \\ & \max_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(i,M-j)}\} - \min_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(i,M-j)}\}, \\ & \max_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(N-i,M-j)}\} - \min_{0\le k\le N,0\le l\le M}\{\sum_{i=0}^{k}\sum_{j=0}^{l} I_{(N-i,M-j)}\} \Bigg\} \end{aligned} \quad (1)$$

For registration and template matching purposes the discrepancy measure is applied on the difference of the corresponding images. To be more precise, the images are considered as two-dimensional functions on the lattice of integers with with default values 0 outside the proper frame of the images.

This measure satisfies the following desirable registration properties (see also [8, 9]):

[R1] a vanishing distance entails a vanishing extent of misalignment and vice versa,

[R2] the distance measure behaves continuously at least with respect to arbitrary small misalignments,

[R3] an increasing extent of misalignment implies an increasing distance measure and vice versa (monotonicity).

It is interesting that it can be shown that these natural properties are not satisfied simultaneously by commonly used matching and registration techniques [1]. Figure 1 illustrates its robustness by applying this measure as fitness function for finding the best match between a reference and a test image. In this demonstration the discrepancy measure is directly applied without any image preprocessing or denoising. As this measure relies only on the evaluation of integral images and max/min operations, it is well-suited for parallelization. An efficient implementation can be tackled by means of of the concept of a summed-area table [10] which is a matrix generated from an input image in which each entry in the matrix stores the sum of all pixel values between the entry location and the lower-left corner of the input image. For applications of summed-area table see also [11] and related concepts based on integral images e.g. [12]. The power of the summed-area table comes from the fact that it can be used to perform filters of different widths at every pixel in the image in constant time per pixel. This makes SAT very useful for embedded vision purposes.

The paper is organized as follows: Section 2. introduces a two-dimensional definition of the discrepancy norm and makes algorithmic optimizations to reduce computation effort. Section 3. presents a vectorization concept for the previously optimized algorithm. Section 4. presents the speedup achieved by the optimizations.

2. Algorithmic Analysis for Implementation

While in the 1-dimensional case partial sums over intervals are evaluated in the 2-dimensional case rectangles are taken instead of intervals. As shown in [1] is suffices to restrict on the rectangles with one corner being coincident with a corner of the image. This suggests to use integral images spreading of each of the four corners of the image. However, a single integral image already contains the information of the remaining integral images from the other corners. This leads to the first optimization step by deducing the values for the four integral images from the original integral image with the top left corner as the starting point. Assume point P_1 is our current index. The value at this position naturally corresponds with the first integral image in the definition of discrepancy norm. The second integral image in the definition has the top right corner as a reference. This corresponds to area II_2 in the figure, which can be computed by subtracting sums $P_2 - P_1$. The third and fourth integral images are very similar. Equations (2) to (3) provide a mathematical formulation of the integral image

transformations as explained above:

$$\begin{aligned}
\Pi_1(x,y) &= \Pi(x,y) && (2)\\
\Pi_2(x,y) &= \begin{cases}\Pi(W,y)-\Pi(x,y) & \text{if } x \neq W,\\ \Pi(W,y) & \text{if } x = W.\end{cases}\\
\Pi_3(x,y) &= \begin{cases}\Pi(x,H)-\Pi(x,y) & \text{if } y \neq H,\\ \Pi(x,H) & \text{if } y = H.\end{cases}\\
\Pi_4(x,y) &= \begin{cases}A(x,y) & \text{if } x \neq W, y \neq H,\\ \Pi(x,H)-\Pi(x,y) & \text{if } x = W, y \neq H,\\ \Pi(W,y)-\Pi(x,y) & \text{if } x \neq W, y = H,\\ \Pi(x,y) & \text{if } x = W, y = H.\end{cases}\\
A(x,y) &= \Pi(W,H)+\Pi(x,y)-\Pi(W,y)2-\Pi(x,H) && (3)
\end{aligned}$$

W indicates the last valid x index of a row and H the last valid y index of a column. Care has to be taken, if any index lies on the edge: Here, some components simply refer to the same area and are equal.

Now, we make use of transformations which the summation terms are invariant to: we are interested in the difference between maximum and minimum of the summation. Thus, adding a constant factor to all elements within the integral image will not affect the difference between maximum and minimum. When it comes to (3), the term $\Pi(W,H)$ clearly is constant, neither depending on index variable x nor on y. As a result, it can be omitted for the discrepancy norm calculation. Note that this needs to be compensated in the others cases of Π_4, too. Taking the constraint for the indexes into account, Equations (4) and (5) are obtained:

$$\tilde{\Pi}_4(x,y) = \begin{cases}\tilde{A}(x,y) & \text{if } x \neq W, y \neq H,\\ -\Pi(W,y) & \text{if } x = W, y \neq H,\\ -\Pi(x,H) & \text{if } x \neq W, y = H,\\ 0 & \text{if } x = W, y = H.\end{cases}, \quad (4)$$

$$\tilde{A}(x,y) = \Pi(x,y)-\Pi(W,y)-\Pi(x,H). \quad (5)$$

2.1. Reducing the compare operations

So far, the discrepancy norm in 2D has been reduced to computing a single integral image and expressing the other forms based on this single one. It still requires 8 compare operations per pixel: one for the minimum, one for the maximum and this has to be done four times for the different components. Some of these operations are redundant if we concentrate on a specific row, meaning y is constant. Applying this method to Π_2 of Equation (2), we obtain equations (6) and (7) which differ only by a constant and a sign. The negative sign will swap minimum and maximum. As a result, the second component can be deduced with simple operations that are only necessary at the end of each

row:

$$\Pi_{p1}(x) = \Pi_p(x) \tag{6}$$

$$\Pi_{p2}(x) = \begin{cases} \Pi_p(W) - \Pi_p(x) & \text{if } x \neq W, \\ \Pi_p(W) & \text{if } x = W. \end{cases} \tag{7}$$

Equations (8) and (9) provide a mathematical formulation of this, in both cases c corresponds to the constant $\Pi_p(W)$ per row. In other words, we only need to compute a single integral image, and compute the minimum and maximum per row, by which we have half of the computation done to get the discrepancy norm according to (1):

$$\max\{\Pi_{p2}(x)\} = \max\{c, c - \min\{\Pi_{p1}(x)\}\}, \tag{8}$$

$$\min\{\Pi_{p2}(x)\} = \min\{c, c - \max\{\Pi_{p1}(x)\}\}. \tag{9}$$

The third component behaves similar to the second one — the only difference is that the constant now is per column and we need the maximum and minimum for each column. Unfortunately, the fourth and last component is more complex. Here, each value is based both on the last value per column and the last value per row. The problem is that for normal maximum or minimum we only take care of numbers that are larger or smaller but not the equal ones. Yet, for the problem mentioned above, we would need all maximized subexpressions with the same value and their corresponding position consisting of the x and y index pair. Further research is necessary to check whether the minimum and maximum of the fourth component could be determined in a more convenient and less complex way. However, a subexpression refers to the third component and only one addition is necessary to get the fourth component.

2.2. Proposed algorithm

Based on the previous findings, we will now consider the complete algorithm and compare it to the base implementation in terms of runtime complexity. The base algorithm consists of four passes over the data; each will compute one integral image component and, simultaneously, yield minimum and maximum by compare operations. The optimized version consists of only two passes. The first pass will calculate one integral image and get the first and the second component at the same time. Here, the second component needs a small overhead at the end of each row. The second pass will deduce the third and fourth component based on the previously computed integral image.

Basically, both versions have $\mathcal{O}(n \cdot m)$ complexity, where n is the image width and m the image height. If we take a closer look at it, the proposed version has $\mathcal{O}(2n \cdot m)$, compared to the initial $\mathcal{O}(4n \cdot m)$. The optimized version will show further improvements if we consider the number of operations more precisely. Additions and subtractions will be considered as the same operation from the view of complexity. The base version has four very similar passes, consisting of integral image computation and comparisons. Computing an integral image point takes three additions, though, there is an optimized version needing only two which requires extra storage for cumulative row sums [13]. The reference version from [14] is implemented with three operations and will be used for real performance comparison.

Table 1 compares both version in terms of the overall operation count. Additions are reduced heavily down to less than a half. Comparisons are brought down by a fourth approximately. As each pass consists of a double-nested loop that produces overhead, the column *Passes* is very important. Another

	Passes	*Additions*	*Comparisons*
Ref	4	$12n \cdot m$	$8n \cdot m$
Opt	2	$5n \cdot m + 2m$	$6n \cdot m + 4m$

Table 1. Comparing the number of operations for the reference and the optimized version.

factor is storage, given the fact that both versions require additional storage of $n \cdot m$. But the base version writes four times to this area, whereas the optimized version only once. If this storage area is accessible to the user, calculating the discrepancy norm in 2D always yields the according integral image for free.

3. Parallelization

The proposed algorithm seems to be well-suited for parallelization methods. Computing and comparing the other components of the discrepancy norm is highly independent. When it comes to parallelization, modern computers offer various options. A common classification in this area comes from [15]. The classification is based on the number of parallel instruction and data streams. A traditional processor belongs to SISD, whereas multi core or multi processor systems are MIMD. Instruction set extensions like SSE and AVX, also referred to as vector units, belong to SIMD.

A similarity measure like the discrepancy norm will normally be applied many times. Pattern matching requires evaluating the discrepancy norm at many different positions of a patch. Therefore, SIMD is a promising approach. It is especially suitable for applying the same kind of operation to several data values at once. Furthermore, SIMD means choosing certain special instructions. At runtime, they do not have any overhead, compared to normal SISD instructions. On the other hand, making use of multiprocessing would lead to an overhead due to the fact that it involves spanning threads, distributing data and synchronizing at the end. As shown by [16], using multi core processors is complex. On the one hand, the work succeeded in using multiple cores to improve performance. On the other, hand the processor topology has an impact. The authors had to bind the threads to cores sharing the same L2 cache in order to improve performance. Not fulfilling this requirement results in a significant performance penalty.

SIMD instructions operate on a dataset or a so called vector. For example, a traditional add would perform $a := a + b$. The SIMD version of this instruction would perform the same operation, but a would be a vector. Typical vector sizes of SIMD units range from 2 to 8 elements. Normally, vector units have registers of a fixed size. Depending on the size of the data type, they can process a certain amount of elements in one step. Vector units are not designed to operate horizontally, which would mean combining elements within a vector register. We will concentrate on the common SIMD extensions for the x86/x64 architecture. There are two extensions in this area: SSE and AVX - both exist in different versions, with each new version extending the previous one by adding new computing capabilities [17].

AVX doubled the vector size comparesd to SSE. Yet, in terms of data shuffling, the situation became much more complex: with vector registers and operations split into lanes, one AVX register consists of two 128-bit lanes which simplified implementing the architecture for the designers. It does not make any difference for vector operations like additions. Nevertheless, for instance, the SSE shuffle operation takes an immediate value that allows indexing of up to four elements. The according AVX

input	v_3	v_2	v_1	v_0
shuffle	v_2	v_3	v_0	v_1
max	$\max_{32}$	$\max_{32}$	$\max_{10}$	$\max_{10}$
shuffle	$\max_{10}$	$\max_{10}$	$\max_{32}$	$\max_{32}$
max	$\max_{3210}$	$\max_{3210}$	$\max_{3210}$	$\max_{3210}$

Table 2. Getting minimum / maximum of a vector register holding 4 elements.

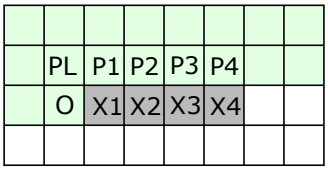


Figure 2. Dependencies for computing new values for an integral images with vector units.

instruction has the same indexing capabilities but operates on a doubled data amount. The AVX instruction simply takes the immediate value and applies the shuffle for each lane. Only a small number of special instructions allow crosslane data exchange in some restricted ways [17, Volume 1 Chapter 14].

3.1. Vectorization

Taking into account the points mentioned above, we will now develop a vectorization scheme. Using vector units for the comparisons is straight forward. Vector units will help us to compare n elements at once. Finally, we have to get the maximum and minimum of the vector itself, which results in overhead because of operating horizontally. To be precise, additional comparisons of $\log n$ are necessary and the same amount of data permutations. The basic idea is to compare pairs of values and then use the result again for pairwise comparisons but with half the number of pairs. Vector units permit comparing several pairs with a single instruction at the same time. Data shuffling assures we are comparing different pairs in the next step. Table 2 illustrates the procedure for a vector unit holding 4 elements.

More complicated is the vectorization of computing the integral image, which can be interpreted as a 2D version of the prefix sum. [18] provides a good summary of prefix sums in general, their applications and a parallel version. The proposed parallelization model is well suited for using GPU acceleration. This was proven by [19]. On the other hand, the GPU version turned out to be only useful for large dataset. Moreover, this was tested for traditional prefix sums and not for 2D versions suitable for integral images which would consist of two passes: one prefix sum over the rows, a second one over the columns, taking the result of the first pass as the input. A similar two-pass algorithm for integral images using GPU was developed and tested by [20]. A notable speedup was not gained for images with a pixel count less than 0.5 million. Furthermore, the data transfer to and from the GPU was not included in the time measurement. Using SIMD extension for integral image computation is a quite new approach, leading to the fact that there are few literature reference that use SIMD extensions. [16] applies SSE for a part of the computation algorithm. Nonetheless, finally, this version is slower than a sequential algorithm presented in the same work. We will show an implementation consisting of a single pass instead of two traditional prefix sum passes. Figure 2 shows the dependencies for computing the new pixels X1 .. X4 in the integral image. All new pixels

have the same row offset (O) and the same compensation factor PL, which is optimal for a vector unit. The same is true for adding the different previous values P1 .. P4. Using PL and O for all vector elements requires one additional instruction to broadcast the single value to all vector elements. We will refer to the summing of X1 to X4 as partial prefix sum. The approach is similar to horizontal minimum / maximum within a vector register. The difference is that a shuffle would not help, but shifting solves the problem. Table 3 lists the single steps. Two additions and two shifts are necessary. This definitely is an improvement over [16] as their scheme required three additions and the same amount of shifts. It might be the reason that their SSE version was slower than an enhanced serial algorithm.

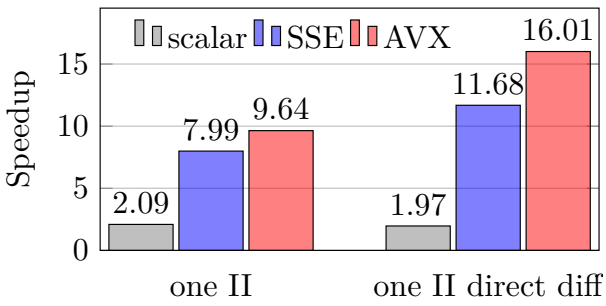
Unfortunately, as AVX does not have shift operations, they are considered to be only useful for integer data. So the shift has to be emulated by combining a shuffle and a blend. The shuffle rearranges data and the blend masks the first element with zero, which is not supported by the shuffle or other operations. AVX2 added integer support and shift operations at the same time. Due to the lane concept, a special cross-lane operation is necessary. The idea is to do the partial sum for each lane. In the last step, the overall sum of the lower lane is broadcasted to all elements in the higher lane of a register and added. What is helpful is that the partial prefix sum in the first step is independent from the other values. Without any doubt, $O - PL + P_x$ does not depend on the sum at first. In the final step, both temporal results have to be merged with a vector addition. In the first place, we have two independent data streams, which helps exploring instruction level parallelism. This is especially important due to the fact that — as stated before — summing within the vector is not ideal for vector units. The data preparation is another step optimal for vector units. If pattern matching is done using a norm without an inner product, the similarity measure is applied to the difference between pattern and test candidate.

We can estimate the expected speedup. For the regular version, we require 3 additions (or subtractions). The vectorized version has an overhead of $2 \cdot \log n$, where n is the number of vector elements. Then, there are three additions and one broadcast, but this already computes n pixels at once. Note, that this is a very rough estimation. We have not taken into account instruction level parallelism. This means instructions differ in latency and throughput. Moreover, the processor might have more operational units for some instructions than for others [21]. Another fact we did not consider is moving data around. SSE and AVX are — like the whole x86 instruction set — based on load and store. The normal version requires a load for each single element, however, the corresponding instructions for vector units load data chunks as large as the vector unit in a single step at the same time. Making the process faster, the bandwidth is also exceeded faster.

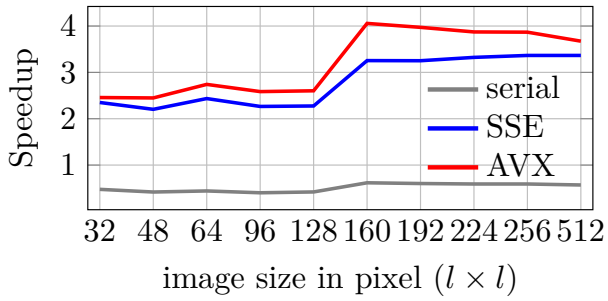
The complexity of the analysis above should make it clear that it is nearly impossible to give an estimated speedup for the whole discrepancy norm calculation. Thus, we will use practical tests to evaluate the performance impact.

input	v_3	v_2	v_1	v_0
shift	0	v_3	v_2	v_1
add	v_3	v_{3+2}	v_{2+1}	v_{1+0}
shift	0	0	v_3	v_{3+2}
add	v_3	v_{3+2}	v_{3+2+1}	$v_{3+2+1+0}$

Table 3. Computing partial prefix sum for a vector register holding 4 elements.



(a) Details about performance if algorithmic optimization and vectorizationis applied to discrepancy norm.



(b) Details about performance if vectorization is applied to integral image computation with the OpenCV algorithm serving as references.

Figure 3. Results of performance evaluation tests

4. Performance Analysis and Evaluation

Coding is done with C++, whereas *Visual Studio 2013* from Microsoft severs as the compiler. The only adjustment is the setting *Enable Enhanced Instruction Set* in the group of *Code Generation*. The selected target architecture is 64-bit. The test system is based on an Intel *i5-4460*. The computer runs *Windows 7 Professional Service Pack 1 64-bit*. The test algorithm applies the discrepancy norm in a sliding window approach, that the implementation is executed many times. Furthermore, the whole test setup is run several times to eliminate random influences. As we measure similarity compared to a pattern, reference subtraction has to be applied for each window. We include this step in time measurement as it is vital for this task and can not be omitted.

Figure 3a summarizes the speedup with the test setup. The direct difference approach outperforms the other implementations by far. With the AVX vectorization leading to a speedup of 16 and SSE vectorization to a speedup of 12. The algorithmic optimized serial version already doubled performance. The average execution time of the AVX version is 0.612 seconds matching a 64×64 data patch within a 512×512 image. AVX can process eight 32-bit integer values at the same time, which is exactly the speedup gained by the vectorization compared with the serial version. On the other hand, SSE produces super linear speedup exceeding theoretical maximum. The data indicates that reference subtraction has a big impact on the runtime. Embedding the difference building in the algorithm itself improved the performance about 50% for SSE and 65% for AVX.

Another comparison concentrates on the vectorization of the integral image algorithm alone. Many common algorithms like *SURF* are based on this intermediate representation [22]. The OpenCV implementation for integral image is compared to the vectorized implementation of the authors and a straight forward serial version. Figure 3b shows the results. The OpenCV algorithm serves as the reference and is twice as fast as a simple serial implementation. This suggests that OpenCV uses the approach from [13] that requires extra storage but reduces the necessary additions. Nonetheless, the vectorized version outperforms OpenCV at any image size, gaining a speedup of 2.5 to 4, depending on the image size.

For embedded applications it is interesting whether the vectorization scheme is applicable to other architectures, too. In embedded computing the ARM architecture plays a crucial role. Here, the ARM Cortex-A series offers SIMD capabilities with *NEON* technology offering a data width equal to SSE. [23]. All in all, the whole vectorization can be coded with the NEON instructions. Unfortunately,

both SIMD architectures have totally different instructions when it comes to data reordering. The most powerful data rearrange instructions of *NEON* are *VTBL* and *VTBX*. One the one hand, they are expensive in terms of execution cycles. On the other hand, several cases can be replaced by faster instructions. For example, the vector shift can be achieved using the NEON instruction *VEXT* and a register containing zero; broadcasting a single element can be done with *VDUP*. The newest ARM instruction set named *ARMv8* should allow further performance optimizations by adding cross-lane instructions providing functionality exactly needed by discrepancy norm calculation like horizontal summation and taking minimum or maximum [24]. Though, it is impossible to estimate the achievable speedup without practical tests.

5. Conclusion

We analyzed a variant of an image similarity measure based on Hermann Weyl's discrepancy from the point of view of efficient implementation by exploiting redundancies in computing multiple integral images. Finally, we proposed an implementation based on vectorization of prefix sums and summed-area tables which results in a speed-up factor 16 compared to a standard integral image based computation. Future research is left for checking parallelization and implementation optimizations also of other variants of Weyl's discrepancy.

References

[1] B. A. Moser, "A similarity measure for image and volumetric data based on Hermann Weyl's discrepancy.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2321–9, nov 2011.

[2] B. Moser, G. Stübl, and J.-L. Bouchot, "On a non-monotonicity effect of similarity measures.," in *SIMBAD* (M. Pelillo and E. R. Hancock, eds.), vol. 7005 of *Lecture Notes in Computer Science*, pp. 46–60, Springer, 2011.

[3] H. Weyl, "Über die Gleichverteilung von Zahlen mod. Eins," *Mathematische Annalen*, vol. 77, pp. 313–352, Sept 1916.

[4] B. Chazelle, *The discrepancy method: randomness and complexity*. Cambridge University Press, 2000.

[5] B. A. Moser and T. Natschlager, "On stability of distance measures for event sequences induced by level-crossing sampling," *Signal Processing, IEEE Transactions on*, vol. 62, no. 8, pp. 1987–1999, 2014.

[6] B. A. Moser, "Stability of threshold-based sampling as metric problem," in *Event-based Control, Communication, and Signal Processing (EBCCSP), 2015 International Conference on*, pp. 1–8, IEEE, 2015.

[7] B. A. Moser, "The range of a simple random walk on $\mathbb{Z}$: An elementary combinatorial approach," *The Electronic Journal of Combinatorics*, vol. 21, no. 4, pp. P4–10, 2014.

[8] J.-L. Bouchot, G. Stübl, and B. Moser, "A template matching approach based on the discrepancy norm for defect detection on regularly textured surfaces," in *10th International Conference on Quality Control by Artificial Vision*, pp. 80000K–80000K, International Society for Optics and Photonics, 2011.

[9] G. Stübl, B. Moser, and J. Scharinger, "On approximate nearest neighbour field algorithms in template matching for surface quality inspection," in *Computer Aided Systems Theory-EUROCAST 2013*, pp. 79–86, Springer, 2013.

[10] F. C. Crow, "Summed-area tables for texture mapping," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 207–212, 1984.

[11] J. Hensley, T. Scheuermann, G. Coombe, M. Singh, and A. Lastra, "Fast summed-area table generation and its applications," in *Computer Graphics Forum*, vol. 24, pp. 547–555, Wiley Online Library, 2005.

[12] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. 511–518 vol.1, 12 2001.

[14] G. Stübl, "Robust defect detection for near-regular textures based on Hermann Weyl ' s discrepancy measure," 12 2013.

[15] M. J. Flynn and K. W. Rudd, "Parallel architectures," *ACM Comput. Surv.*, vol. 28, no. 1, pp. 67–70, 1996.

[16] N. Zhang, "Working towards efficient parallel computing of integral images on multi-core processors," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 2, pp. 30–34, April 2010.

[17] Intel, *Intel 64 and IA-32 Architectures Software Developer's Manual, Combined Volumes: 1, 2A, 2B, 2C, 3A, 3B and 3C*, June 2015.

[18] G. E. Blelloch, "Prefix sums and their applications," tech. rep., Synthesis of Parallel Algorithms, 1990.

[19] M. Harris, S. Sengupta, and J. D. Owens, "Parallel prefix sum (scan) with cuda," *GPU gems*, vol. 3, no. 39, pp. 851–876, 2007.

[20] B. Bilgic, B. Horn, and I. Masaki, "Efficient integral image computation on the gpu," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pp. 528–533, June 2010.

[21] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*. San Francisco, California: Morgan Kaufmann, 4 ed., 2007.

[22] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*, pp. 404–417, Springer, 2006.

[23] ARM, *Cortex-A9 NEON Media Processing Engine - Technical Reference Manual*, June 2012.

[24] ARM, *ARM Cortex-A Series - Programmer's Guide for ARMv8-A*, March 2015.

Towards Agricultural Robotics for Organic Farming *

Georg Halmetschlager[1], Johann Prankl[1], and Markus Vincze[1]

Faculty of Electrical Engineering, Automation and Control Institute, Vision for Robotics Laboratory, Vienna University of Technology, A-1040 Vienna, Austria.
lastname@acin.tuwien.ac.at

Abstract
In big scale agricultural farming complex machines with advanced technology shape already the daily routine. In opposite, the field of organic farming is still characterized by multiple manual tasks that include heavy labor. Our vision is that the fields of automation and robotics offer the necessary technology to lift the burden of back-breaking work off the worker's shoulders. Hence, we propose a scalable and modular agricultural robotic concept that advances farming to the next higher technology level. We provide a low-cost and flexible design in order to realize different autonomous applications, specialized for light weight agricultural work. As proof of concept the proposed configuration is integrated and validated as the experimental platform FRANC. All experiments are performed in real-life outdoor scenarios as vegetable fields that are sowed or planted in row structures. Therefore, we utilize a local navigation system based on a self-parameterizing crop row detection, that enables a local, adaptable, and GPS-independent navigation. The tests show that the hardware and software of the designed system is able to handle rough terrain, offers a high maneuverability, and is adaptable to different row-structures.

1. Introduction

Within the last decades new automation technologies, industrial robots and sophisticated automation machineries entered the food production chain and led to a higher efficiency and increased the productivity of the harvesting process.

Sensors and software that transform classic agricultural machineries into semi-autonomous systems are already available on the market [11]. We belief that robotics has the ability to advance this semi-autonomous systems to the next higher technological level and promises to answer the question how the production chain can be fully automated in each single step of the food production, starting already at the cultivation of the crops. Therefore we developed a scalable, and modular agricultural robotic systems suitable to better support light-weight agricultural work.

As stated by [17], one way to increase the economic efficiency in future crop production may be done better and cheaper with a swarm of small machines than with a few large ones. By the means of our modular concept re-designing existing solutions can be avoided and it becomes possible to enhance existing solutions with robotic modules, for instance a conventional finger cultivator can be turned into a robot by attaching the respective robot module.

*This work was funded by Sparkling Science a programme of the Federal Ministry of Science and Research of Austria (SPA 04/84 - FRANC).

Figure 1: Experimental platform and proof of concept FRANC.

In this article we present our modular system design and concepts for more flexible agricultural robots as well as its realization with the platform FRANC (cf. Fig. 1) as proof of concept. Our contributions are (i) a modular robotic system concept that (ii) can be use for the robotizing of existing farm facilities and (iii) a generic row detection algorithm that does not need any a-priori information. Moreover, we present field trial results of its performance on vegetable fields.

2. Related Work

Most of the state-of-the-art agricultural automation systems are either focused on the (semi-) automation of big land machines or support the farmer during different field manipulation procedures with additional sensor information [11].

Research groups robotized already "standard platforms" as golf carts or other small scale vehicles to focus on algorithms and sensor technologies without the need of the re-development of the grounding vehicle [16, 10, 5]. Contrary there are also completely designed robotic systems such as BoniRob which are suitable for highly specialized solutions as occur in the area of precision farming [3]. As described in [3] BoniRob Apps are comparable to the classical implements. These Apps can be directly integrated on the robot. Existing agricultural machineries have to be redesigned if they have to be used be used with the robot. However, we aim to develop a solution which can be used in combination with existing implements and farm facilities, with little or no additional product development needed.

The contribution presented in this article is a system design concept that sets out to close the gap of existing agricultural robotic systems: rather than focusing on one large multi-purpose autonomous machine, we offer a flexible solution for cheaper crop row production which might be even more acceptable for smaller farms as it allows robotizing existing machines. We present our contribution in the form of a detailed system description and the results of preliminary field trials. Our results should help other researchers and engineers to solve the existing challenges in agricultural robotics with respect to development of robotic systems for light-weight agricultural work.

3. Approach

We approach a robotic system, including a row guidance an autonomy module that adapts by itself to any kind of row organized fields. Our concept includes individually replaceable subsystems that will be presented here.

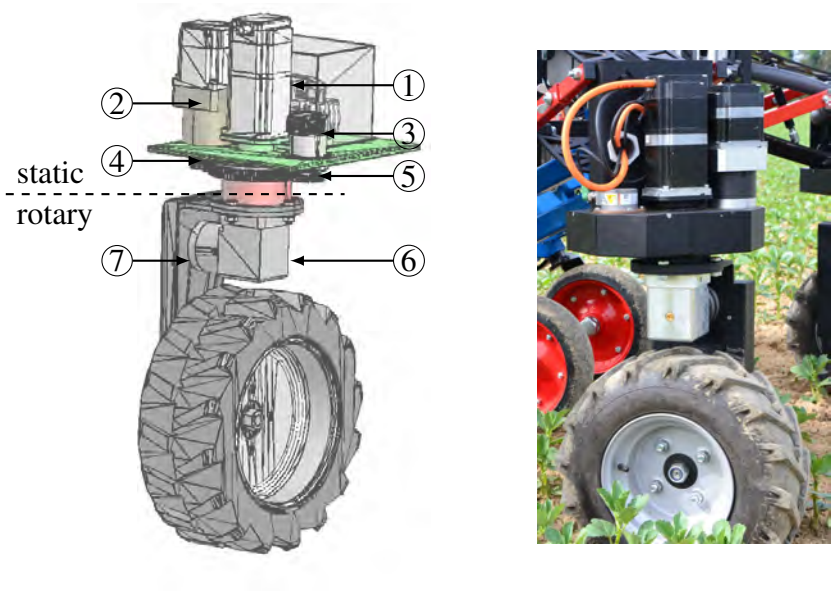


Figure 2: Mechanical realization and parts of the powertrain. ① drive motor and break, ② steering motor, ③ powertrain electric, ④ steering gear, ⑤ optical absolute rotary steering encoder, ⑥ 90° gear, ⑦ chain drive.

3.1. Mechanical Realization of the Powertrains and n-Wheeled Drive Kinematics

As each car-like vehicle, the robot needs at least three degrees of freedom (DoF). The classic kinematic realization of service robots are differential drives, in opposite we decided to implement a n-wheeled steering to combine tractive power, maneuverability, and scalability of the robot. Hence, we propose a kinematic encapsulated powertrain that can be equipped with or without a motor for the steering or tractive power. The wheel can be realized as free running wheel without any motor, can be equipped with a single motor for pure tractive power, or as fully powered, independent steerable wheel (cf. Fig. 2).

Most of the already realized systems use wheel hub motors [2, 3]. Wheel hub motors need a wired connection from the static part to the rotary part. That connection constrains the number of possible wheel turns respectively the maximum steering angle and makes the inverse kinematic complex, because the algorithms have to consider the prior steering motions. We approach a cable free rotary part that allows infinite wheel turns in order to remove these constraints for the trajectory planing.

Vehicles that are equipped with more than one steerable wheel, need a interconnected steering that fulfills the Ackerman-constraint [4]. Summarized, the perpendicular line of each wheel has to intersect at one point. However, pure mechanical realizations of steering systems go hand in hand with comparable complicated mechanical constructions. Hence, we replace the mechanical connection by a electronic connection and an intelligent control that is able to handle the steering maneuverer independent from the amount of steerable wheels, based on their position in the kinematic constellation. Based on the Ackerman-constraint and the "instantaneous center of curvature" $\mathbf{P_{ICC}}$, the necessary steering angle θ_n and different velocities v_n of the single wheels can be calculated with (1)-(4c). The approached equations result automatically in valid trajectories and steering angle configuration if $\mathbf{P_{ICC}}$ is linearly interpolated. Figure 3 depicts an exemplary kinematic configuration and depicts the nomenclature used for the equations. Different drive behaviors for different in field use cases as a

pure back or front steering can be realized dependent on the position of $\mathbf{P_{ICC}}$. The virtual coordinate system $[\mathbf{x_V}, \mathbf{y_V}]$ is used to shift the zero position and the privileged direction.

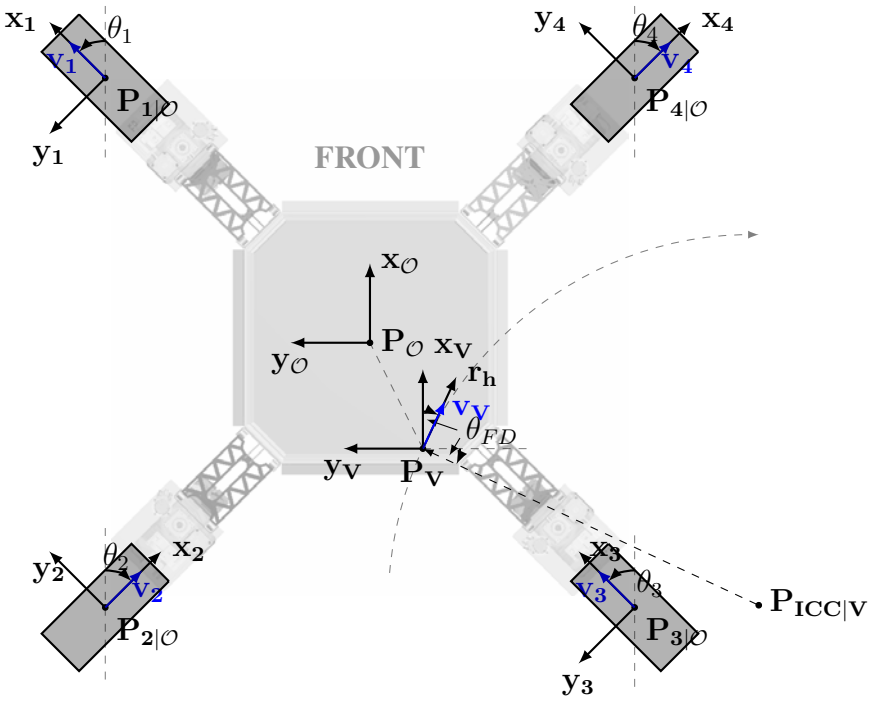


Figure 3: Exemplary kinematic configuration with four independently steerable wheels.

$$\theta_{FD} = \text{atan2}\left(\text{P}_{\text{ICC,x}} + \text{P}_{\text{V,x}}, -\text{P}_{\text{ICC,y}} - \text{P}_{\text{V,y}}\right) \tag{1}$$

$$P_\alpha = (P_{ICC,x} + P_{1,x})^2 + (P_{ICC,y} + P_{1,y})^2 \tag{2}$$

$$P_{n,a} = (P_{ICC,x} + P_{V,x}) \cdot P_{n,y} - (P_{ICC,y} + P_{V,y}) \cdot P_{n,x} \tag{3a}$$

$$P_{n,b} = (P_{ICC,x} + P_{V,x}) \cdot P_{n,x} + (P_{ICC,y} + P_{V,y}) \cdot P_{n,y} - P_\alpha \tag{3b}$$

$$\theta_n = \text{atan2}\left(\text{P}_{\text{n,a}}, \text{P}_{\text{n,b}}\right) + \theta_{\text{FD}} \quad , \tag{3c}$$

With θ_{FD} as the forward direction, $\mathbf{P_{ICC}}$ as the instantaneous center of curvature, $\mathbf{P_v}$ as the origin of the virtual coordinate system, $\mathbf{P_n}$ as the position of the wheel in the kinematic configuration, and $\mathbf{P}_\alpha, \mathbf{P}_{n,a}, \mathbf{P}_{n,b}$ as auxiliary variables. The speed of the single wheels can be calculated based on the distance of the origin of the wheels $\mathbf{P_n}$ to $\mathbf{P_{ICC}}$ with (4a)-(4c).

$$r_n = |\mathbf{P_{ICC}} - \mathbf{P_n}| \tag{4a}$$

$$r_{max} = max(r_n) \tag{4b}$$

$$\mathbf{v}_n = \mathbf{v}_m \cdot \frac{r_n}{r_{max}} \tag{4c}$$

with v_n as the speed of the n^{th} wheel and v_m as the intended maximum speed of the fastest wheel.

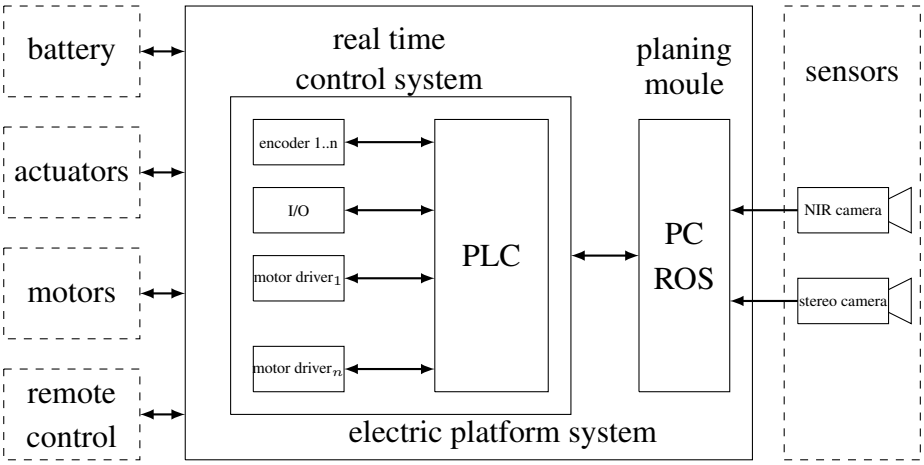


Figure 4: Electrical platform system and the adjacent systems.

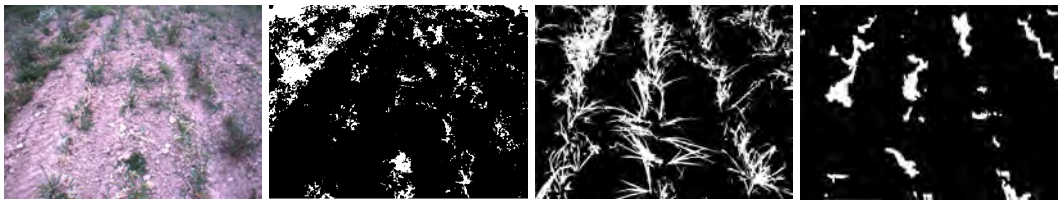
3.2. Electronics and Control System

The vehicle electronics is the bridge between the robot kinematic, including the motors, and the autonomy and row guidance software. Figure 4 shows an overview of the system parts. The necessary sensors system is closely connected to the implemented row guidance system. Based on the review of the prior work [13, 11] we consider that vision systems provide the information for an adaptable navigation and in field task execution. Hence, we approach a vision system that observes light within different ranges of the electromagnetic spectra and is mounted on the robot front. The sensor system consists of two stereo cameras and a NIR camera. A NIR pass filter and the sensitivity of the built in chip form in combination a band pass filter that enables a detection of light from 850nm to 1000nm.

3.3. Row Guidance and Autonomy Software

The row guidance system consists of a segmentation step, followed by a detection of the rows and a parameter extraction. The images are segmented based on NIR and depth data that are provided by the camera system [7]. The extraction of the height information is realised with an online plane calibration that allows determining the camera pose relative to the estimated ground plane.

Several machine vision based row guidance approaches [1, 8, 12] consider pure RGB or NIR information for the segmentation of the plants and soil, while 3D information is omitted and the other way round [9, 14]. Pure RGB-data-based segmentations often fail to segment crops from the soil if they stopped already the production of chlorophyll and lose their green color, while NIR light is still reflected by the cell structure of the leaf (cf. Fig. 5 (b) and (c)). Otherwise, a pure height-based segmentation fails e.g. in early growing stages of the plants, the spectral information can be used as soon as small plants are visible. We approach in [7] a segmentation that fuses both, NIR and depth information together and utilizes the advantages of the one method to compensate the shortcomings of the other. The height information improves the results especially for fields where plants are sowed on dams and allows to filter out small plants and weeds that would add noise to the segmented image (cf. Fig. 5 (d)). Further, the available 3D information enables a projection of the segmentation result to the online estimated ground plane and enables a height-bias-free crop row detection. The row guidance system detects the rows based on a geometric row model and a particle-filter-based row parameter estimation as approached in [7]. The row model describes with three parameters a parallel pattern of lines in the 2D space. The first two parameters α and p represent the 2D normal vector $\mathbf{p}$



(a) (b) (c) (d)

Figure 5: Comparison of different segmentation methods. (a) RGB image, (b) 2G-R-B segmentation [15] , (c) NIR segmentation, (d) NIRD segmentation. (a), (b), (c), and (d) show the same scene under different field of views.

of the closest line and points to the origin of the coordinate system. The third parameter is the scalar d which describes the distance between the lines of the repetitive pattern. The filter samples a 3D parameter space with N hypotheses. Each hypothesis is weighted based on the segmented image. In opposite to other methods the approached crop row detection does not need any prior information on the row structure. Moreover, the particle-filter-inherent properties in combination with the selected geometric row model enable a tracking of the crop rows and improve the results even and especially if natural row irregularities occur. Finally, the negotiable track is extracted out of the row information and is further filtered and processed for the steering information. To achieve the modularity of the whole system, the row guidance is wrapped in the robot operating system (ROS) and can be replaced by another guidance system if necessary. In our recent work we have investigated in [6] how the fusion of odometry and row guidance information can improve the detection results.

4. Tests and Results

As proof of concept we built with the developed subsystems the robotic platform FRANC (cf. Fig. 6). It consists of a frame that carries the electronic and sensor system and is powered with four independent steerable wheels. The algorithms, controller, and the security concept including the remote control with the emergency stop function were implemented to form a whole system with minimal effort.

As stated by [13] the integration task can be a significant effort on its own. The modular concept reduced the integration of the single modules into an overall system to a few mechanical engineering steps as the preparation of the frame including the mounting points and an one-time parametrization of the electrical system and the control algorithms. The parameterizable and adaptable algorithms and interface design simplifies the integration of the subsystems into a working solution and overcomes several integration problems that have to be faced in traditionally designed systems.

FRANC was successfully tested in rough terrain and recorded in-field data for the evaluation of the row guidance algorithm that is used by the autonomy software. The tests proved the feasibility, maneuverability, and rigidity of our modular concept for real-life applications.

The crop row detection algorithm and row guidance software is tested with data recorded during in-field tests of the robot. The robot was maneuvered within row organized fields, parallel to the rows. With this information, parameter windows for p and α can be defined to evaluate the crop row detection algorithm. Correct row structure estimations have to end in a parameter configuration that describe rows within the given windows. Since the row distance has to be constant during

the whole procedure, the error of the row distance estimation e_d is directly determined based on manually measured ground truth data. The particle filter is initialized with $N = 1000$ randomly generated hypotheses that represent parameter configurations with $\alpha = \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $p = [-0.75\text{m}, 0.75\text{m}]$, and $d = [0.2\text{m}, 1.5\text{m}]$.

Figure 6: FRANC during in-field trials.

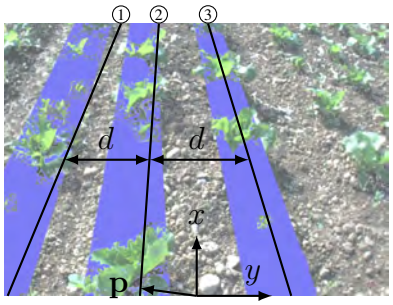


Figure 7: Crop rows and parameter windows.

The parameter windows are defined with $p_w = [0.2\text{m}, -0.2\text{m}]$, $\alpha_w = [+0.2\text{rad}, -0.2\text{rad}]$, and the manually measured ground truth data for the row distance $d_{GT} = 0.45\text{m}$. The experiments show that the particle filter based crop row detection ends in average after five cycles in correct estimations for all three parameters (cf. Fig. 8). The steps within the row offset can be ascribed to the normalization algorithm that searches for the closest line of the pattern to the origin of the coordinate system that was slightly shifted to the right side during the recordings. Hence, the orientation and the offset of the row pattern is either described with line ② or ③ (cf. Fig. 7).

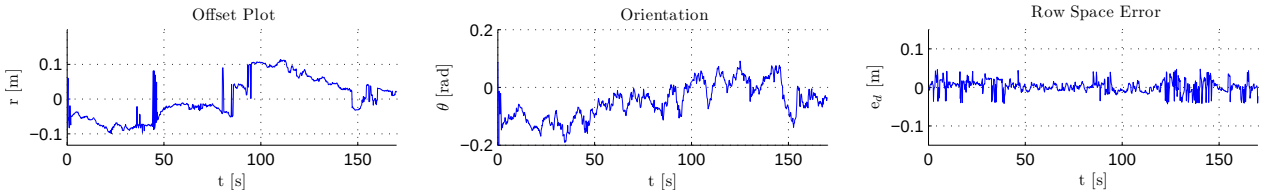


Figure 8: Results of the row detection algorithm with data recorded during in field trials. Offset and orientation has to be within the windows as stated in the text. Average error of the row distance parameter refered to the ground truth.

5. Conclusion

In this article we presented a design concept for a modular agricultural robot and its realisation in the FRANC prototype including results on preliminary field trials.

Testing FRANC in the field proved its maneuverability on rough terrain. The recorded in-field data for the evaluation of the row guidance algorithm revealed that the particle-filter-based crop row detection ends in average after five cycles in correct estimations.

We believe that the conceptual design, its prototypical realization, and the preliminary field trial results presented in this article constitute valuable knowledge for fellow researchers in the field of agricultural robotics and serve as a stepping stone towards developing robotic modules for more flexible agricultural automation.

References

[1] B Åstrand and A Baerveldt. A vision based row-following system for agricultural field machinery. *Mechatronics*, 15(2):251 – 269, 2005.

[2] T Bak and H Jakobsen. Agricultural robotic platform with four wheel steering for weed detection. *Biosystems Engineering*, 87(2):125–136, 2004.

[3] W Bangert, A Kielhorn, F Rahe, A Albert, P Biber, S Grzonka, S Haug, A Michaels, D Mentrup, M Hänsel, et al. Field-robot-based agriculture:"remotefarming. 1" and "bonirob-apps". *VDI-Berichte*, (2193):439–446, 2013.

[4] G Dudek and M Jenkin. *Computational principles of mobile robotics*. Cambridge university press, 2010.

[5] J Gomez-Gil, R Ruiz-Gonzalez, S Alonso-Garcia, and FJ Gomez-Gil. A kalman filter implementation for precision improvement in low-cost gps positioning of tractors. *Sensors*, 13(11):15307–15323, 2013.

[6] G Halmetschlager, J Prankl, and M Vincze. Increasing the precision of generic crop row detection and tracking and row end detection. *In G. Kootstra, Y Edan, E van Henten, and M Bergerman (Eds.), Proceedings of the IROS Workshop on Agri-Food Robotics. Hamburg, October 2, 2015.*

[7] G Halmetschlager, J Prankl, and M Vincze. Probabilistic near infrared and depth based crop line identification. In *IAS-13, Workshop on Recent Advances in Agricultural Robotics, Workshop Proceedings of IAS-13 Conference on*, pages 474–482, 2014.

[8] Guo-Quan Jiang, Cui-Jun Zhao, and Yong-Sheng Si. A machine vision based crop rows detection for agricultural robots. In *Wavelet Analysis and Pattern Recognition (ICWAPR), 2010 International Conference on*, pages 114–118, 2010.

[9] M Kise, Q Zhang, and F Rovira Más. A stereovision-based crop row detection method for tractor-automated guidance. *Biosystems Engineering*, 90(4):357–367, 2005.

[10] R Lenain, B Thuilot, C Cariou, and P Martinet. High accuracy path tracking for vehicles in presence of sliding: Application to farm vehicle automatic guidance for agricultural tasks. *Autonomous robots*, 21(1):79–97, 2006.

[11] Ming Li, Kenji Imou, Katsuhiro Wakabayashi, and Shinya Yokoyama. Review of research on agricultural vehicle autonomous guidance. *International Journal of Agricultural and Biological Engineering*, 2(3):1–16, 2009.

[12] J Romeo, G Pajares, M Montalvo, JM Guerrero, M Guijarro, and A Ribeiro. Crop row detection in maize fields inspired on the human visual perception. *The Scientific World Journal*, 2012, 2012.

[13] D.C. Slaughter, D.K. Giles, and D. Downey. Autonomous robotic weed control systems: A review. *Computers and Electronics in Agriculture*, 61(1):63 – 78, 2008.

[14] U Weiss and P Biber. Plant detection and mapping for agricultural robots using a 3d lidar sensor. *Robotics and Autonomous Systems*, 59(5):265 – 273, 2011. Special Issue ECMR 2009.

[15] DM Woebbecke, GE Meyer, K Von Bargen, and DA Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995.

[16] Yin Xiang and Noboru Noguchi. Development and evaluation of a general-purpose electric off-road robot based on agricultural navigation. *International Journal of Agricultural and Biological Engineering*, 7(5):14–21, 2014.

[17] S Yaghoubi, NA Akbarzadeh, SS Bazargani, M Bamizan, and MI Asl. Autonomous robots for agricultural tasks and farm assignment and future trends in agro robots. *International Journal of Mechanical & Mechatronics Engineering IJMME-IJENS*, 13(03):1–6, 2013.

WS 6: Vision for Robotics II

A Step Forward in Human-Robot Collaboration – The Project CollRob *

Rosemarie Velik[1], Bernhard Dieber[1], Saeed Yahyanejad[1], Mathias Brandstötter[1], David Kirschner[1], Lucas Paletta[2], Ferdinand Fuhrmann[2], Patrick Luley[2], Herwig Zeiner[2], Gerhard Paar[2], and Michael Hofbaur[1]

[1] Institute for Robotics and Mechatronics
JOANNEUM RESEARCH, Austria
first.last@joanneum.at

[2] Institute for Information and Communication Technologies
JOANNEUM RESEARCH, Austria
first.last@joanneum.at

Abstract
*Human-robot collaboration is a novel, hot topic in the field of industrial and service robotics with considerable potential. It offers the possibility to combine human cognitive abilities with the strengths of robot technology in terms of precision and performance, thus opening up a wide range of possibilities beyond the traditional application of robots. The research project "**Collaborative Robotics" (CollRob)** is an initiative focusing on the conceptualization, research, development, and evaluation of novel methods and tools for collaborative and cooperative robots. This article aims at giving an overview about this project in terms of its backgrounds, objectives, and the current status of research covering topics such as machine perception, sensitive redundant kinematic manipulation, dynamic adaptive planning, human-robot interaction and information exchange, human factors, and safety.*

1. Introduction

Since the introduction of robots to factories, approximately 50 years ago, their strength has been to perform well specified and repetitive tasks in constrained environments. Due to the high configuration and programming efforts in addition to significant investment costs, such an approach can only be profitable at large scales. However, current developments in modern production show a trend towards individualized manufacturing and thus small series in robot exploitation. Accordingly, novel innovative strategies and methods are necessary to allow for more flexibility in the production environment. In this context, one highly promising approach is human-robot collaboration [9]. According to this concept, humans and robots shall be enabled to jointly work together in the production process in order to combine human cognitive abilities with the strengths of robot technology in terms of precision and performance [2]. To make such a collaboration efficient and safe, a large range of challenges has to be addressed including topics like machine perception, sensitive redundant kinematic manipulation, dynamic adaptive task planning, human-robot interaction and information exchange,

*This work has been supported by the Austrian Ministry for Transport, Innovation and Technology (bmvit) within the project framework Collaborative Robotics.

human state evaluation, and safety standards. These challenges are addressed in the 4-year research project "Collaborative Robotics" (CollRob), launched in 2015[1]. In this article, an overview about the backgrounds, objectives and current status of this project is given, which shall serve as a reference for proceeding publications and research initiatives. Chapter 2. describes the elaborated hardware and software architecture of the overall CollRob system. Chapter 3. presents different specified levels of complexity for human-robot collaboration as well as envisioned use cases to test developed concepts and methods. Chapter 4. outlines concrete research challenges addressed within CollRob. Finally, Chapter 5. gives a conclusion.

2. Robot System Hardware and Software Architecture

To enable collaboration between a human and a robot, a variety of sensors is required. Figure 1(a) shows an overview of the CollRob hardware setup. Besides the robot itself, which is equipped with torque sensors, we use dedicated visual and proximity sensors for i) monitoring of the workspace and ii) detection and tracking of the human. In addition, we use wearable bio-sensors and eye tracking glasses to monitor the behavior and state of the human. A tablet, augmented reality (AR) glasses, microphones, speakers, and gesture bracelets are used as human-robot interaction devices.

The broad range of different sensors and devices results in complex data-flows, which in turn cause strong dependencies and couplings between the individual application parts. A software architecture for an application like this needs to relax the strong dependencies in order to enable re-usability, scalability and easy exchange of individual modules. Publish/subscribe has proven to be a well suited architectural pattern to accomplish this decoupling. An application is composed of modules which provide data (publish) and consume data (subscribe) from others. However, the transport of this data is handled by a dedicated infrastructure such that the modules need not know providers and consumers of their data. Figure 2 shows the modules which compose the CollRob system. In our architecture, each module is modelled as publisher and/or subscriber. We realize the publish/subscribe system using the Robot Operating System (ROS).

3. Levels of Complexity of Human-Robot Collaboration and Use Cases

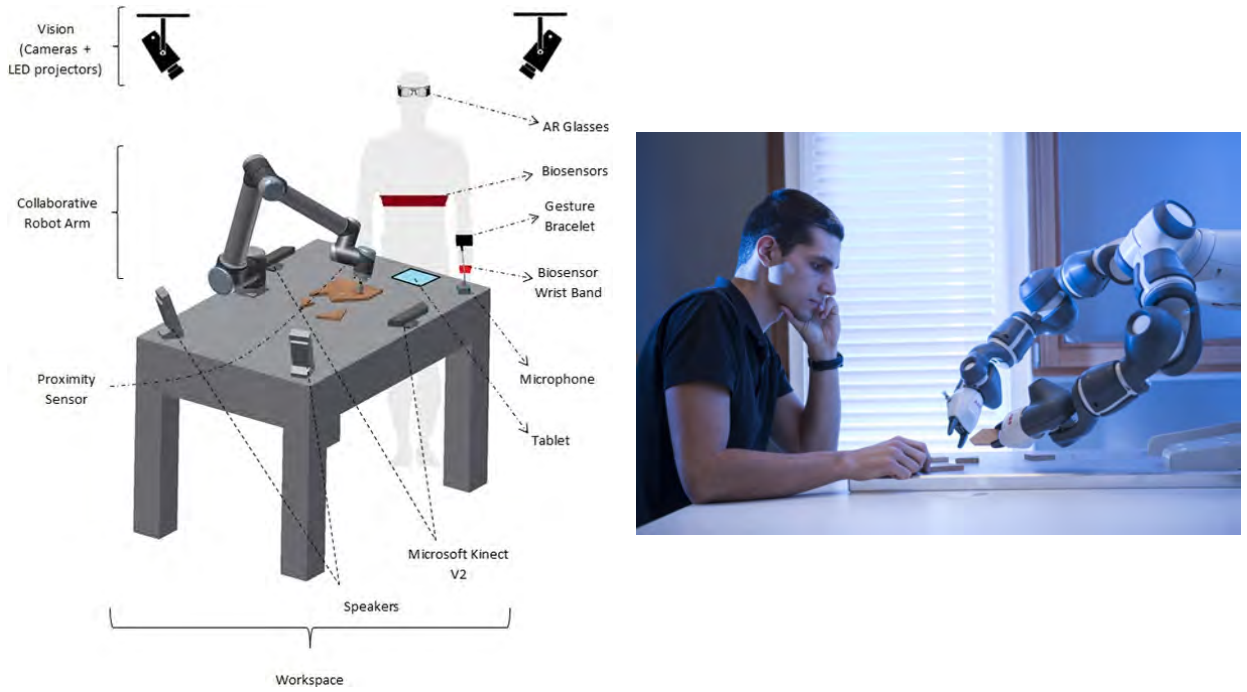
Within the CollRob project, different levels of complexity of collaboration have been specified (see Table 1). The category A is in fact non-collaborative. However, taking it into account can be useful for setting up the general CollRob system architecture before going into collaboration details. The categories B to D consider human-robot interaction with gradually increasing complexity. Category E describes the case of two collaborating robots (or one robot with two arms) and category F the interaction between two robots and one or more humans.

category	**A**	**B**	**C**	**D**	**E**	**F**
umbrella term	encapsulation	H-R co-existence	static H-R collaboration	dynamic H-R collaboration	static/ dynamic R-R collaboration	static/ dynamic H-R-R collaboration
interaction level	interaction-free operation	safety stop	static collaboration	dynamic collaboration	static/ dynamic R-R collaboration	static/ dynamic H-R-R collaboration
actors	robot	human+ robot	human + robot	human + robot	2 robots	2 robots + human(s)
temporal dependence	independent	interrupt	sequential	simultaneous	sequential/simultaneous	sequential/simultaneous
spatial dependence	separated	separated	shared	shared	shared	shared
human-robot contact	none	rudimentary	pronounced	comprehensive	n.a.	pronounced/comprehensive

Table 1. Overview of interaction categories

Concerning the described categories, different application domains will be addressed in CollRob

[1] http://www.joanneum.at/en/robotics/reference-projects/collrob.html



(a) Overview of hardware setup

(b) Human-robot collaboration to build a Tangram puzzle

Figure 1. Overview of hardware setup for puzzle solving

(e.g., industrial applications, entertainment applications, service applications, assistive technology applications). Concerning the applications of choice, it was decided that at least one ”set of use cases” should be chosen for which it was possible to carry them throughout all possible categories (A to D plus optionally E and F) by continuously extending and adding ”human-robot collaboration features”. The use case set of choice for this purpose is to solve a Tangram puzzle while the robot and the human are cooperating toward the goal (see Figure 1(b)). Further use cases, in collaboration with industrial partners, address industrial applications such as human-robot joint assembly and inspection tasks.

4. Addressed Research Challenges

4.1. Dynamic Working Environment Monitoring and Safety

In order to perform a safe and reliable collaborative task in complex and dynamic environments, a robust monitoring system, which provides a real-time status of the target area, is required. Methods from 2D [14] and 3D [24] quality inspection (and combined [29]) can be applied to robotic scenarios as long as sensors are lightweight and compact enough to be used on moving parts of the robot. 2D inspection is sufficient in case of close-to-planar objects. 3D shape comparison e.g., by Iterative Closest Point (ICP) [23], is a robust method in industrial inspection for irregular objects. Texture-based methods are complementary in case of smooth 3D surfaces that do not allow a precise 3D alignment due to shape ambiguities [5]. One major current application for such techniques is the inspection of correct 3D shape, the automatic planning of grasping is still a very rare application under real production conditions. Within CollRob, we use various optical sensors for localization, detection, 3D reconstruction, and depth estimation. The overall system has to cope with challenges such as submillimeter position accuracy for robot grasping, occlusion, surface reflections and complex shapes.

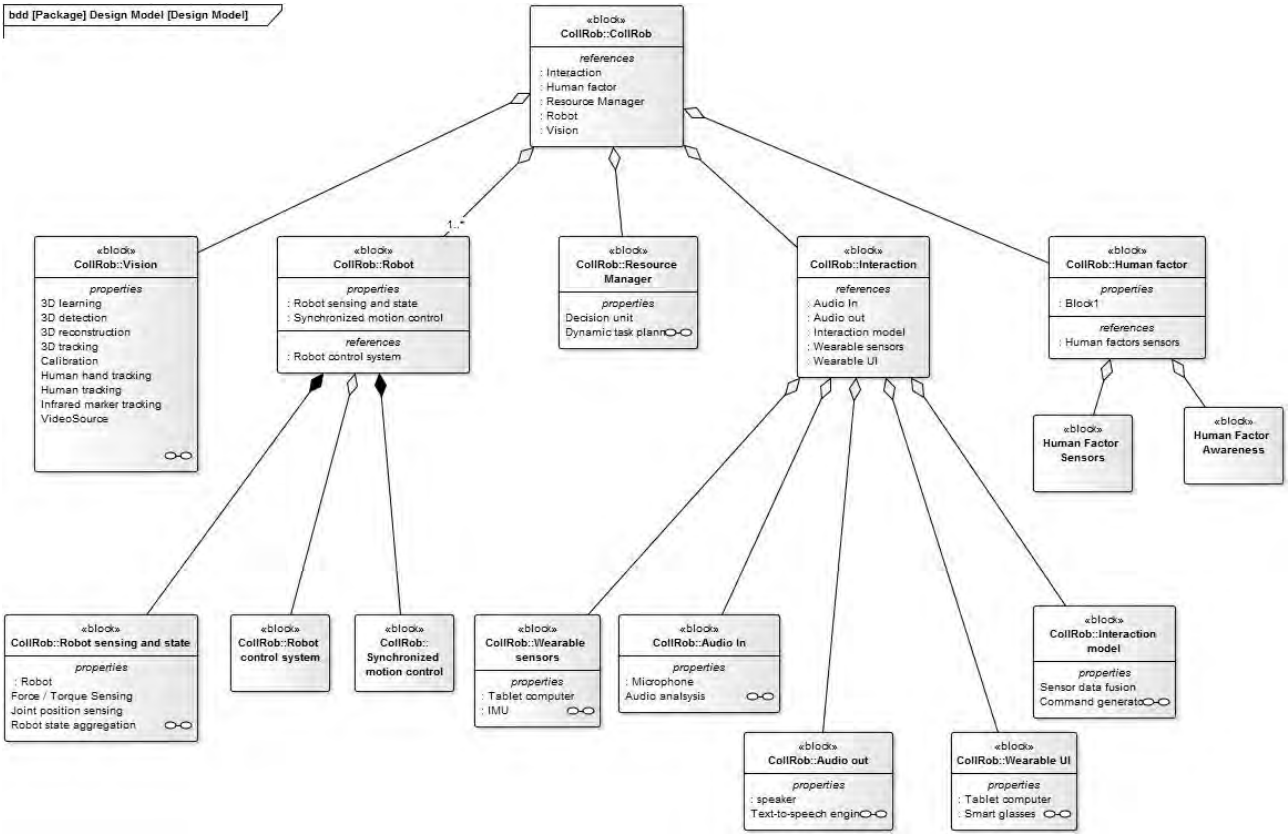


Figure 2. Structure of application modules

The system design fuses the information acquired from the various sensors: (1) Laser scanning (stationary) is used for large-scale mapping of workspace environment. (2) A 3D-Time-of-Flight (TOF) sensor is used for rough dynamic characterization of the scene. Alternatively, an additional stereo camera system with wider FoV can be used. (3) For high-resolution localization & inspection within the workspace of the robot (in order to achieve the required 0.05 mm resolution at specific parts of the scene), an active stereo system with an additional pattern projection unit is used. (4) On one of the Pan-Tilt Units (PTUs) used for sensor pointing, in addition a laser speckle projector is mounted to enhance texture-less regions for high-resolution stereo analysis. (5) All components of the sensor system will be integrated in one unit which allows quick installation and setup in the production environment.

We can benefit from such a system along with other sensory data for safety assurance [16]. Safety is one of the most important factors when considering human-robot collaboration in industrial applications. Various safety features such as vision, proximity sensors [33], laser detectors, touch and collision sensors, force torque sensors, and emergency stops could be exploited to achieve this goal. At the same time, many standards and guidelines throughout the design, robot manufacturing, installation, and final implementation are issued to increase the safety in the system [15, 1]. To provide safety, we need to cope with the sensor failures and their occlusion and also dynamic environments. To achieve this goal, we plan to build a hybrid safety system, which combines multiple safety features and sensors together in multiple layers in a both serial and parallel manner. This way, failure of one sensor will not necessarily compromise the safety as long as other components continue to function.

4.2. Analysis of Human Behaviors, Emotions and Actions

Humans generally interact with robots in the same way they might interact with other people, establishing social relationships and emotional ties with them [4, 8]. As industrial robots are enabling human and robot workers to work side by side as collaborators in manufacturing tasks, a fundamental issue regards the development of methods to assess the user's experience with a robot, while understanding how humans feel during their interaction with it [27]. Furthermore, human-related variables are essential for the evaluation of human-interaction metrics [26]. To work seamlessly and efficiently with their human counterparts, robots must similarly rely on predictions of the human worker's behavior, his/her emotions, task specific actions and intent to plan their actions. In [12] for instance an anticipatory control method using a human-in-the-loop architecture was implemented that enables robots to proactively perform task actions based on observed gaze patterns to anticipate actions of their human partners according to its predictions.

In the project CollRob, there is a focus on advancing models of human-related variables that directly refer to the evaluation of levels of autonomy in human-robot interaction, such as situation awareness, trust and workload, which have a long history in the automation literature [3, 6, 31]. CollRob undertakes to elaborate situation awareness in the manufacturing domain of human-robot interaction (HRI) on the basis of human attention measures. It specifically considers the dynamic estimation of current and predicted gaze in the context of collaboration affordances. Affordances have already been thoroughly studied in robot control [22]. However, from the human worker's viewpoint in the manufacturing domain, affordances refer to relations between the human and the manufacturing environment that, through a collection of stimuli, afford the opportunity for the worker to perform an interaction. CollRob intends to estimate various levels of human attention in the 3D environment [20] – in the context of collaboration affordances – and from this become capable to derive parameters for decision making: as low levels of situation awareness would decrease speed in safety-critical task processing, high levels would need to increase the throughput or to increasingly consider production quality related processing. As a first step, CollRob developed methodologies for the efficient, robust and low-cost method for the continuous localization of human gaze in industrial work cells. One application is to estimate gaze directly from eye tracking glasses based on the visual recognition of artificial random dot markers [28]. Additionally, a spatiotemporal model of attention was developed that estimates human gaze solely from egocentric vision [21]. Further activities in the frame of human behavior measurements will focus on the worker's context being estimated from psychophysiological measurements [27] and developing a metric for HRI situation awareness in the manufacturing domain.

4.3. Resource Managing including Dynamic Task Optimization and Decision Making

Human-robot collaboration requires robust and time-aware dynamic planning and scheduling strategies for robot-human teams. In the last few years, key contributions to make robot-human team collaboration more fluent stem from [25, 18]. Within the CollRob project, we will focus on the development of algorithms to deal with geometric issues, consider time based planning strategies and provide a robust implementation. Currently, our focus is to find a robot model for geometric and time-aware scheduling strategies by building a puzzle together. Our long term research goal is to deeply integrate the social interaction models (e.g., the fair distribution of team members, conflict solution strategies etc. for individual team members). A key issue will now be how we model such aspects.

4.4. Human-Robot Interaction and Information Exchange

Current research on HRI investigates different ways how robots and humans interact, the main ones being voice control [13], interaction primitive [7], motion recognition [10], force adaptation [17] and shared presence [30]. However none of these methods is well established in commercial applications. In this project, we focus on new paradigms of HRI in the context of collaborative industrial robotics and emphasize the distinction between collaboration and other forms of human-robot interaction, which usually view the problem as robot control or human-robot communication via tele-operation [11]. We include natural interaction mechanisms (acoustic and gestural interface), human factors [32] and a visualization component supported by augmented reality functionalities in an intelligent, context-sensitive control system. While human collaborators primarily interact using speech and gestural input, they receive situation-dependent information about current and future tasks, the robot's movement path and possible dangers. Here we explore the use of a head-worn AR system to ensure unobtrusive, hands-free collaboration and explore the optimal information flow to avoid cognitive load and distraction from the task. Moreover, the physical state of the human affects both the robot behavior and the feedback channel. This all strengthens the collaborative aspect of the interaction by increasing communication quality, trust, security awareness, and work efficiency. Recently, a first implementation of the interaction system was set up including the speech interface, a basic dialogue manager and a tablet application, serving as basic sensor interface. The human collaborator is thus able to interact with the robot using speech input, including basic robot task queue manipulation commands, while monitoring the sensory data. As a next step, the acoustic interface will be extended to include natural language understanding together with a more advanced dialogue manager. Moreover, a tablet application and wearable sensors will be used to analyse behaviors, emotions and actions of human collaborators. We will use the results of this analysis to implement the context-sensitive visualization and control system.

4.5. Redundant Sensitive Robotic Manipulation

Whenever a physical human-robot interaction is supposed to take place to fulfill a shared task, human ergonomic operation and safety are important aspects which must be observed. Robot safety is provided by the electromechanical system in different ways and often in a redundant fashion. In practical terms, the main options to reduce the risk of human injuries are safety-related monitored stops, speed and separation monitoring, and power and force limitations, as manifested in [ISO/TS 15066:2016]. The implementation of these options is done (i) by measuring the direct energy transfer between the robot and an object or (ii) by monitoring the environment using electromagnetic or sound based sensors. More specifically, a physical contact can be recognized by measuring the force, torque or current at the end effector, the robot's base, or at each joint. In addition, a sensitive skin applied on the manipulator can measure a contact force as well. If a direct contact is not desired, the environment can be perceived by different sensors operating at distance (see Section 4.1.). All sensor data can be fused to expand the knowledge of the environment, thus being used to control the robot's movement in a human-safe manner. That means that the dynamic movement of the robot must be planned and executed in an adaptive and reactive way.

If a kinematically redundant system has to perform a given task, the additional freedom can be used to enhance safety (by increasing the distance of the manipulator's parts to a human or reducing the velocity of the robot's arm segments) and ergonomics for a human operator (by configuring the robot joints in a way that the robot does not disturb ergonomic human motion). Such systems allow a

change of the manipulator configuration without influencing the end effector's trajectory.

Although the method of redundancy resolution is well understood for local optimization using the Jacobian matrix (see, e.g., [19]), we try to expand the knowledge of redundant robot systems onto a global view on redundancy. Based on this goal, cost functions for any type of manipulators (e.g., mobile, serial) can be formulated and computed for the entire system. Moreover, a compliance control scheme should be developed which uses this extensive description of the kinematic behavior. In case of a higher number of freedoms in the system (more than one) a multi-priority control can be used in a meaningful way. Thus, the aim of this work package is to realize the computation of a primary task (given trajectory of the end effector), a compliance control of the whole robot system and a desired optimization (e.g., energy minimization) based on one mathematical concept.

5. Conclusion

In this article, the research project CollRob was presented proposing novel methods for human-robot collaboration. Research covers machine perception, sensitive redundant kinematic manipulation, dynamic adaptive planning, human-robot interaction and information exchange, human state evaluation, and safety. To integrate different methods and system components into one working setup, a software architecture has been developed based on a publish-subscribe principle implemented in ROS. Different use cases have been identified and will be addressed for evaluation purposes of the developed methods including a demo use case of joint human-robot Tangram puzzle building and various industrial applications in collaboration with company partners.

References

[1] ISO 13849-1: 2015, Safety of machinery – Safety-related parts of control systems.

[2] *Robotiq Collaborative Robot Ebook*. Robotiq, 2016. 6th edition.

[3] Jenay Beer, Arthur D Fisk, and Wendy A Rogers. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, 3(2):74, 2014.

[4] Cynthia Breazeal. Toward sociable robots. *Robotics and autonomous systems*, 42(3):167–175, 2003.

[5] Sébastien Druon, Marie-José Aldon, and André Crosnier. Color constrained icp for registration of large unstructured 3d color data sets. In *Information Acquisition, 2006 IEEE International Conference on*, pages 249–255. IEEE, 2006.

[6] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[7] Marco Ewerton, Gerhard Neumann, Rudolf Lioutikov, Heni Ben Amor, Jan Peters, and Guilherme Maeda. Learning multiple collaborative tasks with a mixture of interaction primitives. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1535–1542. IEEE, 2015.

[8] David Feil-Seifer and Maja J Matarić. Defining socially assistive robotics. In *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, pages 465–468. IEEE, 2005.

[9] Michael A Goodrich and Alan C Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[10] Yihsin Ho, Yoshihiro Kawagishi, Eri Sato-Shimokawara, and Toru Yamaguchi. A human motion recognition using data mining for a service robot. In *Advanced Robotics (ICAR), 2011 15th International Conference on*, pages 229–234. IEEE, 2011.

[11] Guy Hoffman and Cynthia Breazeal. Collaboration in human-robot teams. In *Proc. of the AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, USA*, 2004.

[12] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration.

[13] Simon Keizer, Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, and Oliver Lemon. Handling uncertain input in multi-user human-robot interaction. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 312–317. IEEE, 2014.

[14] Elias N Malamas, Euripides GM Petrakis, Michalis Zervakis, Laurent Petit, and Jean-Didier Legat. A survey on industrial vision systems, applications and tools. *Image and vision computing*, 21(2):171–188, 2003.

[15] George Michalos, Sotiris Makris, Panagiota Tsarouchi, Toni Guasch, Dimitris Kontovrakis, and George Chryssolouris. Design considerations for safe human-robot collaborative workplaces. *Procedia CIRP*, 37:248–253, 2015.

[16] Carlos Morato, Krishnanand N Kaipa, Boxuan Zhao, and Satyandra K Gupta. Toward safe human robot collaboration by using multiple kinects based real-time human tracking. *Journal of Computing and Information Science in Engineering*, 14(1):011006, 2014.

[17] Bojan Nemec, Andrej Gams, Miha Denisa, and Ales Ude. Human-robot cooperation through force adaptation using dynamic motion primitives and iterative learning. In *Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on*, pages 1439–1444. IEEE, 2014.

[18] Masahiro Ono, Brian C Williams, and Lars Blackmore. Probabilistic planning for continuous dynamic systems under bounded risk. *Journal of Artificial Intelligence Research*, pages 511–577, 2013.

[19] Christian Ott, Alexander Dietrich, and Alin Albu-Schäffer. Prioritized multi-task compliance control of redundant manipulators. *Automatica*, 53:416–423, 2015.

[20] Lucas Paletta, Katrin Santner, Gerald Fritz, Heinz Mayer, and Johann Schrammel. 3d attention: measurement of visual saliency using eye tracking glasses. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 199–204. ACM, 2013.

[21] Patrik Polatsek, Wanda Benesova, Lucas Paletta, and Roland Perko. Novelty-based spatiotemporal saliency detection for prediction of gaze in egocentric video. *IEEE Signal Processing Letters*, 23(3):394–398, 2016.

[22] Erich Rome, Lucas Paletta, Erol Şahin, Georg Dorffner, Joachim Hertzberg, Ralph Breithaupt, Gerald Fritz, Jörg Irran, Florian Kintzler, Christopher Lörken, et al. The MACS project: an approach to affordance-inspired robot control. In *Towards affordance-based robot control*, pages 173–210. Springer, 2008.

[23] S Rusinkiewicz and M Levoy. Efficient variants of the icp algorithm. In *Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152, 2001.

[24] Giovanna Sansoni, Marco Trebeschi, and Franco Docchio. State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation. *Sensors*, 9(1):568–601, 2009.

[25] Julie A Shah. *Fluid coordination of human-robot teams*. PhD thesis, Massachusetts Institute of Technology, 2011.

[26] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40. ACM, 2006.

[27] Lorenza Tiberio, Amedeo Cesta, and Marta Olivetti Belardinelli. Psychophysiological methods to evaluate userÆs response in human robot interaction: a review and feasibility study. *Robotics*, 2(2):92–121, 2013.

[28] Hideaki Uchiyama and Hideo Saito. Random dot markers. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 35–38. IEEE, 2011.

[29] George Vosselman and Johan WH Tangelder. 3d reconstruction of industrial installations by constrained fitting of cad models to images. In *Mustererkennung 2000*, pages 285–292. Springer, 2000.

[30] Johann Wentzel, Daniel J Rea, James E Youn, and Ehud Sharlin. Shared presence and collaboration using a co-located humanoid robot. In *Proceedings of the 3rd International Conference on Human-Agent Interaction*, pages 273–276. ACM, 2015.

[31] Chris Wickens, Jason McCarley, Lisa Thomas, Model In DC Foyle, A Goodman, and BL Hooey. Attention-situation awareness (a-sa) model. In *NASA Aviation Safety Program Conference on Human Performance Modeling of Approach and Landing with Augmented Displays*, page 189, 2003.

[32] Christopher D Wickens, John D Lee, Yili Liu, and Sallie Gordon-Becker. Introduction to human factors engineering. 1998.

[33] Martin Zirkl, Anurak Sawatdee, Uta Helbig, Markus Krause, Gregor Scheipl, Elke Kraker, Peter Andersson Ersman, David Nilsson, Duncan Platt, Peter Bodö, et al. An all-printed ferroelectric active matrix sensor network based on only five functional materials forming a touchless control interface. *Advanced Materials*, 23(18):2069–2074, 2011.

Industrial Grasping - An Autonomous Order Picking System*

Julia Nitsch and Gerald Steinbauer

Institute for Software Technology
Graz University of Technology, Austria
{*jnitsch,steinbauer*}*@ist.tugraz.at*

Abstract
Automated storing, retrieving, and delivering items is an important part of Industry 4.0 application. For low-volume this task is done usually manual. In this paper we present an architecture and a proof-of-concept implementation for order picking using the robot Baxter from Rethink Robotics. The main contribution besides providing full functioning prototype is a dependable control architecture.

1. Introduction

Industry 4.0 is one of the keywords, when we talk about the next level of production. Industry 4.0 represents the 4^{th} industrial revolution and promises improvement of productivity through automated, self-organizing and self-optimizing processes. It addresses the needs of high-quality products which are also highly customized but still ready for mass production.
This work contributes to the field of Industry 4.0 by developing an assistant robot for order picking. Such robots share the environments with humans. In a typical warehouse system items can be stored in larger transport boxes. The transport boxes again can be stored in shelves to save space. If a specific item needs to be picked the transport box first needs to be pulled out of the shelf and then the item can be picked and delivered. This procedure is called order picking. For items with a moderate frequency this type of picking is usually done by hand which is a monotonic and time consuming task. In our scenario we tend to automatize that task.
The system we propose is based on a 3-TIER architecture. The planning layer uses an artificial intelligence (AI) planner to generate a list of skills the robot has to execute. The planner outputs a list of skills, the robot needs to execute in order to achieve its goal. Skills are composed of skill primitives. These primitives can perform perception, manipulation, grasping tasks or any combination of those. Failures are already detected at the level of the primitives where local recoveries can be performed. If these recoveries fail too, these errors are reported to the executive layer. This architecture ensures the detection and recognition of failures. Together with appropriate steps for recovery dependable execution is achieved. The proposed architecture was realized as a proof-of-concept implementation using the two arm robot Baxter from Rethink Robotics. For details about the realized system we refer the interested reader to [10].
The reminder of this paper is organized as follows. In the next sections we briefly discuss related research and the target environment. In Section 4. the proposed system architecture is presented. Due to the space constraints we focus on skill primitives. In the next section we briefly present an evaluation focused on the skill primitives. In section 6. we draw some conclusions.

*This work was supported by incubed IT GmbH.

2. Related Work

Numerous works exist about high-level planning for robot systems solving complex tasks using sets of simpler system capabilities. This system capabilities are called skills. Dividing a complex task in such skills has multiple benefits like flexibility, reuse of skills and good software portability.
Pederson et al. shows in [13] the division of a complex task a sequence of multiple subtasks (= skills). Skills are described being the *fundamental building blocks* or the *system capabilities*. If a new complex task should be executed, the system needs not being reprogrammed. It is sufficient to simply reorder the skills. Similar to our approach, an execution monitor surveys the outcome of the skills.
In [12] skills are ported to different robotic platforms. Skills get further decomposed into skill primitives. With this detailed decomposition the hardware level is abstracted from the skills itself. The advantage of modularity and the abstraction of tasks is pointed out.
The authors in [14] introduce a 4-TIER architecture, with the same idea of abstraction for skills and skill primitives as in the previous papers. The lowest layer ensures the hardware abstraction and so the re-usability on different platforms. The next layer contains action and perception primitives. The top layers handles the planning task. As the previous addressed work, this abstraction is used for easing the human robot-interaction. All these papers show a clear distinction between tasks, skills and primitives and focus on portability and easy execution of complex new tasks. But their focus is on human-robot interaction. The human in the loop defines a new task through reordering skills. The next works present a successful task planning utilizing artificial intelligence (AI) planner instead of humans in the loop ordering skills. In [6] Huckaby defined skills with preconditions and effects in the model space of the problem. The initial state and goal are stated in the process space. They proposed PDDL [7] as planning language.
In [6] the focus lies on the high-level. It is assumed that skills and their primitives always succeed. In [11] a system is proposed which transfers the high-level description from the AI planner to a behavioral state machine. Failures in the primitive execution are detected by a vision system and recoveries are performed.
Finally some works addressing the order picking problem are discussed. The authors in [9] present a mobile bin picking system. Items are picked from a box standing on the ground and placed at a delivery station. The high-level of this system is a finite state machine (FSM).
In [3] a software architecture and their implementation for grasping objects is presented. Some of these concepts are used in our work too. The collision environment is a 3D occupancy grid excluding robot parts. Known and recognized objects are represented as geometric primitives or as mesh models of the objects. In [1] the authors present a pick and place approach where they have to deal with known and unknown objects, cluttered workspace and noisy sensor data.

3. Target Environment and System

For the a proof-of-concept implementation of the proposed order picking system we use the robot Baxter from Rethink Robotics (see Fig. 1a). It is a two-arm robot with internal sensors such as cameras and proximity sensors in the wrists. In order to get a global overview of the environment we added a RGBD camera on top. The environment Baxter operates in is shown in Figure 5b. It is a mock-up of a typical manual storage.

(a) Baxter with its inbuilt and additional mounted sensor.

(b) Environment in which Baxter performs the order picking task.

Figure 1: Robot Baxter and the environment it operates in.

4. Architecture

In Figure 2 the conceptional overview of the proposed 3-TIER architecture is shown. The 3 layers for planning, executive and behavioral control are separated. The communication is clearly defined. The top layer represents the planning layer. The planner uses the information of the domain and the problem to generate a plan. The plan is a sequence of skills that have to be executed to reach a given goal. The plan is forwarded to the next layer. The executive layer takes care of the execution of each skill. It knows about the composition of the skill primitives. The primitives are located in the behavioral layer. The advantage of this abstraction is its clear structure and its modularity. For further reading about the 3-TIER architecture please see [8, p. 244–277].

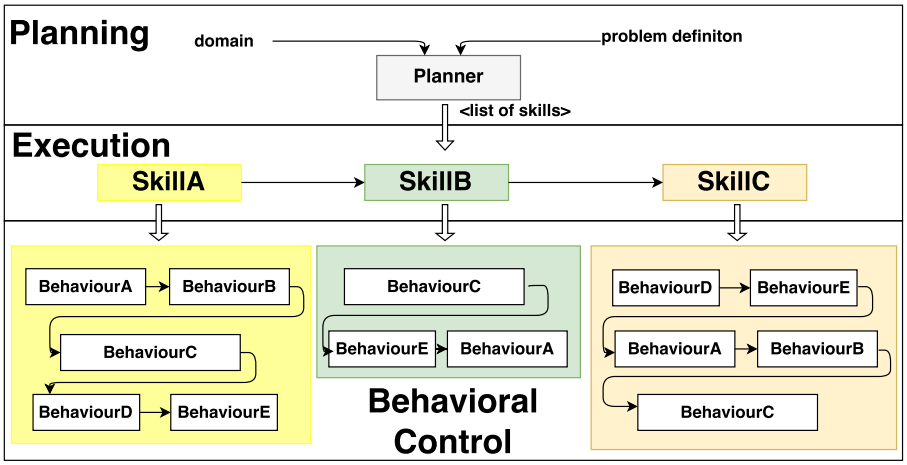


Figure 2: Overview of proposed system's architecture.

4.1. Planning Layer

The top layer of the 3-TIER architecture is the planning layer. The planning layer uses a domain and problem description of the given environment and task. It is based on the Planning Domain Definition Language (PDDL) modeling the system capabilities (further on called skills), the current state of the environment and the goal state. The authors of [6] showed that PDDL is an appropriate choice for

robotic tasks. The domain contains information about all the objects which can appear in the environment. Further it knows about the skills the robot is able to perform. A skill is defined through its name and the its parameters. A skill has a precondition and an effect. The precondition is the state the environment needs to be in before the action can be performed. The effect of a skill is the description of the environment after the skill is performed.
The problem describes the initial state s_0 and the goal g, which are a set of propositions [4]. The goal state s_g is a state, that satisfies g. First object instances are defined, which occur in the environment and their initial properties are stored. The planner takes the domain and the problem description and generates a list of skills, which need to be performed to solve the given problem. We use the planner SGPlan6 [5]. This list of skills is forwarded to the executive layer of the 3-TIER architecture.
For solving the order picking task, the following skills are required: *moveBoxFromLevelToTray*, *movBoxFromTrayToLevel* and *graspItem*. Lets assume an environment containing transport boxes *BOX_A*, *BOX_C* and a shelf with levels *LEVEL_1*, *LEVEL_2* and the goal of picking one item from *BOX_C* placing it at the delivery box *DBOX_C* and picking two items from *BOX_A* placing it at delivery box *DBOX_A*. The planner comes up with the following plan (see Listing 1).The name of the skill is the first parameter, followed by the parameters the skill requires. So the first skill, which has to be performed is *moveBoxFromLevelToTray*. The *BOX_C* is moved from *LEVEL_2* to the *TRAY*.

Listing 1: Output of planner for example domain and problem.

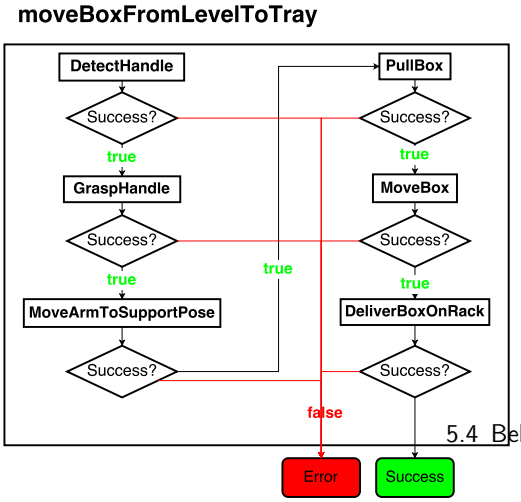
```
0 (MOVEBOXFROMLEVELTOTRAY BOX_C LEVEL_2 TRAY)
1 (GRASPITEM BOX_C DBOX_C TRAY)
2 (MOVEBOXFROMTRAYTOLEVEL BOX_C LEVEL_2 TRAY)
3 (MOVEBOXFROMLEVELTOTRAY BOX_A LEVEL_1 TRAY)
4 (GRASPITEM BOX_A DBOX_A TRAY)
5 (GRASPITEM BOX_A DBOX_A TRAY)
6 (MOVEBOXFROMTRAYTOLEVEL BOX_A LEVEL_1 TRAY)
```

4.2. Executive Layer

The executive layer receives a list of skills from the planner. The executive layer handles the execution of single skills. Each skill is composed of skill primitives, which are the fundamental building blocks of each skill. The executive layer knows about this decomposition and ensures that primitives are executed in right order to guarantee a successful skill execution. This decomposition of the skills *moveBoxFromLevelToTray*, *moveBoxFromTrayToLevel* and *graspItem* is shown in Table 1. The composition of skill primitives for each skill is intrinsic knowledge of this layer. Further it monitors the outcome of each primitive. This layer has also the opportunity to perform recovery behaviors, if primitives fail. If no recovery can be performed or the recovery fails, this failure is reported to the planning layer. The decomposition of the *moveBoxFromLevelToTray* and its execution is shown in Figure 3.

skills	**moveBoxToRack**	**graspItem**	**moveBoxToLevel**
skill primitives	detectHandle graspHandle moveArmToSupportPose pullBox moveBox deliverBoxOnTray	detectItem graspItem deliverItem	detectHandle graspHandle moveArmToSupportPose pullBox moveBox deliverBoxOnLevel

Table 1: Within this table the skill primitive composition of all skills are listed.



5.4 Behavioral control

Figure 3: Decomposit *ToTray* into its primitives and how executive layer handles the executio

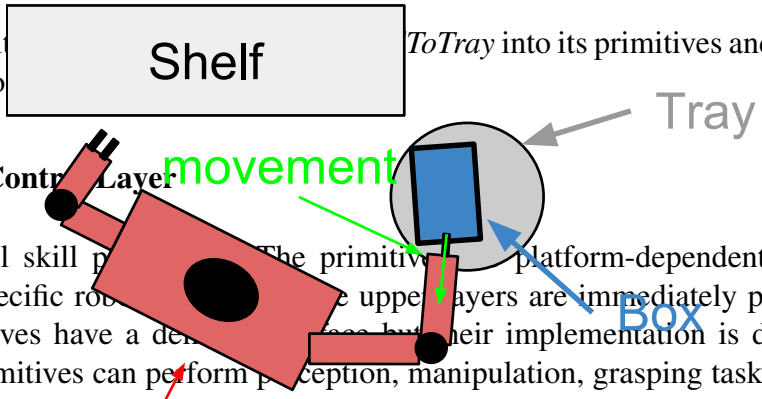


Figure 5.22: After the box is turned the box's handle is released and the grasping arm moves towards it to grasp the handle.

4.3. Behavioral Contr Layer

This layer holds all skill p The primitive platform-dependent and have to be re-programmed for specific ro e upper ayers are im ediately portable to other platforms. The primitives have a de eir implementation is different for different platforms. Skill primitives can perform p ception, manipulation, grasping tasks or any combination of those. Local recoveries are performed in this layer. If a recovery is impossible skill primitives report their error. A skill primitive can be used by different skills.
Figure 4 depict for instance the primitive of looking for items once the box is on the tray. For the box detection the point cloud of the top RGBD camera and the PCL implementation [16] of FPFH features [15] (initial detection) and ICP (fine alignment) are used. The control of the arms are realized using the MoveIt! framework [2]. Items are detected using the RGB cameras in the wrist. Figure 5a shows the inspection of a box while Figure 5a depict the internal representation of the situation.

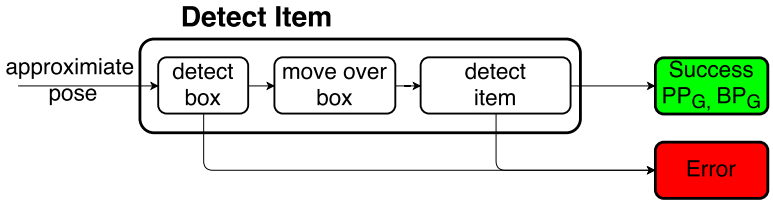


Figure 5.23: In this figure the different blocks of the detect item skill primitive are shown. First it detects the transport box, then it moves the arm over the box and detects the item using the hand camera. If the skill succeeds the *item pose* PP_G and the *box pose* BP_G ares returned.

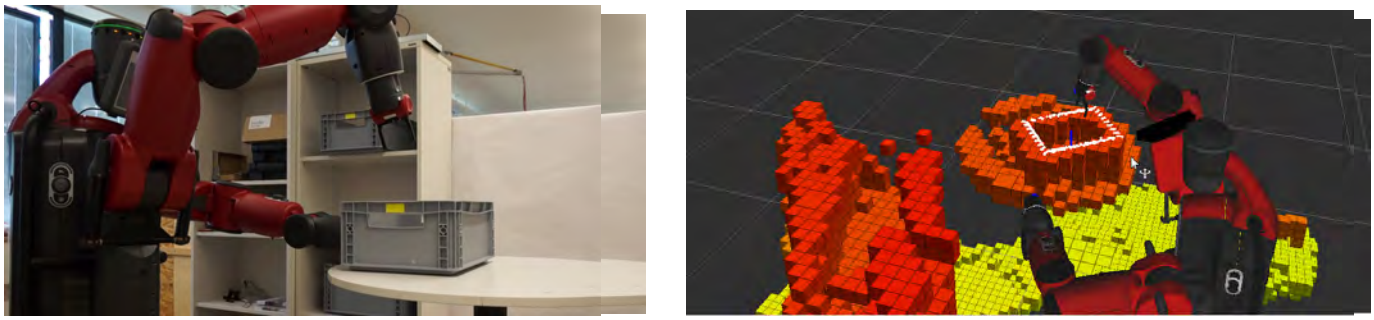
Figure 4: Detect Item primitive. PP_G represents the global item pose. BP_G represents the global box pose.

5. Results

The major result of this work is a working prototype implementation of the proposed order picking system based on the robot Baxter and standard software packages such as ROS or MoveIt!. But we are interested in particular in the dependability of the system. Therefore, we performed a detailed evaluation of the individual skill primitives.
For the evaluation we executed individual skill primitives multiple times (around 50 trials each) in

Figure 5.25: This figure shows the inspection points in more detail.

Figure 5.25: This figure shows the inspection points in more detail.



(a) Baxter inspecting the box with it's endeffector camera.

(b) Baxter inspecting the box visualized in RViz. The white point cloud indicates the detected box. The orange voxels visualize the collision scene.

(a) Baxter inspecting the box with it's endeffector camera.

(b) Baxter inspecting the box visualized in RViz. The white point cloud indicates the detected box. The orange voxels visualize the collision scene.

Figure 5: Realization of the primitive detectItem.

Figure 5.26: This figure shows Baxter inspecting the box in reality as well as visualized in RViz.

given setups. Details about the evaluations can be found in [10]. In Figure 6 the results of the individual skill primitives are shown. The green bar indicates the successful execution rate, the gray bar marks the failure executions which are detected by the system and the red bar shows the failures which are not detected by the system. Even if the success rate of some primitive is not overwhelming, the system detects the failure and reacts to it. The recognition of failures is one fundamental ability of reliable systems. As soon as the errors are detected, the robot can react to it autonomously.

72

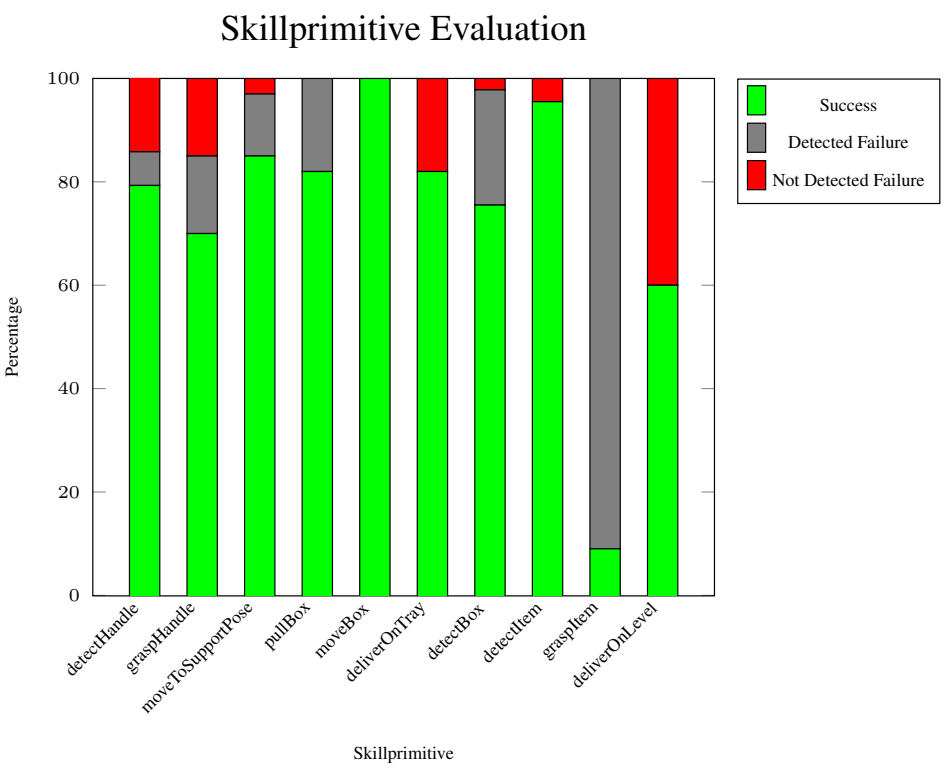


Figure 6: Skillprimitive evaluation

6. Conclusion

In this paper an autonomous order picking system was presented. In order to keep the system modular and portable a 3-TIER architecture was developed. The planning layer utilized an AI planner, which uses a PDDL description of the planning problem. The planner received the description of system skills, as well as start and goal state and provided a list of planned skills. Each skill is composed of skill primitives. These skill primitives are needed to address manipulation of the box, grasping items and perceptual tasks. The decomposition of skills into primitives enriched with monitoring and recovering capabilities contribute to the dependability of the system. The proposed system was implemented as a prototype using the robot Baxter and standard robotics software libraries.
Using this prototype implementation the concept of the skill primitives was evaluated. Although most primitives worked quite well, the evaluation pointed out some problems of this proof-of-concept system. Within most primitives, the major problem was that the execution of planned trajectories was aborted because Baxter was not able to execute them precisely enough. However these errors were detected by our system and reported to the high-level controller. For future work a more reliable execution of arm motions by Baxter needs to be addressed.

References

[1] S. Chitta, E.G. Jones, M. Ciocarlie, and K. Hsiao. Mobile manipulation in unstructured environments: Perception, planning, and execution. *IEEE Robotics & Automation Magazine*, 19(2):58–71, June 2012.

[2] Sachin Chitta, Ioan Sucan, and Steve Cousins. Moveit! *IEEE Robotics & Automation Magazine*, 1(19):18–19, 2012.

[3] Matei Ciocarlie, Kaijen Hsiao, Edward Gil Jones, Sachin Chitta, Radu Bogdan Rusu, and Ioan A Şucan. Towards reliable grasping and manipulation in household environments. In *Experimental Robotics*, pages 241–252. Springer, 2014.

[4] Malik Ghallab, Dana Nau, and Paolo Traverso. *Automated planning: theory & practice*. Elsevier, 2004.

[5] Chih-Wei Hsu, Benjamin W Wah, Ruoyun Huang, and Yixin Chen. Handling soft constraints and goals preferences in SGPlan. In *International Conference on Automated Planning and Scheduling (ICAPS)*, 2006.

[6] J. Huckaby, S. Vassos, and H.I. Christensen. Planning with a task modeling framework in manufacturing robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5787–5794, Nov 2013.

[7] D. Mcdermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins. PDDL - The Planning Domain Definition Language. Technical Report TR-98-003, Yale Center for Computational Vision and Control, 1998.

[8] Robin Murphy. *Introduction to AI robotics*. MIT press, 2000.

[9] Matthias Nieuwenhuisen, David Droeschel, Dirk Holz, Jorg Stuckler, Alexander Berner, Jun Li, Reinhard Klein, and Sven Behnke. Mobile bin picking with an anthropomorphic service robot. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2327–2334. IEEE, 2013.

[10] Julia Nitsch. Industrial Grasping. Master's thesis, Faculty for Computer Science and Biomedical Engineering, Graz University of Technology, Graz, Austria, 2016.

[11] K. Okada, Y. Kakiuchi, H. Azuma, H. Mikita, K. Murase, and M. Inaba. Task compiler : Transferring high-level task description to behavior state machine with failure recovery mechanism. In *IEEE International Conference on Robotic and Automation (ICRA)*, May 2013.

[12] Mikkel Rath Pedersen, Lazaros Nalpantidis, Aaron Bobick, and Volker Kruger. On the integration of hardware-abstracted robot skills for use in industrial scenarios. In *2nd International IROS Workshop on Cognitive Robotics Systems: Replicating Human Actions and Activities, (Tokyo, Japan)*, Nov 2013.

[13] M.R. Pedersen, D.L. Herzog, and V. Kruger. Intuitive skill-level programming of industrial handling tasks on a mobile manipulator. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4523–4530, Sept 2014.

[14] Francesco Rovida, Casper Schou, Jens Skovand Dimitris Chrysostomou Andersen, Rasmus Skovgaardand Damgaard, Simon Bøgh, Mikkel Rathand Bjarne Grossmann Pedersen, Ole Madsen, and Volker Kruger. Skiros: A four tiered architecture for task-level programming of industrial mobile manipulators. In *International Workshop on Intelligent Robot Assistants, (Padova, Italy)*, 2014.

[15] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3212–3217. IEEE, 2009.

[16] R.B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.

User-centered Assistive Robotics for Production
-
The AssistMe Project

Gerhard Ebenhofer[1†], Markus Ikeda[1†], Andreas Huber[2†], and Astrid Weiss[2†]

[1]Profactor GmbH.
gerhard.ebenhofer@profactor.at
markus.ikeda@profactor.at

[2]ACIN-Institute of Automation and Control
Vienna University of Technology
huber.cognition@yahoo.com
astrid.weiss@tuwien.ac.at

† These authors contributed equally to this work.

Abstract
In this paper we present the first results of the AssistMe project which aims at enabling close human-robot cooperation in production processes. AssistMe develops and evaluates different means of interaction for programming and using a robot-based assistive system through a multistage user-centered design process. Together with two industrial companies human-robot cooperation scenarios are evaluated in two entirely different application areas. One field of application is the assembly of automotive combustion engines while the other one treats the machining (polishing) of casting moulds. In this paper we will describe the overall project methodology, followed by a description of the use cases and a detailed outline of the first robotic prototype set up. The paper closes with an overview on the results of the first user trials that show very similar findings for both use cases and gives an outlook on the next expansion stage of the human-robot cooperation scenario.

1. Introduction
The idea that industrial robots need to leave their working cells and pre-programmed routine tasks in order to become more flexible in use and also more applicable for SMEs with smaller lot sizes and often changing production processes is nothing new. Robots, such as the collaborative robots from Universal Robots[1] and Baxter from Rethink Robotics[2] are entering the market with exactly that aim to offer robotic solutions for a closer human-robot collaboration, in which the strengths of the humans (e.g. problem solving, decision making) can get combined with the strengths of the robot (e.g. efficient fulfilment of reoccurring tasks) [1]. Companies such as KUKA start investing more and more in user-centered development (UCD) and usability standards such as ISO/TR16982:2002 were developed to support safe and close cooperation. Nevertheless, little user-oriented research has been performed so far outside the laboratory in the industrial context to understand what makes operators

[1] http://www.universal-robots.com/de/
[2] http://www.rethinkrobotics.com/

accept or reject robotic assistance (e.g. [2]). Similarly, little is known about best practices of user-centered development in the industrial context [3] [4].
Up to now, robot-based assistive systems are not widely spread in the manufacturing industry, as there is still research missing to uncover their full potential, and room for improvement in terms of usability, user experience, and subsequently user acceptance. Assigned purpose of the project AssistMe is the user-centered development and evaluation of innovative means of interaction for human-robot cooperation to improve usability and user experience of robot-based assistive systems in order to flexibly automatize selected production steps in an economically viable way.
The aim of the AssistMe project is to develop innovative haptic and optic concepts for human-robot cooperation in two different applications contexts, namely the assembly of automotive combustion engines while the other one treats the machining (polishing) of casting moulds. These concepts can be used during set up and interaction with a robot-based assistive system.
The project consists of three major development cycles. In a first iteration an assistive robot system, more or less out of the box is implemented for the use cases by application of process equipment. User studies regarding teaching and use of the systems are carried out. User-centred improvements in terms of usability and programmability are implemented as technical components in order to reduce programming complexity and programming duration as well as to improve system reliability and process quality.
Therefore different technology options are foreseen by the project frame. Force feedback technology will support programming and the usage of robot programs in order to make better use of robot articulated machining tools supporting the navigation through the real world by position-based haptic force feedback. Optic interaction technology, 2D and 3D sensors (and the corresponding machine vision algorithms) integrated with projection devices will render spatial augmented interaction e.g. textual feedback – instructions and explanations, during use. Apart from visualization, spatial augmented reality concepts with position and object-based projected information will be developed in order to be able to define virtual light barriers and projected buttons. Tools will be automatically positioned relative to objects (due to object pose recognition technology). In combination with haptic interaction technology interaction concepts will be evaluated that prevent users from the violation of 3D collision contours that have been captured and automatically interpreted as such by optical reconstruction technology beforehand. These interaction paradigms will be developed in a multi-stage process, together with operators in the two different testbeds and subsequently they will be evaluated in different expansion stages of the interaction technologies.

2. User-centered Design Approach

The two-years project is based on the concept of iteratively evaluating the same robotic assistance in different stages of expansions for the two different use cases. Stage 1 is an off-the-shelf robotic arm from Universal Robotics . Stage 2 will be further enhanced with a 3D sensor and Stage 3 with force feedback. Every stage of expansion will be evaluated together with representative target users from our industrial partners with respect to usability, user experience, and acceptance. After every evaluation implications for improvement for the next expansion stage will be derived to keep the operators' point of view in the development process. The AssistMe project thereby follows a very similar user-centered design approach as presented in [3] and the evaluation activities are methodologically grounded in the USUS evaluation framework [5]. The work presented in this paper are the general use cases for both application contexts, as well as the first expansion stage and its evaluation. Before we go on detail with our use case implementation, we will give a short overview on related state-of-the-art assistive robot systems.

3. Assistive Robotic System

Robot-based production nowadays is essential for industrial manufacturing. Due to safety reasons industrial robots are placed in a cell behind spatially separating safety equipment such as fences. As precise playback machines for movements, industrial robots remain insensitive towards their environment and repeat predefined sequences of actions. Industrial robots cannot react to changes in their environment and require reprogramming. [6] differentiate automatic and manual robot programming systems. In industrial scenarios robot specialists do reprogramming and reconfiguration partly with text based, controller integrated, teach pendant based (online) tools as well as with CAD-based graphical robot simulation tools. Results are, apart from some sensor signal inputs, more or less inflexible robot programs. Recently, a new class of industrial robots hit the market namely, [7] [8] [9] to mention a few, which can be potentially used in the same environment as human co-workers if relevant norms (A,B level norms that define Safety Integrity levels, performance levels, application specific C level standards) are fulfilled. [10] [11] define four modes of human-robot coexistence and collaboration as relevant for robotic applications. [12] specifies safety requirements for collaborative industrial robot systems and the work environment, and supplements the requirements and guidance on collaborative industrial robot operation. Programming of collaborative robot systems is equivalent to standard industrial robots since trained robot programmers are target on the one hand. On the other hand programming is simplified using macros to support unexperienced users. [9] provides the possibility of hand guidance during system teach in. This input modality is evaluated in the project, but gear friction renders exact hand guided teach-in difficult. Industrial installations of collaborative robots remain (until the integration of the project results) inflexible and unintelligent playback machines for movements and process technology such as intelligent cameras etc. It remains complicated and almost impossible with commercially available systems to integrate that renders in adaptive behavior. The AssistMe projects wants to enable naïve operators to manually teach a robotic arm for their purposes with little pre-knowledge requested. Afterwards a safe and user-friendly cooperation with the robot in the production process should be possible.

4. Use Cases

4.1 Assembly of automotive combustion engines (use case A)

The assembly of a combustion engine includes the installation of a cylinder head cover. The installation is carried out manually by stacking the cover with pre-inserted screws onto the motor block and tightening the screws with a manual power tool. The electronic screwdriver of the manual workplace is fitted with a push start mechanism, electronic control unit and a shut-off clutch and therefore starts rotating when pushed onto the screw and stops motion when retracted respectively when a predefined torque is reached. The working instruction of the workstation includes several additional process steps. An automatic screw tightening system is expected to provide assistance and to reduce the workload at the workstation for the human worker.
A state-of-the-art collaborative robot system [10] [11] is equipped with the power tool (Figure 1) and programmed to perform screw tightening operations in the required order and accuracy to meet a defined process quality (screw-in depth, torque,…). In the first expansion stage, the project evaluated the effectivity and simplicity of the user interface as implemented by the robot manufacturer and proposed modifications, which will inform the implementation of expansion Stage 2 and 3.

Figure 1- collaborative screwdriver robot system

Figure 2-manual polishing workplace

4.2 Machining (polishing) of continuous casting moulds (use case B)

Continuous casting of profile bars requires high precision moulds with excellent surface finish. Casting moulds are crafted from flat material by wire electro discharge machining that leaves eroded surfaces without the required surface finish quality. Manual polishing (Figure 2) by air pressure driven oscillating polishing machines is extraordinary labour intensive, unergonomic and harmful to health. Prolonged exposure to hand transmitted vibration from powered processes or tools is associated with an increased occurrence of symptoms and signs of disorders in the vascular, neurological and osteoarticular systems of the upper limbs [13]. Setup and programming time is crucial for the use case since continuous casting molds are usually one of a kind products, manufactured in lot size one, with polishing being by far the most labor intensive production step causing umpteen hours of labor per mold. Therefore an assistive system, easy to program and setup, is desirable that can reduce the amount of labor especially for ergonomic and health reasons.
A state-of-the-art collaborative robot system [10] [11] is equipped with a polishing tool and programmed to perform polishing operations. In the first expansion stage, the project evaluated the effectivity and simplicity of the user interface as implemented by the robot manufacturer and proposed modifications, which will inform the implementation of expansion stages 2 and 3.

5. Preliminary User Trials

The first expansion stage of the two use cases was evaluated in the first year of the project. In total three user trials were conducted in order to get feedback from the workers who actually used the new robotic system. Participants were recruited by our industrial partners and we explored the teaching of the robotic arm in Trial 1 and 2 and the actual collaboration with the robot in Trial 3 (see Table 1 for an overview).

All participants in Trial 1 and 2 successfully completed both teaching tasks (only one participant did not finish the second task due to time constraints). The gathered data showed that the touch panel in its off-the-shelf version was experienced as not feasible and too complex to control the robot during the teaching task for participants of both use cases. There was a strong tendency to omit the panel as an intermediate device and to try to directly control the robot using kinesthetic teaching. However, this type of control was also limited in feasibility as the robot arm was experienced as too bulky and

unprecise for teaching positions that way. Overall, the teaching of expansion stage 1 was rated as low with respect to usability, user experience, and acceptance, which can be explained by the fact that the actual teaching was only a fraction of the whole process, which was experienced as too complicated due to the touch panel. Trial 3 revealed that in the actual collaboration with the robot its working pace was perceived as not flexible enough, which bears the risk to re-establish a rigid production line logic. More details on the studies can be found in [14].

	User Trial 1 **Use case A**	**User Trial 2** **Use case B**	**User Trial 3** **Use case A**
Environment	Factory	Laboratory	Factory (assembly line)
Task	Teaching of screw positions	Teaching of polishing positions	Cooperative screwing
Duration	1 day	2 days	3 weeks
No. of Participants	5	5	5
Research Methods	Oberservation, Questionnaires	Oberservation, Questionnaires	Interviews

Table 1. Overview of the three user trials.

6. Inferred usability improvements

6.1 Technical project outlook: Expansion stage 2

Usability studies yielded requirements regarding robot hand guidance. Gear friction yields stacking and imprecise movement. Locking of certain degrees of freedom (e.g. rotation or translation,…) is asked for by users as well as semiautomatic tool alignment and expected to improve both programming time and process quality. A state of the art force torque sensor was integrated as well as buttons to call perpendicular realignment (Figure 4) or locking of rotational or translational degrees of freedom.

6.2 Technical project outlook: Expansion stage 3

Collaboration can be improved by adding visual feedback on the robot and the work piece during the teaching (to reduce the burden of switching attention between the robot and touch panel). [15] [16] introduce the notion Spatial Augmented Reality (SAR) and describe it as enhancement or aggregation of several Augmented Reality (AR) technologies. One formulation [17] might be a depth camera projector based system to project (correctly distorted) information on three dimensional objects instead of flat screens (Figure 3) and may be used for projection of buttons.

(Applied) robotics does not make use of SAR methods extensively. [18] introduces a projection-based safeguard system for robotic workspaces especially for collaboratively used workspace. [19] gives an overview on Tangible User Interfaces (TUI) which denote interfaces that can be manipulated physically, and which have an equivalent in the digital world and represent a mean for interactive control. The project proposes a combination of TUI and SAR methods. Hand guided positioning of the robot might be uncomfortable or time consuming due to inappropriate input modalities (friction afflicted robot drives, unintuitive touch screens,…).

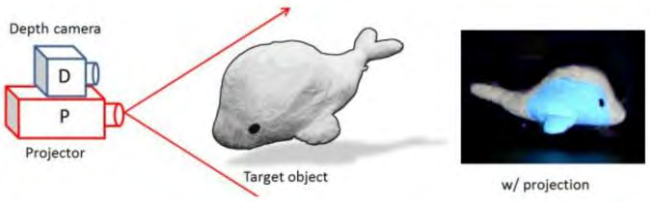


Figure 3- Depth Camera based tracking for Spatial Augmented Reality [17]

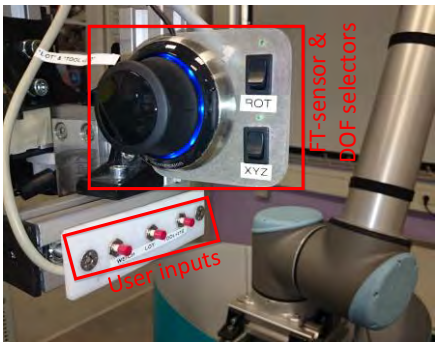


Figure 4 - projected buttons

Tightening order and poses of screws might be programmed by pointing to the screws with the finger [20]. Object and pose recognition introduces TUI to the world of AR. We propose a system that provides spatial detection functionality of teach-in devices (e.g. a spherical marble) that can be manipulated by the human. The system consists of one or more 3D sensors and a calibrated projector. Information on dynamic marble pose (resting marble may denote an underlying screw to be tightened) may be used for the programming of process points. A marble is placed on the cylinderhead cover. A projected interface element is pushed by the programmer who has to hold in order to avoid accidental acknowledgements. Once acknowledged, the 3D pose of the marble and thus the underlying screw is programmed to the system.

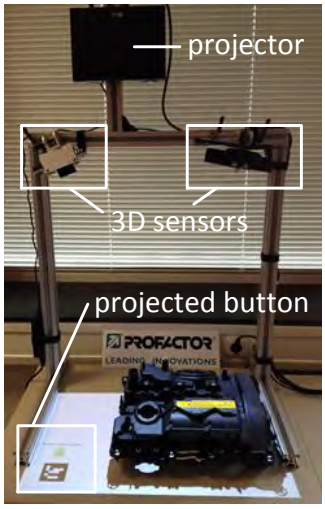


Figure 5 - system setup

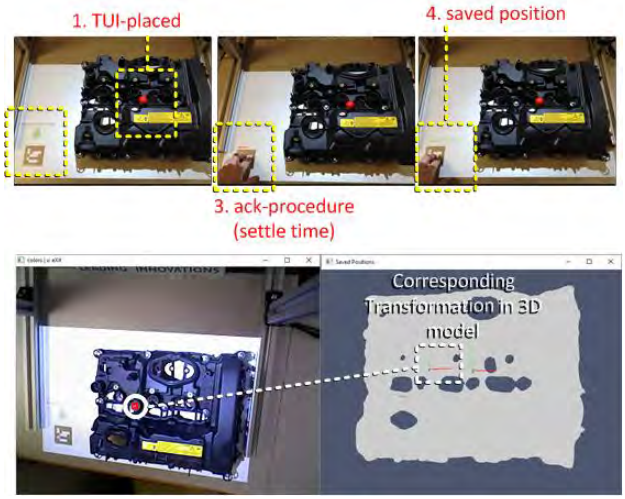


Figure 6 - SAR-TUI based process point programming

The system architecture (Figure 8) motivates advanced functionality in terms of robot hand guidance. Figure 7 shows first results of the environmental modelling system. An arbitrary placed object is recognized by the 3D camera and a virtual environmental model is updated (in realtime) with the model of the recognized object in its correct pose.

Force feedback algorithms are planned to render intuitive feedback for the user according to elements of the environmental model. E.g. boundary surface of volumes containing obstacles should make it impossible for the user to move the robots to or through such areas. Therefore negative force (as far as movement direction is concerned) has to be exerted to the hand guiding user. Positive forces may attract the user to process points contained by the model.

Virtual obstacles (attached to real world 2D – markers can therefore easily be integrated into the environmental model).

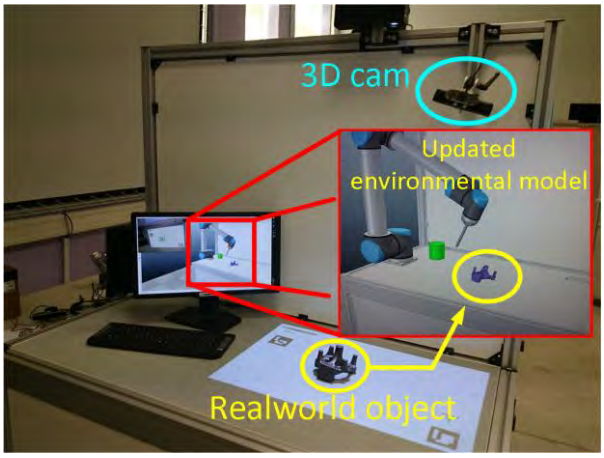


Figure 7- Augmentation of Virtual Model by Real World Objects

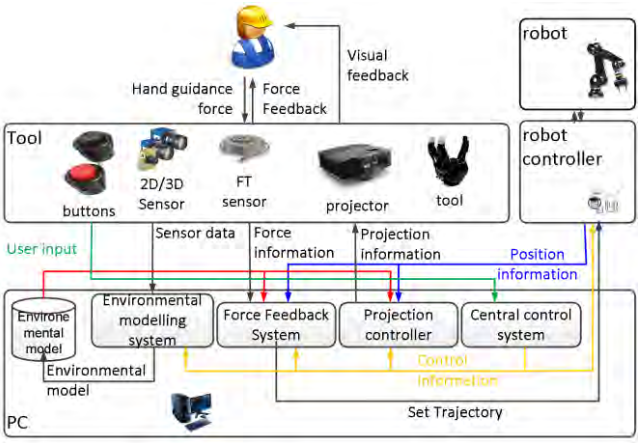


Figure 8 -system architecture

The force feedback system may not only render force feedback for the user during teach-in. It also can control process forces as e.g. required by the polishing process. Force-Torque information is acquired from the FT-sensor. An external sensor (Figure 9) is used instead of built in robot [9] functionality since force values estimated from required motor currents are too inaccurate due to e.g. gear friction. Process forces can be controlled e.g. for a touch up operation to exact 5N which is well below the detection threshold of the robot (Figure 10).

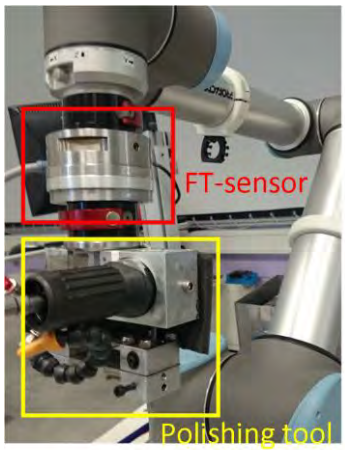


Figure 9- FT-sensor and polishing tool

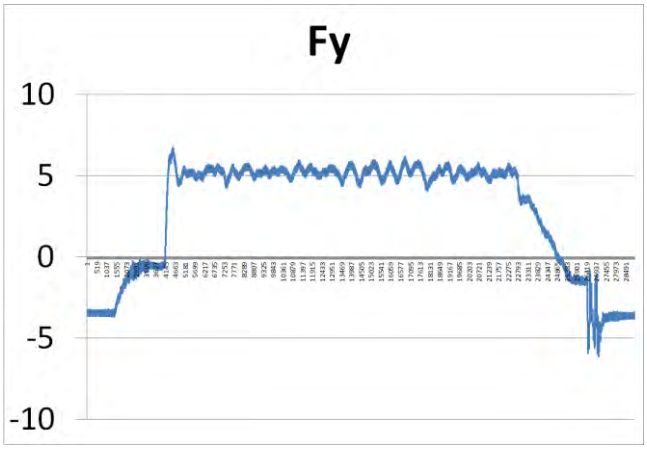


Figure 10- measured process forces

7. Conclusion

In this paper we presented the AssistMe project, which aims at enabling more flexible human-robot collaboration in the industrial context through a user-centred design approach. We outlined the overall approach of the project as well as its two use cases: (1) Assembly of automotive combustion engines and (2) Machining (polishing) of continuous casting moulds. We roughly described the main findings from the first end user evaluations which used a state-of-the-art robotic system. An outlook on expansion stages 2 and 3 were motivated and described.

8. Acknowledgements

This research has been funded by the FFG via the project AssistMe.

8. References

[1] A. Weiss, R. Buchner, M. Tscheligi und H. Fischer, „Exploring human-robot cooperation possibilities for semiconductor manufacturing," in *Collaboration Technologies and Systems (CTS), 2011 International Conference on*, 2011.

[2] D. Wurhofer, T. Meneweger, V. Fuchsberger und M. Tscheligi, „Deploying Robots in a Production Environment: A Study on Temporal Transitions of Workers' Experiences," in *Human-Computer Interaction--INTERACT 2015*, Springer, 2015, pp. 203-220.

[3] R. Buchner, N. Mirnig, A. Weiss und M. Tscheligi, „Evaluating in real life robotic environment: Bringing together research and practice," in *RO-MAN, 2012 IEEE*, 2012.

[4] S. Griffiths, L. Voss und F. Rohrbein, „Industry-Academia Collaborations in Robotics: Comparing Asia, Europe and North-America," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, 2014.

[5] A. Weiss, R. Bernhaupt und M. Tscheligi, „The USUS evaluation framework for user-centered HRI," *New Frontiers in Human--Robot Interaction,* Bd. 2, pp. 89-110, 2011.

[6] G. Biggs und B. MacDonald, „A survey of robot programming systems," in *Proceedings of the Australasian conference on robotics and automation*, 2003.

[7] *http://www.kuka-robotics.com/en/products/industrial_robots/sensitiv/lbr_iiwa_7_r800/start.htm.*

[8] *http://www.mrk-systeme.de/index.html.*

[9] *https://en.wikipedia.org/wiki/Universal_Robots.*

[10] *ISO 10218-1:2011 Robots and robotic devices -- Safety requirements for industrial robots -- Part 1: Robots.*

[11] *ISO 10218-2:2011 Robots and robotic devices -- Safety requirements for industrial robots -- Part 2: Robot systems and integration.*

[12] *ISO/TS 15066:2016 Robots and robotic devices -- Collaborative robots.*

[13] M. Bovenzi, „Health effects of mechanical vibration," *G Ital Med Lav Ergon,* Bd. 27, Nr. 1, pp. 58-64, 2005.

[14] A. Huber, A. Weiss, J. Minichberger und M. Ikeda, *First Application of Robot Teaching in an Existing Industry 4.0-Environment. Does it Really Work? Societies,* 2016.

[15] O. Bimber und R. Raskar, Spatial augmented reality: merging real and virtual worlds, CRC Press, 2005.

[16] R. Raskar, G. Welch und H. Fuchs, „Spatially augmented reality," in *First IEEE Workshop on Augmented Reality (IWAR'98)*, 1998.

[17] K. Tsuboi, Y. Oyamada, M. Sugimoto und H. Saito, „3D object surface tracking using partial shape templates trained from a depth camera for spatial augmented reality environments," in *Proceedings of the Fourteenth Australasian User Interface Conference-Volume 139*, 2013.

[18] C. Vogel, M. Poggendorf, C. Walter und N. Elkmann, „Towards safe physical human-robot collaboration: A projection-based safety system," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, 2011.

[19] H. Ishii, Tangible user interfaces, CRC Press, 2007.

[20] C. Harrison, H. Benko und A. D. Wilson, „OmniTouch: wearable multitouch interaction everywhere," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011.

Experiences with RGB-D based navigation in real home robotic trials*

P. de la Puente[1], M. Bajones[1], C. Reuther[2], D. Fischinger[1], D. Wolf[1], M. Vincze[1]

[1] ACIN Institute of Automation and Control.
Technical University of Vienna. Austria.

[2] MetraLabs Robotics GmbH Ilmenau. Germany.

Abstract
Autonomous robot navigation is an important and challenging component that is still missing in many real applications. In particular, home environments present open challenges that differ notably from one user apartment to another. Laser sensors cannot perceive objects at all heights commonly found in homes, we investigated the feasibility and suitability of using RGBD sensors for 2D autonomous navigation and a variety of tasks at real user homes. We use the concept of virtual laser scans to integrate RGBD data into mapping and localization methods. For realistic user interaction in actual homes we designed and improved ,over several pilot studies, the robot behavior for tasks such as approaching the user. In this paper, we report the adaptations needed to cope with home-specific challenges using RGBD sensors as a solution to perceive 3D environments.

1. Introduction

In many areas of application of service robots, mobility plays a role of great importance. Non-industrial real environments present increased complexity and in general are very hard to handle [18, 14].

In particular, autonomous navigation in user homes is a challenging aspect of care robotics projects. The SRS (MultiRole Shadow Robotic System for Independent Living) project pointed this out and focused on the development of remotely-controlled, semi-autonomous robotic solutions [11]. In other projects, such as Giraf and Giraff++, the robots were also externally teleoperated [2] and there was no autonomous navigation. This missing autonomous mobility has been identified as a key next aspect that needs to be solved. In the Companionable project, autonomous navigation to fixed predefined places was incorporated, but a controlled test home was used [17] instead of different real home environments.

RGB-D navigation poses additional difficulties [7, 3, 10]. The reduced field of view, the blind detection area and the small maximum range of this kind of sensors provide very limited information about the robot's sorroundings. Noisier points, spurious measurements and scale issues in the depth data also affect the perception capabilities.

This paper presents our developments, adopted solutions and identified issues to overcome the challenges of home environments using RGB-D sensors. We studied different navigation tasks and the adaptive approach to the user and conducted trials in several homes of older adults. The contributions

*This work was partially funded by the European Commission, project HOBBIT (FP7-288146).

to resolve the different tasks are described and specific problems of homes are addressed.

The paper is organized as follows. Section 2. describes the robot platform and sensor setup configuration. Section 3. presents our system architecture and implementation overview focusing on the navigation related tasks and components, which are explained in more detail in Section 4.. In Section 5., observed navigation problems are identified and addressed, while Section 6. provides an initial overview of the different navigation functions usage during the trials. Finally, Section 7. includes conclusions, final remarks and future challenges.

2. Robot platform and sensor setup

The PT2 mobile robot platform (prototype 2) was developed for the Hobbit project by partner Metralabs [12]. Two symmetric drive units and one supportig castor wheel constitute the low level locomotion system. A safety edge bumper protects the platform and blocks the motors while pressed, preventing the robot from moving while it is hitting an obstacle.

The sensor setup is based on two main RGB-D sensors, keeping a configuration similar to the one proposed for a previous prototype of the robot [5, 3]. The bottom camera -used for localization- is placed at a height of 35 cm. The head camera -used for obstacle avoidance, user detection, object and gestures detection and recognition- is mounted inside the robot's head, and can be tilted. 2D virtual lasers are created from each of the sensors, considering the largest measurements for localization with the bottom sensor (since they correspond to obstacles further away, like walls) and the closest measurements for obstacle avoidance with the top sensor. A height interval within the whole 3D point clouds is considered for the generation of the virtual scans.

3. System architecture and implementation

The whole system architecture has high modularity. To facilitate development, code reuse, communications management and integration, the popular Robot Operating System (ROS) [16] framework was used. Metralabs robots run the MIRA (Middleware for Robotic Applications) [4] framework, which manages low level control aspects of the platforms and also provides autonomous navigation functionalities and the Miracenter tool, which runs a complete instance of the framework with a graphical user interface.

In order to integrate MIRA into our ROS based system, several interfaces were required. The basic infrastructure of the new interfaces was based on existing interfaces from the STRANDS[1] project, modified and extended for our choices and needs. In our case, we decided to use MIRA navigation instead of ROS navigation because it was already well tuned for the current prototype platform and for reasons similar to the ones outlined by the Robot-ERA project team, such as enhanced support and robustness [9]. The required interfaces are implemented in different classes and run as a single ROS node.

In the first place, the virtual laser scans generated by ROS nodes had to be read, converted and adapted to be used by MIRA for localization and obstacle avoidance. In the other direction, an interface to provide the current localization pose as a ROS topic was required as well, and the corresponding trasformations are also computed and broadcasted.

[1]STRANDS project, EC 7th Framework Programme. Grant agreement num. 600623

Also, goal poses had to be sent from ROS nodes and SMACH to be processed by MIRA. This function was implemented as a ROS action server from which a MIRA navigation task including position, final orientation, and preferred driving direction subtasks -with their corresponding tolerances- is started. This way, a navigation task to a given goal can be preempted from the behavior side and the necessary feedback about the task status is provided by the actionlib server. The interface to start the action is the same as when sending a goal to ROS MoveBase action server, and the provided feedback is also translated to similar terms. Interfaces for discrete motion commands (distance to move, angle to turn) were also created to run in a separated ROS thread based on the distance mode experimental feature of MIRA.

Another action was created in order to start and interrupt docking on and off from the charging station. These procedures were implemented internally within MIRA.

Fig. 1 shows a simplified overview of the system architecture, including the most important modules and data flows with regard to navigation related tasks. More details about these tasks and methods are included in Section 4..

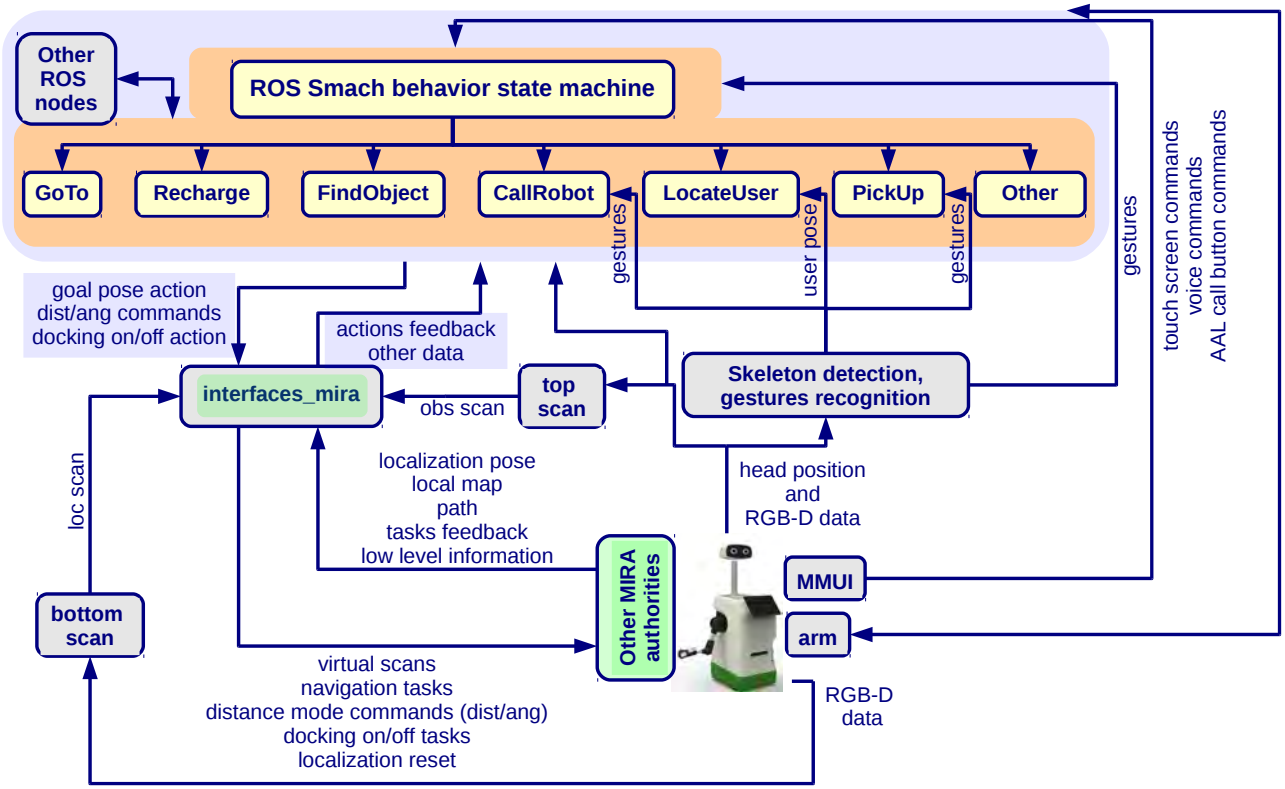


Figure 1. System components overview, focusing on navigation related tasks and modules (other details omitted for the shake of clarity). ROS Smach behavior state machines can interact with MIRA through the interfaces node directly or through other nodes.

The mapping process took place in advance, during the setup phase at each new trial site. For building new maps, we used the ROS implementation of Gmapping [8] and converted the generated maps into the MIRA corresponding format.

4. Navigation-related functions

Several functions desired by the users required navigation capabilities. This section describes these functions. More particular details of navigation between fixed places in real user homes are described in Section 5..

4.1. Go to place

This is the most basic navigation function. When started by the user, the robot should move from the current position to a given place, using predefinid positions and labels. The user can select the desired place/room name from a list displayed on the user interface or can use a voice command.

4.2. Recharge

During the setup phase, the charging station must be placed in a suitable place, which is not always easy to find in real apartments. Enough space for the station itself and its supporting plane that prevents it from moving is needed. There should also be enough space in front of it, so that the robot can detect the station with the bottom RGB-D sensor from a distance (the minimum recommended distance is 50 cm). Obstacles at the sides of the station can also reduce manouverability, increase the risk of false positives in the detection and result in a higher number of failures. Proper wall sockets and satisfactory conditions in the room are required, and visibility should be good, with no direct light coming from nearby windows. Last, but not least, it is very important that localization in front of the station is good so that the template is within the field of view (but the error distance to it is not so critical). Therefore, the station should not be placed along a featureless wall with few references in the orthogonal direction. Places in front of doorways are preferred over positions with a lower degree of geometrical variance in the alignment direction.

This task comprises several actions. In the first place, the robot has to reach a predefined position in front of the charging station. From this position, a docking action is started. The docking action starts the MIRA docking procedure, which first of all applies template matching between the bottom virtual scan and the station shape template (recorded from that point during the setup phase). If the template is found, template-based localization is activated and when the robot approaches the station obstacle avoidance is disabled. This procedure was specifically adapted by Metralabs to work with RGB-D sensors, since during the last part of the movement the robot is blind and open loop commands are applied then. When the docking movement is completed, the state machine checks whether the robot is indeed charging or not. If docking succeeded, the robot looks down and the user is notified, otherwise the robot should dock off and try again. If the template is not found, the action is considered aborted and then the state machine should apply a small rotation to one side and start the action again. If the station template is not detected again, then a rotation towards the other side should be applied and the action is started once more. If detecting the template is still not possible the whole docking task is aborted and the user is notified.

Once the robot detects that it starts recharging (either after autonomous docking or when manually placed into the station), localization is reset to the position of the station recorded and saved during the setup phase. This recovery mechanism was very useful both for testing and during the trials.

4.3. Find object

This function requires navigation to a set of predefined searching positions, usually located in front of tables, chests of drawers and other horizontal surfaces at intermediate height intervals. The searching positions were placed around 60-80 cm away from the surfaces border, trying to cover a large area of the horizontal plane with the head sensor. For this function, the path to all the searching positions was requested by setting a navigation task and muting the navigation until a path is received or a given timeout is reached. This way, it was possible to obtain the path from MIRA, which does not allow arbitrary path planning. If all paths are received, the searching positions are checked in increased length order, otherwise goals to which a path is not provided are visited first.

4.4. Call robot

At first, a fixed predefined place was associated to each call button id, so that the robot would come to that place when the call button was pressed. This simple method was not flexible enough, hence a novel approach was developed for this function.

It was convenient for the users to call the robot while sitting on their favourite chairs, armchairs, sofas, beds, etc. However, since the prototype platform was not able to rotate on the spot and the back side of the robot could collide with the furniture when turning around to drive away, the predefined positions could not be very close to the actual sitting place. It is also important to take into account that there can be different localization errors, plus the allowed error to decide whether the goal has been reached, so the real position is not always exactly the same (the uncertainty accumulates). A compromise was often needed so that the user could reach the touch screen while allowing the robot to detect the obstacle and drive away safely. Furthermore, a more significant limitation was the fact that chairs and armchairs usually can move and the sitting position of the user along a sofa or a bed will most likely vary from time to time. So the desired final position should not be fixed in an improved method.

The new approach we present incorporates user detection and interaction, remembered obstacles and discrete motion commands for coming closer to the user with better, adapted positioning. Discrete motion commands towards the detected user were chosen over a planned path for three main reasons: 1) the movement is more direct and predictable; 2) with our sensor setup, obstacle avoidance while following the path would require the head to look down whereas user detection and a nicer interaction require the head to look straight forward; 3) existing navigation frameworks, to our knowledge, are not directly suitable for planning a path to a goal out of the fov or blocked and inside an occupied area, so defining a proper goal pose in free space would be challenging, depending on the environment characteristics and requiring higher level knowledge about the specific furniture (orientation, feasible approach directions and so forth). Discrete motion commands based solely on the user detection with no obstacle avoidance whatsoever, on the other hand, can be risky.

Our solution works as follows. Predefined positions are defined further away from the sitting zone so that the user's skeleton should be inside or close to the head sensor fov. The maximum distance limit is given by the defined range of the virtual scan for obstacle avoidance or by the maximum distance for a discrete motion command. When a predefined position is reached, the navigation parameters are changed so that occupied areas in the current local map are not degraded or overwritten by new sensor data (user navigation mode). Then the head can move up for user detection[15], and ray-tracing on the local map is performed. The local map's origin is located at the odometry reference system and the

origin for ray-tracing is defined at the robot hull front, based on the robot's corrected pose. The local map is used instead of the current measurements alone because it also contains obstacles remembered from before, even if they are already inside the sensor's blind zone. The robot turns to face the user and moves towards the sitting position up to a distance given by the closest projected measurement within an angle around the detected user, considering a safety margin. Fig. 2 illustrates this.

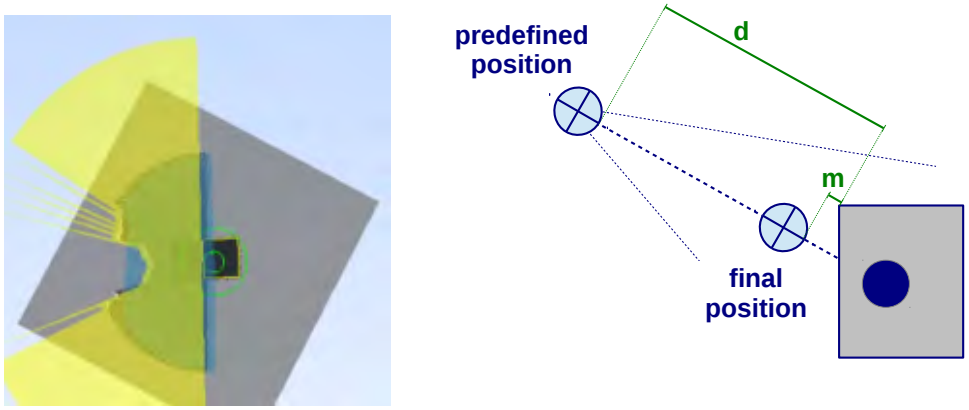


Figure 2. Obtaining the distance to move towards the user. Left: ray tracing on the local map. The current virtual scan when the head is looking up is depicted in blue, while the scan obtained from ray tracing is shown in yellow. The ray tracing scan is considered to be obtained from the frontal part of the robot (excluding the bumper, which is taken into account separatedly). The obstacle in front of the robot is not detected by the head sensor but is present in the remembered local map and therefore in the scan obtained by ray tracing. Right: simplified diagram of how the distance to move is computed. The minimum distance from ray tracing is denoted d and m is the safety margin.

For enhanced flexibility, reliability and better adaptation to different users, after this process is executed the robot asks whether to come even closer. When there is a positive answer the robot moves 15 cm more, completely blind now and trusting the user's input, and then the question and subsequent movement can be repeated up to three times. The user can reply by means of voice, gestures, or the touch screen commands. The robot remembers if it moved closer and if so it moves backwards again before starting a new navigation task.

The described method for approaching the user was developed for the call robot function, but it can also be applied in other tasks and contexts. For instance, we also incorporated it at the end of a fitness function scenario so that the robot comes closer to the user for further interaction after exercising.

4.5. Pick up object

For this function, in the first place, navigation to a goal obtained from the user's pointing gesture is needed. Using autonomous navigation in this way provides a much wider applicability and can be useful for picking up objects not directly inside the robot's field of view. The static map of the environment is also used for checking whether a detected object is too close to walls or funiture, since that means risk of collision and picking up the object is not possible then. Once the precomputed pose is reached, fine positioning with respect to the detected object is performed, based on discrete motion commands with sufficient accuracy. More details about the pick up algorithms can be found in [6].

4.6. Locate user

This function requires navigation to a set of predefined searching positions which in this case were usually located in the middle of the rooms or were the same as the call button places. Since detecting the user while the robot is rotating is hard, several shorter rotations were performed, stopping to call and detect the user in between. Using discrete motion rotations with our settings resulted in a more abrupt movement that affected localization, so a navigation task including only an orientation subtask was preferred. Still, the orientation estimate was sometimes not so good after the short rotations and the uncertainty associated to the orientation measurement had to be adjusted. Also, it was usually better to define this kind of positions in places where distinctive references were present and in relatively open space so that localization could get better before reaching narrower areas that require better accuracy.

5. RGB-D based navigation in home environments

One of the first things to check with the proposed sensor setup, depending on the environment, is that the height interval to be considered for the bottom sensor virtual scan generation may need to be changed. In general, we used a fixed width of 8 cm around the horizontal plane at the sensor's height. In the presence of low sofas and similar furniture, however, this interval had to be lowered since otherwise the virtual scan generated with the largest measurements included irregular surfaces such as cushions, etc. Fig. 3 illustrates this kind of problem.

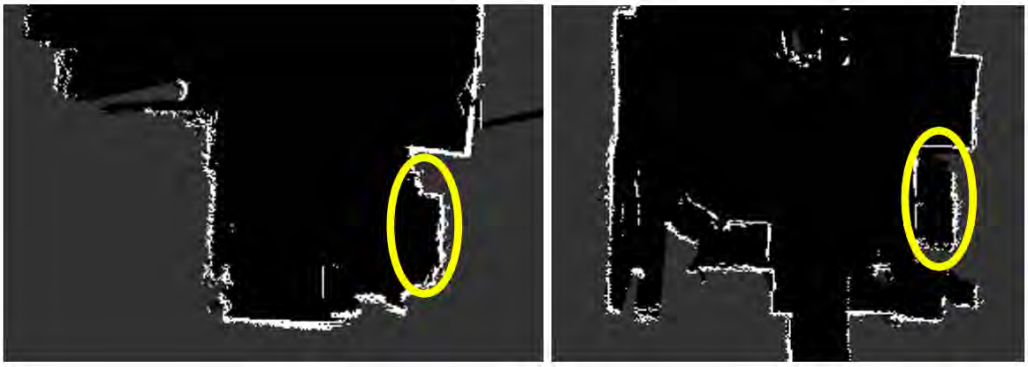


Figure 3. If the height interval considered for the bottom virtual scan is too high in the presence of low sofas, irregular borders and cushions will be present in the generated scan.

Regarding complete map building, the main limiting factors are the small field of view and the range properties of the RGB-D sensors. Since home environments may have narrow areas and excessive rotations during the mapping process can lead to less accurate results and even cause distortion, some walls may be missing in the resulting maps. This fact, together with the possibility that current sensor data overwrite the global map, can bring about problems in global path planning (see Fig. 4 for an example).

One first thing that can be done to prevent this problem when the robot is not facing the obstacle is carefully edit the map manually, to include the complete walls without risking the map building process overall result. This, however, does not help when the current sensor data are used to overwrite and clear the map. In that case, with this feature included within the MIRA based implementation, one option is to add MIRA `nogo` areas in order to avoid undesired plans when the robot is too close to a wall. The problem is that in very narrow areas some margin must be allowed for possible localization

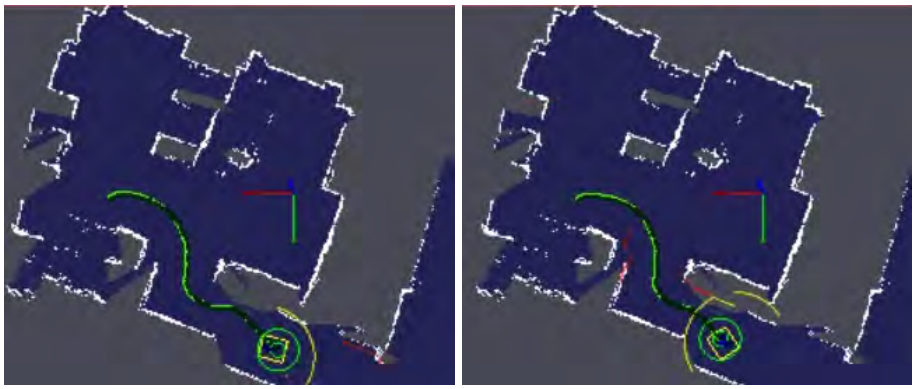


Figure 4. Global path planning in a narrow corridor in a home-like lab. One wall is missing in the map and is not always observed by the obstacles virtual laser (yellow). The initial path may be too close to the wall (left) and only be corrected if the robot gets to face the wall and is not already too close (right).

errors, so a compromise is needed. Fig 5 shows a couple of examples from real user homes. The environment on the left presents more critical localization conditions when entering the corridors that look horizontal in the image, since they are reached after traversing a long featureless corridor. Along this corridor, however, lateral localization is more accurate because the robot should not accumulate so much uncertainty before accessing it from these lateral, horizontal, shorter corridors.

Figure 5. MIRA nogo areas can be added in order to avoid undesired global paths when the static map is cleared. In very narrow corridors, however, some margin must be left to account for posible localization errors.

Today's laser based localization methods assume that features of the environment can be detected. This is not true in long corridors where the robot could be anywhere along the parallel walls unless an end is observed, which is sometimes not feasible if the corridor is long or when the corridor ends in a room and hence a further opening. This problem can be avoided by using laser sensors, with a significantly longer range. Another option is to combine geometric methods with visual methods [13], but if the walls are uniform the same problem remains. A possible addition of a few external references in difficult areas such as corridors, very wide or narrow spaces, and ambiguous places could also help in the mapping and subsequent localization processes. In the previous case (Fig 5, left), we

ended up adding a small extra piece of furniture in the long corridor as an additional reference. Visual markers with a fixed pose could help completely correct localization in those areas.

When entering a room, it is important that the robot is correctly localized in the transversal direction to the doorway and that the doorway is approached from the front, so doors located at one side of a corridor may cause problems, while doors located at the beginning or end of a corridor are better. In order to approach doors from the front, avoiding getting too close to the corner sides, a useful strategy to try in wide enough places is adding `nogo` areas at sides of a doorway entrance or at sharp corners (Fig. 6). This way, it is possible to have safer navigation behaviour in wide areas while keeping the capability to go through narrower areas. This provides more flexibility than methods with fixed security margins for the whole operational area.

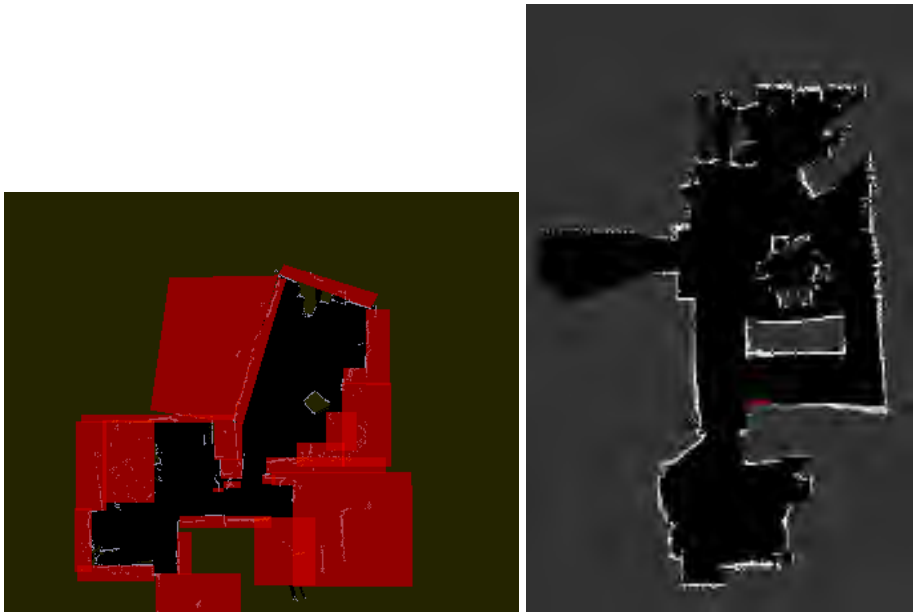


Figure 6. MIRA `nogo` areas can be used to avoid getting too close to corners and door sides. Note that good enough localization in the transversal direction is still very important.

`Nogo` areas were also useful to avoid difficult and not allowed areas and rooms in the environments. A few examples are shown in Fig 7. Areas with cables and thin obstacles on the floor and very narrow rooms (usually kitchens) where this non-omnidirectional robot could not manouver were also avoided. It is worth noting, of course, that `Nogo` areas are only useful if localization is good enough.

Other challenging situations were caused by thresholds, bumps on the floor and carpets. In the case of thresholds, we tested commercial and home made ramps when possible (Fig. 8). After testing different configurations and finding proper steep limits, the robot was usually able to pass thresholds. In a few cases problems were observed with standard planning methods if a new plan caused the robot to turn while driving on a ramp. A direct behaviour control may actually improve over plan-based approaches in narrow passages and in these particular cases, but it needs separate implementation and specific, situation-dependent triggering. Regarding localization, it would be useful to switch between 2D and 3D localization methods, but again the complexity would be increased.

Another important thing to take into account is that dealing with the dynamic nature of real environments is an open research issue to which new projects are dedicating big efforts [19, 1]. Our experience with the previously described approaches showed that minor changes in the position of

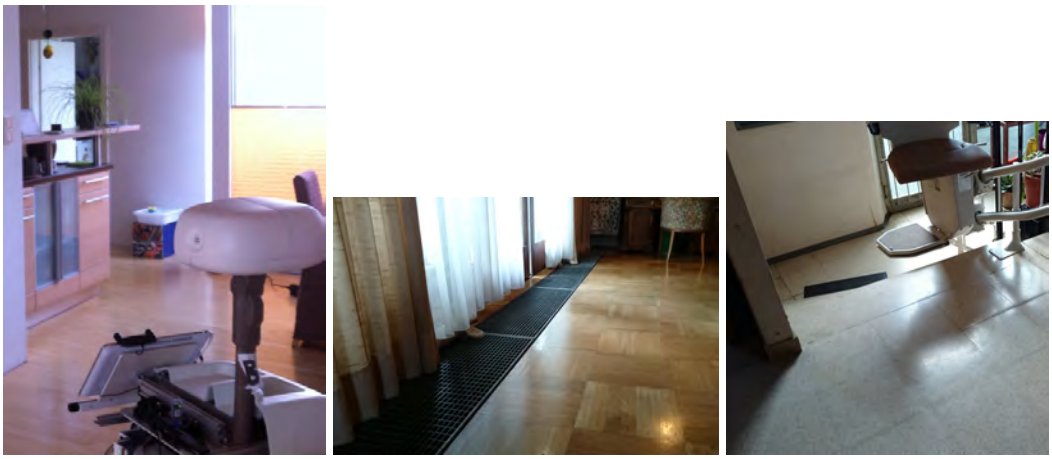


Figure 7. Examples of difficult areas in the environment to be avoided. Left: high outer shelves canot be observed with this sensor setup. Middle: the robot should not go through the uneven ventilation area close to the window. Right: areas with stairs are particluarly dangerous and should not be allowed.

Figure 8. Using ramps to go through thresholds.

chairs and small items did not usually have a significant influence on the localization quality, as long as discrepancies with respect to the static map were limited. Removing or changing furniture that provided relevant geometrical references present in the map, however, led to serious differences, especially in narrow areas and corners with little space to recover while moving. Detecting when the robot is actually lost and trying to apply autonomous recovery methods are very challenging problems that require specific research.

6. Functions usage

Table 1 provides results on how often different functions were used in the user trials conducted in Vienna. Preliminary results show that most of the users evaluated these functions overall as "Good", but there were also important failures and negative comments. Results in the lab were significantly better and we think that likely reasons are related to the following facts: there was more time for setup, preparation and improvements; the environment is easier and better known; we know the system better than the real users; robot transportation and long term operation can degrade platform performance and calibrations, etc.

Table 1. FUNCTIONS USAGE DURING THE TRIALS

User	`GoTo`		`CallRobot`		`Recharge`
	Used	Cancelled	Used	Cancelled	Used
V1	89	33	340	190	49
V2	29	8	172	81	186
V3	18	5	74	21	126
V4	46	26	97	75	16
V5	117	24	71	42	88
V6	386	85	146	60	312
V7	41	17	349	263	168

7. Conclusions and future work

This paper presented a summary of contributions and findings for navigation tasks of care robots operating in real home environments, using an RGB-D based sensor setup. Increased flexibility and adaptability over existing solutions were provided for the tasks execution, also addressing different limitations of the sensors used.

Our experience indicates that autonomous RGB-D navigation in real home environments is usually feasible and is much appreciated but presents drawbacks, open problems and further challenges. Directions for future work were already highlighted in the paper.

References

[1] R. Ambrus, N. Bore, J. Folkesson, and P. Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[2] S. Coradeschi, A. Cesta, G. Cortellessa, L. Coraci, C. Galindo, J. Gonzalez, L. Karlsson, A. Forsberg, S. Frennert, F. Furfari, A. Loutfi, A. Orlandini, F. Palumbo, F. Pecora, S. von Rump, A. Stimec, J. Ullberg, and B. ¿Otslund. Giraffplus: A system for monitoring activities and physiological parameters and promoting social interaction for elderly. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*, volume 300, pages 261–271. Springer International Publishing, 2014.

[3] P. de la Puente, M. Bajones, P. Einramhof, D. Wolf, D Fischinger, and M Vincze. RGB-D sensor setup for multiple tasks of home robots and experimental results. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[4] E. Einhorn, T. Langner, R. Stricker, C. Martin, and H. Gross. Mira - middleware for robotic applications. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[5] D. Fischinger, P. Einramhof, W. Wohlkinger, K. Papoutsakis, P. Mayer, P. Panek, T. Koertner, S. Hoffmann, A. Argyros, M. Vincze, A. Weiss, and C. Gisinger. Hobbit - the mutual care robot. In *Workshop on Assistance and Service Robotics in a Human Environment at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[6] D. Fischinger, A. Weiss, and M. Vincze. Learning grasps with topographic features. 2015.

[7] J. González-Jiménez, J. Ruiz-Sarmiento, and C. Galindo. Improving 2D reactive navigators with Kinect. In *10th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2013.

[8] G. Grisetti, C. Stachniss, and W. Burgard. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Transactions on Robotics*, 23:2007, 2007.

[9] N. Hendrich, H. Bistry, and Jianwei Z. PEIS, MIRA, and ROS: Three frameworks, one service robot -A tale of integration. In *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2014.

[10] M. Jalobeanu, G. Shirakyan, G. Parent, H. Kikkeri, B. Peasley, and A. Feniello. Reliable kinect-based navigation in large indoor environments. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[11] M. Mast, M. Burmester, E. Berner, D. Facal, L. Pigini, and L. Blasi. Semi-autonomous tele-operated learning in-home service robots for elderly care : A qualitative study on needs and perceptions of elderly people, family caregivers, and professional caregivers. In *Proc. of the 20th International Conference on Robotics and Mechatronics*, 2010.

[12] Metralabs.

[13] P. Panteleris and A. Argyros. Vision-based SLAM and moving objects tracking for the perceptual support of a smart walker platform. In *Computer Vision - ECCV 2014 Workshops*, volume 8927 of *Lecture Notes in Computer Science*, pages 407–423. Springer International Publishing, 2015.

[14] E. Prassler, R. Bischoff, W. Burgard, R. Haschke, M. Hñgele, G. Lawitzky, B. Nebel, P. PlÂger, U. Reiser, and M. ZÂllner, editors. *Towards Service Robots for Everyday Environments. Recent*

Advances in Designing Service Robots for Complex Tasks in Everyday Environments. Springer Tracts in Advanced Robotics. Springer Berlin Heidelberg, 2012.

[15] A. Qammaz, N. Kyriazis, and A. A. Argyros. Boosting the performance of model-based 3d tracking by employing low level motion cues. In *Proc. of the British Machine Vision Conference (BMVC)*. BMVA Press, 2015.

[16] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, 2009.

[17] C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, C. Huijnen, H. van den Heuvel, A. van Berlo, A. Bley, and H.-M. Gross. Realization and user evaluation of a companion robot for people with mild cognitive impairments. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pages 1153–1159, 2013.

[18] A.J. Soroka, Renxi Qiu, A. Noyvirt, and Ze Ji. Challenges for service robots operating in non-industrial environments. In *Proc. of IEEE International Conference on Industrial Informatics (INDIN)*, 2012.

[19] G.D Tipaldi, D. Meyer-Delius, and W. Burgard. Lifelong localization in changing environments. *International Journal of Robotics Research*, 32(14):1662–1678, 2013.

WS 7: Poster OAGM & ARW

Localization of an Automated Guided Vehicle (AGV) by Stereo Based Visual Odometry and Artificial Landmark Detection

Daniel Klingersberger and Gerald Zauner

FH OÖ Forschungs & Entwicklungs GmbH, Stelzhamerstr. 23, 4600 Wels
gerald.zauner@fh-wels.at

Abstract
Localization in a known environment is an essential topic in the field of robotics – consequently a variety of methods (e.g. Visual Odometry or SLAM) are scientifically well established. Compared to experimental robotics, the determination of position based on Machine Vision approaches is not yet fully implemented in the domain of Automated Guided Vehicles (AGV). Thus, the aim of this master's thesis is to design, realize and test a localization system exclusively based on Machine Vision for the use in AGVs. The x- and y-axis positioning as well as the determination of orientation of the vehicle in all three axes is carried out by a stereo camera based Visual Odometry approach and a supporting detection of artificial landmarks placed on the ceiling. Both methods complement each other perfectly: while Visual Odometry bridges distances without landmarks, drift caused by Visual Odometry is corrected by artificial landmarks. Test series have shown that the localization error falls below ±20mm if the distance between camera and landmark does not exceed 4500mm. Also the inclination of the vehicle is equalized. This localization system has various advantages compared to well established methods: designing and installation efforts can be reduced, while the flexibility for route changes can be increased compared to traditional magnetic guidance systems. The interference immunity is higher compared to contour matching methods due to the use of absolute reference points placed on the ceiling. The proposed system is not suitable for use in halls because the distance between camera and ceiling-landmarks should not exceed 4500mm. Nevertheless, this localization system is an interesting alternative to well established methods primarily for the use in the public sector, e.g. hospitals or libraries.

A Holonomic Robot for Rescue Applications

R. Edlinger[1], M. Zauner[2], W. Rokitansky[2]

[1] FH OÖ Forschungs & Entwicklungs GmbH, A-4600 Wels, Stelzhamerstraße 23
[2] FH OÖ Studienbetriebs GmbH, A-4600 Wels, Stelzhamerstraße 23
{raimund.edlinger, michael.zauner, walter.rokitansky}@fh-wels.at

Abstract
For autonomous mobile robots it is important to have the capability to plan and reach a defined goal. In this article, we present a novel mobile robot for urban search and rescue, capable of achieving a high level of locomotion. The preliminary aim is to build rescue robots which are able to drive in an unstructured environment and search for victims. Mobile robots have been an essential element in search and rescue scenarios and especially in space exploration to perform science on lunar and planetary objects. With advancements in research and technology many mobile robots have been developed with different configurations, geometries, sizes and flexibility of locomotion. These systems share different performance qualities under certain operational conditions. A new mechanism is developed to drive sideways which could be helpful especially in difficult curved staircase or uneven terrain.

Low Cost Remote Control for SAR Applications

A. Pointinger[1], B. Fuchs[1], R. Edlinger[2], M. Zauner[1], W. Rokitansky[1]

[1] FH OÖ Studienbetriebs GmbH, A-4600 Wels, Stelzhamerstraße 23
[2] FH OÖ Forschungs & Entwicklungs GmbH, A-4600 Wels, Stelzhamerstraße 23
{armin.pointinger, bernd.fuchs2}@students.fh-wels.at
{raimund.edlinger, michael.zauner, walter.rokitansky}@fh-wels.at

Abstract

This poster presents a low-cost remote control for SAR applications. The use of multi robot systems makes it difficult to control all robots from one operator base. The case described herein with a maximum weight of <10 kg is easy to handle and transport and fits the requirements for cabin baggage by airlines. To save space current low-cost embedded systems are used which are very energy efficient and provide a long operation time. In order to build a flexible and modular system, the communication and energy supply are able to work with different sources. The communication between robot and operator base is possible with a LAN cable or via wireless LAN. The energy can be delivered by a battery or an external energy grid. The batteries have enough power to run the operator station for 2 hours and enables rescue operations to be fulfilled under the harshest conditions. Because of the "Spacemouse" and the ergonomic control elements used, the unit is user-friendly and can be operated with gloves and in dark environments. The elements are clearly structured and make using the robots more intuitive. The control elements are focused in three groups: the engine, the arm control and special functions. All components and joints are sealed with rubber seals. Therefore, rain or dusts in harsh environment guess no problem for the remote control unit. Experts from first responder organizations will test the control in the coming years and contribute their experience to its further development.

New Algorithm to Speed up the Computation of a Visibility

M. Zauner[1], R. Edlinger[2], W. Rokitansky[1]

[1] FH OÖ Studienbetriebs GmbH, A-4600 Wels, Stelzhamerstraße 23
[2] FH OÖ Forschungs & Entwicklungs GmbH, A-4600 Wels, Stelzhamerstraße 23
{michael.zauner, raimund.edlinger, walter.rokitansky}@fh-wels.at

Abstract

This poster describes a new algorithm called B# which is needed to find the visibility graph of a polygonal region with obstacles defined by simple polygons. It focuses on finding the entire visibility graph among polygonal obstacles which has been tuned in a variety of test cases. The obstacles are restricted to simple cases, i.e. where no edge intersects any other edges. The visibility graph problem itself has long been studied and has been applied to a variety of areas. A common use for it has been for finding the shortest path. The B# algorithm has been implemented adjustments made and experimental comparisons via time measurements carried out. A comparison between "Naïve algorithm" and "Naive algorithm with B#" was performed with different numbers of vertices. The B# algorithm by itself doesn't calculate the visibility graph. It selects the next best obstacle where the calculation should proceed.

Ridge Point Extraction with Non-Maximum Suppression on Irregular Grids

Richard Schönpflug and Hubert Mara
r.schoenpflug@stud.uni-heidelberg.de
hubert.mara@iwr.uni-heidelberg.de

Ruprecht-Karls-Universität Heidelberg
IWR – Interdisciplinary Center for Scientific Computing
FCGL – Forensic Computational Geometry Laboratory
Klaus-Tschira-Platz, 69120 Heidelberg, Germany

Abstract
Assyriology is the study of cultures related to cuneiform writing, which was used for more than three millennia before Christ in the ancient Middle East. Drawing hundreds of thousands of documents with cuneiform script manually is a tedious task and leads to a demand for automated tools assisting the daily work of assyriologists. The cuneiform script is a handwriting using wedges (Latin: cunei) imprinted into clay tablets. Therefore the digitization of cuneiform tablets is increasingly using 3D-scanners that provide irregular triangular grids in $\mathbb{R}^3$*. These grids i.e. meshes are discrete manifolds, which are first filtered by using* Multi-Scale Integral Invariants *(MSIIs) for visualization. Secondly the MSII filter results are used to extract points along the or ridges within the 3D-model leading to a digital drawing of e.g. a cuneiform tablet. Therefore we choose the idea of the* non-maximum suppression *as used by the Canny edge detector for raster images. In contrast to the Canny edge detector we had to (i) to adapt to an arbitrary number of neighboring vertices, which have to be reduced locally in case of flat areas; (ii) to implement an estimator for the gradient direction, which cannot be provided by the MSII filter; and (iii) to provide a border treatment as real world meshes have missing parts. All the work was embedded within our modular* GigaMesh *software framework. Results are shown for synthetic and real data, demonstrating a computational complexity of* $O(n)$*, which requires only one parameter. Finally a summary and an outlook are given.*

1. Introduction

Cuneiform script was used for more than three millennia before Christ and is one of the oldest known writing systems. It is a handwriting in 3D, where imprints were made into clay tablets, using a reed styli [13]. This results in groups of wedge shaped imprints forming the characters. The name *cuneiform*, originates from the word *cuneus* for wedge. Drawing a replication of the cuneiform tablets is an integral part of their decipherment. This drawing step is traditionally done by manually tracing photographs of the tablets and can take hours or even days. This is an almost impossible task taking into account the hundreds of thousands of unpublished tablets. These tablets are important for many other disciplines as they provide insights into a wide variety of topics ranging from the economics of ancient societies to the first great works of literature, e.g. the epic of *Gilgamesh* [10].

This work is motivated by the task to extract cuneiform characters and other imprinted features out of 3D-models of tablets. The models are acquired using optical scanners based on the principle of structured light [12]. Having a robust filter using *Multi-Scale Integral Invariants* (MSIIs) [7] we extended, the filtering using the principle of the non-maximum suppression as known from the *Canny edge detector* [2]. Therefore we had to extend the algorithm for an arbitrary number of neighboring vertices as there is no fixed number of neighboring pixels/vertices. As MSII filtering does not provide a gradient direction we had to add an estimator using the normals of the triangles (faces) connecting the vertices. To improve robustness we apply a local mesh simplification for flat areas. These processing steps are described within the next sections and are embedded within our modular *GigaMesh* software framework [8, 9], which provides the MSII – and other filter results – as precomputed function values $f(\cdot)$ for irregular meshes. This work is used for further processing to gain high-level knowledge of cuneiform tablets as known from the domain of *Handwriting Text Recognition* (HTR) [1].

2. Ridge Tracing on Irregular Grids

The acquired 3D-models consist of meshes described by lists of vertices $\mathbf{p_i} = (x_i, y_i, z_i)^T$ and faces (triangles) $\mathbf{t}_i := \{\mathbf{p}_{A_i}, \mathbf{p}_{B_i}, \mathbf{p}_{C_i}\}$ having an orientation. The mesh is a discrete two-dimensional manifold $\mathcal{M}_2$ in $\mathbb{R}_3$ having orientated edges $\{\mathbf{e}_{a_i}, \mathbf{e}_{b_i}, \mathbf{e}_{c_i}\}$, which are implicitly given by the oriented faces [7]. The orientated faces allow to determine the space enclosed by the mesh. The index i is used to address all the elements of the mesh processed consecutively, while j addresses all elements next to the element with index i. Note that computational expensive calculations – especially the MSII filter – are parallelized within *GigaMesh*. The vertices of the 1-ring neighborhood are denoted as $\mathbf{p}_j$ around the central vertex $\mathbf{p_i}$. The 1-ring contains all faces sharing $\mathbf{p_i}$. Additionally each face $\mathbf{t}$ has a normal vector denoted as denoted by $\mathbf{n}$, which are normalized $\hat{\mathbf{n}} = \mathbf{n}/|\mathbf{n}|$ before e.g. computing the dot product $\langle \hat{\mathbf{n}}_i, \hat{\mathbf{n}}_j \rangle$. Furthermore we compute a normal vector $\mathbf{n_i}$ for each vertex $\mathbf{p_i}$ using the normals $\mathbf{n}_j$ of the adjacent faces $\mathbf{t}_j$. Experiments have shown that this approximation is sufficient for our algorithm and more complex methods like normal vector voting [11] are not necessary.

2.1. Retrieval and simplification of ordered 1-rings

For the following steps of the non-maximum suppression the vertices next to each other are required to be in the sequence given by the orientation of the edges. Our algorithm then uses the implicitly given adjacencies of the mesh to fetch all faces ot the 1-ring of $\mathbf{p_i}$ following the orientation of the edges, adding the vertices $\mathbf{p_j}$ to a sorted list without duplicates excluding $\mathbf{p_i}$. *GigaMesh* ensures that non-manifold vertices and edges are removed [7, p. 121] before computing the sorted list. If $\mathbf{p_i}$ is a vertex on the border $\partial\mathcal{M}_2$ of the mesh, a second iteration using the opposite orientation of faces' edges is necessary – otherwise an arbitrary number of vertices of the 1-ring will be missing.

As subsets of consecutive vertices $\mathbf{p_j}$ can be on a plane the 1-ring has to be simplified to provide a robust tracing of ridge points. For this reason each subset of consecutive vertices are reduced to one representative vertex denoted as $\mathbf{p}'$ in the following example, which is shown in Figure 1. It shows consecutive vertices $\{\mathbf{p}_5, .., \mathbf{p}_9\}$, which are located together with $\mathbf{p}_i$ in one plane, i.e. the faces defined by those vertices have the same direction of their normals. Theoretically we can detect flat parts within the 1-ring by pairwise computing the dot product of adjacent triangles' normals. Such sets of triangles could be replaced by one bigger triangle. As triangle normals can only provide gradient directions within its 1-ring, we have chosen to use the vertex normals, which can store arbitrary normals computed from a range of methods, e.g. a weighted average or computed using

normal vector voting. Therefore the dot products of consecutive pairs of $\{\hat{\mathbf{n}}_5, .., \hat{\mathbf{n}}_9\}$ is computed, where values of ≈ 1 indicate flat parts. The color map represents the distance to the xy-plane to show the three-dimensional nature of the 1-ring. Figure 1b shows that the neighboring vertices $\mathbf{p}_4$ and $\mathbf{p}_1$ are added to the simplified mesh creating a slight artificial valley to generally avoid flat areas having no gradient direction. The threshold determining if these dot products are ≈ 1 is called ϵ and it is the only parameter to be set by the user.

The first vertex is stored in the list $\mathcal{L}_{group}$ with label ID 0. For each $\langle \hat{\mathbf{n}}_i, \hat{\mathbf{n}}_j \rangle$ within the range ϵ to the previous entry then $\mathbf{p_j}$ will be added to $\mathcal{L}_{group}$ with the same label ID. If not, the label ID will be incremented before inserting the item. The algorithm continues until all vertices $\mathbf{p_j}$ in the 1-ring are processed. When all items are processed, the dot product of the first and last entry of the adjacent vertices list needs to be compared because they are contiguous. If the condition to group the two vertices is met, the label ID of all elements with the current label is changed to 0. Now all adjacent vertices are traversed and a new vertex is created for every label, which is assigned the average function value, position vector and normal vector of the corresponding vertices. The grouping process is equivalent to a run-length encoding. In Figure 1b, this results in the new vertex $\mathbf{p}'$ which is the average of vertices $\mathbf{p}_5$ to $\mathbf{p}_9$. The reduced 1-ring has to contain at least 3 vertices to be a manifold otherwise $\mathbf{p_i}$ is not further considered to be a maximum. In case $\mathbf{p_i}$ is a border vertex the minimum amount of required vertices in the 1-ring is 2.

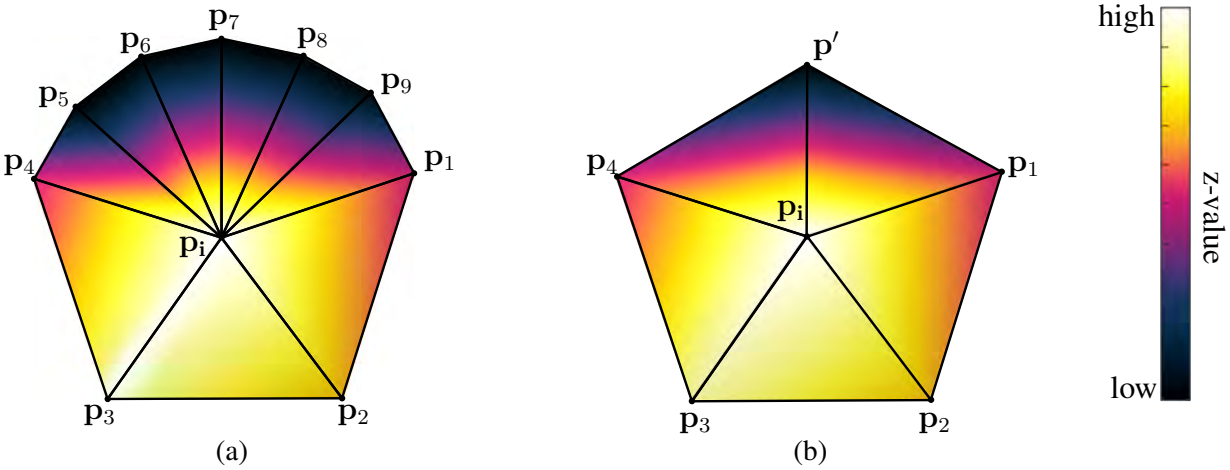


Figure 1: Example of the mesh simplification process. (a) The contiguous vertices $\mathbf{p}_5$ to $\mathbf{p}_9$ lie on a plane. (b) The related faces between vertices have been grouped, resulting in the new vertex $\mathbf{p}'$.

2.2. Principal direction of the gradient value $f(\cdot)$

Analogously to the Canny algorithm, we have to compute the principal direction $\mathbf{t}$ of the gradient. As we typically use the MSII-filter for $f(\mathbf{p_i})$, we have to use the normals to detect $\mathbf{t}$ and its orthogonal secondary direction $\mathfrak{b}$. To achieve this, the dot product $\langle \hat{\mathbf{n}}_\mathbf{i}, \hat{\mathbf{n}}_\mathbf{j} \rangle$ is computed. The vertex $\mathbf{p_j}$ with the largest dot product is the principal direction $\mathbf{t}$ and is saved for later computations. This is illustrated in Figure 2a with $\mathbf{t} = \mathbf{p}' - \mathbf{p_i}$. In Figure 2b $\pm\mathfrak{b} = \pm\mathbf{t} \times \hat{\mathbf{n}}_\mathbf{i}$ is shown. The normal, the principal, and the secondary direction span a *Frenet-Serret frame* (TNB frame) with the planes $\boldsymbol{\tau}_{nt}$ and $\boldsymbol{\tau}_{nb}$.

According to Canny we need the gradient values $\mathbf{p}$ and $\mathbf{q}$ on the secondary directions $\pm\mathfrak{b}$. These are found on the intersections $\mathbf{p}_{\overline{jk}} := \boldsymbol{\tau}_{\mathfrak{bt}} \cap \mathbf{e_{jk}}$ and $\mathbf{p}_{\overline{lm}} := \boldsymbol{\tau}_{\mathfrak{bt}} \cap \mathbf{e_{lm}}$. To compute $f(\mathbf{p}_{\overline{jk}})$ we interpolate linear between the two vertices $\mathbf{p}_j$ and $\mathbf{p}_k$ with the respective function values $f(\mathbf{p}_j)$ and $f(\mathbf{p}_k)$.

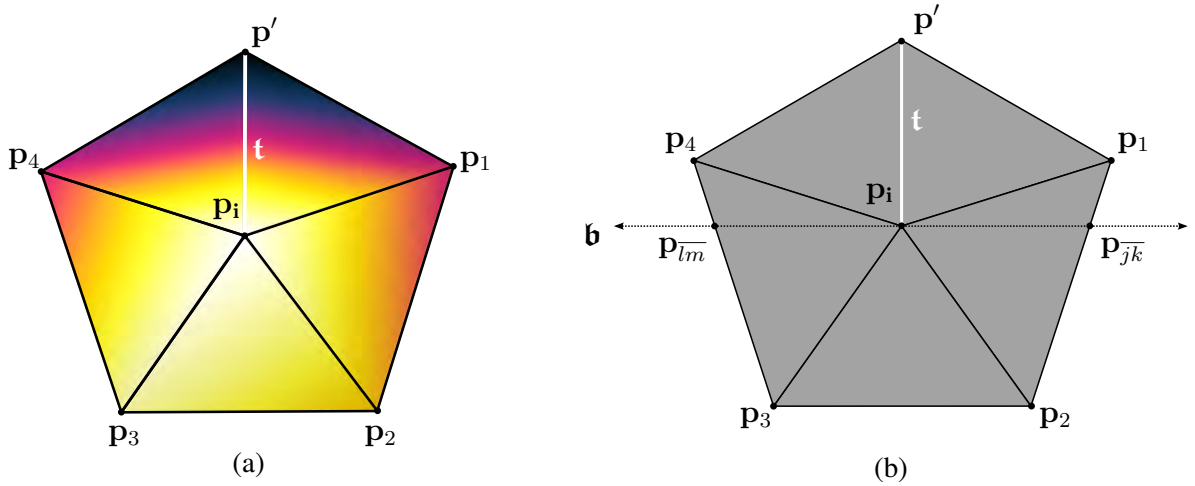


Figure 2: (a) The vector $\mathfrak{t}$ describes the principal direction outbound from $\mathbf{p_i}$. (b) The vertices on the orthogonal secondary direction $\mathfrak{b}$ are $\mathbf{p}_{\overline{jk}} = \mathbf{p}_{\overline{12}}$ and $\mathbf{p}_{\overline{lm}} = \mathbf{p}_{\overline{34}}$.

2.3. Non-maximum suppression with border treatment

Finally we distinguish vertices being maxima from those being non-maxima:

- If either $f(\mathbf{p}_{\overline{jk}})$ or $f(\mathbf{p}_{\overline{lm}})$ is larger than $f(\mathbf{p_i})$, then $\mathbf{p_i}$ is suppressed by discarding this vertex.
- Otherwise $\mathbf{p_i}$ is a maximum and added to the list $\mathcal{L}_{max}$.

In case $\mathbf{p_i}$ is on the border $\partial\mathcal{M}_2$, we treat the vertex by checking the existance of the edges $\mathbf{e}_{jk}$ and $\mathbf{e}_{lm}$ intersecting the plane $\boldsymbol{\tau}_{nb}$. For existing edges we proceed as described above. Otherwise we have to choose a function value of $\mathbf{p_j}$ close to $\boldsymbol{\tau}_{nb}$: If there is an edge $\mathbf{e}_{ij}$ with $\langle \hat{\mathbf{e}}_{ij}, \mathfrak{b} \rangle > 0$ we choose the function value $f(\mathbf{p}_j)$ of the edge having the dot product closest to 1. Having no positive value for the dot product leads to suppression of the vertex. This procedure is repeated using $-\mathfrak{b}$ for the second secondary direction.

3. Results

The execution times for various real world and synthetic test cases behave linear, depending on the number of vertices of the mesh. This heuristically determined computational complexity of $O(n)$ with n being the number of vertices is shown in Table 1. The resulting ridge points on a detail of a three-dimensionally acquired cuneiform tablet is shown in Figure 3. These selected points can be exported using the current view and its underlying *OpenGL* projection matrix within *GigaMesh* either as perspective or as orthogonal projection. The latter is true to scale assuming a calibrated 3D-model. While the surfaces are rendered as raster images, the ridge points are exported as overlays using the *Scalable Vector Graphics* (SVG) [4] file format, which describes their exact location using the *eXtensible Markup Language* (XML), commonly used within the *Digital Humanities*.

Results on a high resolution data set are shown in Figure 4a. Due to the high density of vertices, the algorithm responds to small disturbances i.e. noise of the surface, leading to false positives. These can be eliminated by smoothing the surface prior to the application of our algorithm. In Figure 4b a combination of Taubin and TwoStep smoothing was applied using *MeshLab* [3]. This increase in robustness behaves – as expected – like the Canny edge detector, which has a smoothing step as a prerequisite.

Data set	Type	Vertices	Runtime
Chars4Testing	measured	10,492	0.044s
cuneus_ideal	synthetic	15,521	0.071s
Half4Testing	measured	282,428	0.789s
HOS_G10_Preview	measured	371,711	1.809s
VAT 10908	measured	3,034,899	12.269s
HOS_G10_Full	measured	6,596,964	33.319s

Table 1: Performance of the algorithm on multiple data sets. Dataset name, type, number of vertices and the respective runtime are given.

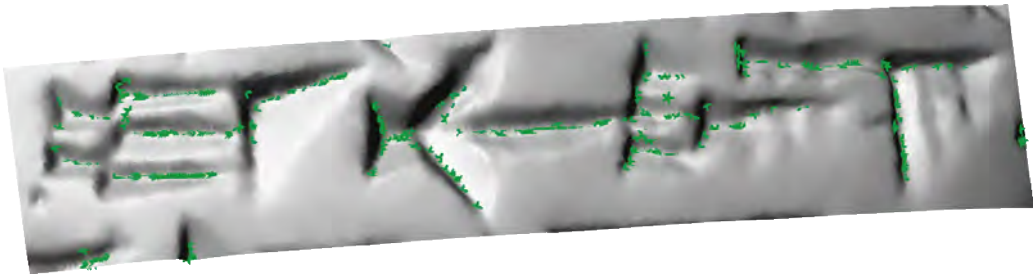
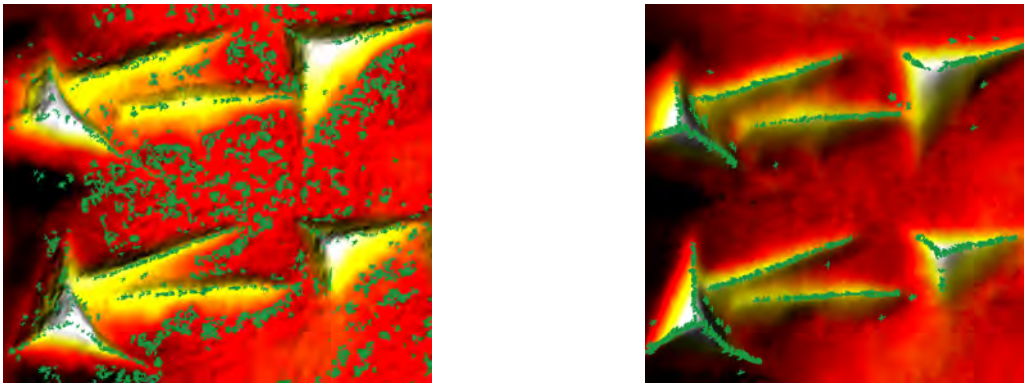


Figure 3: Detected ridge points in the real world data set Chars4Testing. It can be seen that the points follow the ridges of the mesh nicely.



(a) (b)

Figure 4: High resolution dataset HOS_G10_Full (a) before and (b) after smoothing.

4. Outlook and Summary

Future enhancements of our algorithm are the implementation of a marching front to connect the ridge points to lines, making them exportable as SVG. Following the Canny approach, hysteresis tracking is a future extension providing an even more robust selection of feature points. Furthermore smoothing of the function values instead of smoothing the mesh will improve the performance by reducing the computational overhead of processing $\mathcal{M}_2$. The final vision is to have a completely autonomous system, which begins transcribing the ancient tablets immediately after their acquisition, exporting the digital drawings with automated annotations directly into a searchable database [6].

The algorithm implemented in this work succeeds in extracting ridge points from irregular grids, using non-maximum suppression. Although the execution on an irregular surface mesh architecture

contains various challenges, all of them could be resolved. The most important challenges were the mesh simplification step and the determination of the maximum gradient direction. We could show the adaptation of the Canny edge detector, used on regular grids to irregular triangular meshes in $\mathbb{R}^3$. The necessary user input is kept to a minimum, namely only one parameter, which controls the strength of the local and temporary mesh simplification. In general our algorithm delivers robust approximations with high performance used for further processing with methods from machine learning [5].

References

[1] B. Bogacz, M. Gertz, and H. Mara. Character Retrieval of Vectorized Cuneiform Script. In *Proc. of Int. Conf. on Document Analysis and Recognition (ICDAR)*, pages 326–330. IEEE, 2015.

[2] J. Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[3] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Italian Eurographics Conf.*, pages 129–136, 2008.

[4] J. David Eisenberg. *SVG Essentials*. O'Reilly, 1 edition, 2002.

[5] D. Fisseler, F. Weichert, G. Müller, and M. Cammarosano. Towards an Interactive and Automated Script Feature Analysis of 3D Scanned Cuneiform Tablets. In *Proc. of the 4th Scientific Computing and Cultural Heritage*, Heidelberg, Germany, 2013.

[6] B. Groneberg, F. Weiershäuser, T. Linnemann, and D. Ullrich. Digitale Keilschriftbibliothek Lexikalischer Listen aus Assur. In *Max-Planck-Gesellschaft – Jahrbuch*. Max-Planck-Gesellschaft, 2005.

[7] H. Mara. *Multi-Scale Integral Invariants for Robust Character Extraction from Irregular Polygon Mesh Data*. PhD thesis, Ruprecht-Karls-Universität, Interdisciplinary Center for Scientific Computing (IWR), Heidelberg, Germany, 2012.

[8] H. Mara and S. Krömker. Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes. In *Proc. of the Int. Conf. on Document Analysis and Recognition (ICDAR)*, pages 62–66. IEEE, 2013.

[9] H. Mara, S. Krömker, S. Jakob, and B. Breuckmann. GigaMesh and Gilgamesh - 3D Multi-scale Integral Invariant Cuneiform Character Extraction. In A. Artusi et. al., editor, *Proc. VAST Int. Symposium on Virtual Reality, Archaeology and Cultural Heritage*, pages 131–138, Paris, France, 2010. Eurographics Association.

[10] S.M. Maul. *Das Gilgamesch-Epos*. Beck, 2005.

[11] D. L. Page, Y. Sun, A. F. Koschan, J. Paik, and M. A. Abidi. Normal Vector Voting: Crease Detection and Curvature Estimation on Large, Noisy Meshes. *Graphical Models*, 64(3–4):199–229, 2002. Special issue: Processing on large polygonal meshes.

[12] R. Sablatnig and C. Menard. Stereo and Structured Light as Acquisition Methods in the Field of Archaeology. In *Mustererkennung 1992*, pages 398–404. Springer, 1992.

[13] W. von Soden. *The ancient Orient: an introduction to the study of the ancient Near East*. Wm. B. Eerdmans Publishing Co., 1994.

Noise Robustness of Irregular LBP Pyramids

Christoph Körner, Ines Janusch, Walter G. Kropatsch

Pattern Recognition and Image Processing (PRIP)
Vienna University of Technology, Austria
{*christoph,ines,krw*}*@prip.tuwien.ac.at*

Abstract
In this paper, we briefly introduce the SCIS algorithm - a hierarchical image segmentation approach based on LBP pyramids - and evaluate its robustness to uniform, Gaussian, and Poisson distributed additive chromatic noise. Moreover, we study the influence of image properties such as the amount of details and SNR on the segmentation performance. Our evaluation shows that SCIS is robust to Gaussian and Poisson noise for our testing environment.

1. Introduction

Local binary patterns (LBPs) were originally introduced as a texture descriptor by Ojala et al. in 1994 [14]. Due to their computational simplicity and their robustness to varying lighting conditions LBPs have since become popular texture operators. In order to compute the LBP for a certain pixel, this pixel is compared to its subsampled neighbourhood. In case the value of a neighbouring pixel is larger than or equal to the value of the center pixel its bit is set to 1 otherwise to 0. The resulting bit pattern thus describes the neighbourhood relations. The bit pattern may be transformed to a decimal number by encoding each neighbourhood pixel using its position in a binary data item.
Image pyramids provide a multiscale representation of an image by applying smoothing and subsampling to this image repeatedly. Burt proposed in 1981 such an approach using a Gaussian like smoothing [1]. The well known Laplacian pyramid was later introduced by Burt and Adelson in [2]. For the Laplacian pyramid (except for the top level) the difference images of successive layers of a Gaussian pyramid are stored instead of the Gaussian smoothed images itself. A reconstruction of the original image is possible based on its Laplacian pyramid representation. Image pyramids are for example used when computing multi-scale image features as it is done by SIFT (scale invariant feature transform) [11] or for image compression (as described in [2]).
Both LBPs and image pyramids among other applications have been used individually in image segmentation algorithms:
Chen et al. [5] and Heikkilä et al. [8] for example use LBPs for segmentation purposes. These approaches however use LBP histograms, the spatial information of LBPs is therefore lost. Two visually completely different images may have the same LBP histogram - a major drawback of these approaches.
For hierarchical image segmentation a wide range of approaches has been published in the past: Kropatsch et al. present in [9] a hierarchical segmentation method based on minimum weight spanning trees of graph pyramids. A similar approach that allows user interaction during the segmentation process is presented by Gerstmayer et al. in [7]. A hierarchical image segmentation approach based on the feature detector MSER (maximally stable extremal regions) was proposed by Oh et al. in [13].
In this paper we discuss a recent image segmentation approach that combines LBPs and combinatorial pyramids - the structurally correct image segmentation algorithm (SCIS) introduced by Cerman [3]. Using this approach highly textured regions are merged late in the segmentation hierarchy. Thus,

preserving visual information that is important for human perception up to high levels of the pyramid. It is known that standard image pyramids such as Gaussian and Laplacian pyramids as well as hierarchical representations based on these concepts eliminate noise in the image due to the repeated smoothing operation [6]. However, since regions showing noise may also be consider as highly textured regions this may not be the case for SCIS. Therefore, we analyze the noise robustness of SCIS in this paper.

The rest of the paper is structured as follows: A short introduction to the SCIS algorithm is given in Section 2. Its robustness to noise is tested in experiments presented in Section 3. Results of these experiments are discussed in Section 4. Section 5. concludes the paper and gives an outlook to future work.

2. Structurally Correct Image Segmentation

The "structurally correct image segmentation" (SCIS) algorithm was first presented in [3]. Although, SCIS is based on LBPs it does not use histograms. This hierarchical segmentation approach constructs an irregular graph pyramid (sequence of reduced graphs), by iteratively identifying and removing redundant structural information, and merging regions with the lowest dissimilarity first.

The SCIS algorithm represents an image as a directed acyclic graph. Each vertex corresponds to a pixel, a superpixel or a region, and each edge corresponds to an adjacency relationship between two pixels, superpixels or regions. After merging neighboring vertices with equal grayscale-/color values, it is possible to assign each edge a direction, and thus describing the relationship between adjacent vertices as strict inequality relationships. As a result, this merging induces a strict partial order onto the vertices of the image graph. This ordering of the vertices is not a total ordering, because not all pairs of vertices are comparable by following monotonically increasing or decreasing paths. An edge is said to be "structurally redundant", if the removal of this edge does not break the reachability property of the graph. The SCIS algorithm identifies most of these redundant edges in a fast manner by means of a primal and dual topological LBP classification (see [4] for more detailed information) and removes them.

SCIS (see Algorithm 1) employs this idea, mentioned as *simplifying the structure* in algorithm 1 (line 5-8). Structurally redundant dual edges are determined and removed. Subsequently, regions with the lowest dissimilarity are merged. In the case of grayscale images, the absolute region intensity difference $d(x, y) = |g(x) - g(y)|$ is used, where x and y are two regions, and $g(x)$ is the

Algorithm 1 structurally correct image segmentation (SCIS)

input: 2D image
output: combinatorial pyramid

1: $k := 0$
2: initialize base level C of combinatorial pyramid
3: $C' :=$ remove dual saddles in C
4: $C_0 :=$ merge plateaus in C'
5: **repeat**
6: $\quad k := k + 1$
7: $\quad$ simplify structure in current level C_k
8: **until** $C_k = C_{k-1}$

intensity of x. For color images, the current implementation uses the CIEDE2000 color difference [15]. During the merging of regions, the algorithm computes the new value of the region as the mean value of all included pixels, and verifies, if this merging does not break the strict partial ordering of the previous graph.
In practice it is sufficient to remove redundant dual edges and to check that a newly computed value fits the surrounding LBP values. The remaining dual edges are then sorted according to their contrast. The dual edge with the lowest contrast is considered first and checked if it can be merged. Therefore, a new value for the merged dual vertices is computed (in our case this is the color mean) and if this value satisfies the binary relationships stored at the incident dual edges the edge is contracted. If not, the dual edge with the next lowest contrast is considered. This process is repeated until a suitable dual edge for merging is found. This way, by removing edges with the lowest contrast first, regions with low contrast are merged first.
Therefore, visual information that is important for humans is preserved even at high levels of reduction since highly textured regions are merged late in the hierarchy. This way, around 70 percent of regions can be merged in an image, with only a minimal loss of information important to humans. In many cases, merging up to around 94 percent is possible with visually acceptable results.

3. Experimental Setup

Depending on the technology for capturing a picture (analog or digital) and storing the picture (compressed or uncompressed) various types of noise are introduced to the resulting image. In this paper, the SCIS algorithm is evaluated regarding its robustness to common types of noise in digital images, such as quantization noise, sensor noise and shot noise.

3.1. Types of Noise

Quantization noise is introduced when the sensor of a digital camera maps the incoming light intensity to quantized levels of color values for each pixel. For the experiments, this noise type is modeled as uniform distributed chromatic additive noise with an amplitude of 50; the average SNR of the test images is $-9.03 \pm 0.34dB$. Figure 1a shows the normalized distribution of the quantization noise model.
Sensor noise can be caused by multiple environmental effects in a digital image sensor, such as bad lighting conditions, thermal conditions, and many more. For the experiments, this noise type is modeled as a Gaussian distributed chromatic additive noise with a σ of 5 and centered around 0; the average SNR of the test images is $6.20 \pm 0.34dB$. Figure 1b shows the normalized distribution of the sensor noise model.
Shot noise is introduced due to the fluctuations of the amount of photons that hit the sensor for a given exposure. For the experiments, this noise type is modeled as Poisson distributed chromatic additive noise with a λ of 50 centered around 0; the average SNR of the test images is $3.19 \pm 0.33dB$. Figure 1c shows the normalized distribution of the shot noise model.
Figure 2 shows these noise types applied to a sample image.

3.2. Image Data Set and Ground Truth

The SCIS algorithm is tested on 26 animal and landscape images of the Berkeley Image Segmentation data set [12] (see Figure 2a for a sample test image). This data set includes multiple ground truth segmentations per test image which have been segmented by humans. For the evaluation of the segmentation error for each test image, the first ground truth reference in alphabetic order is used (see Figure 2b for a sample ground truth image).

The segmentation error is evaluated on the reconstructed grayscale images of the LBP pyramid on a range from 0 to 5000 segments. Figure 3 shows the reconstructed images of the LBP pyramid for 200, 500, 1000 and 2000 segments as well as the magnitudes of the 2D Fourier transform. It is well visible that the frequencies are evenly distributed. A Fourier transform of a Laplacian pyramid shows circular structures due to the bandpass effect of this pyramid, for a Gaussian pyramid lowpass effects are visible in the frequency domain. Since these effects are not visible in the Fourier transforms of the LBP pyramid we conclude that the LBP pyramid does not have a bandpass nor a lowpass effect.

3.3. Validation Methodology

For estimating the empirical segmentation error against the ground truth images, the region-based segmentation measurement Global Consistency Error (GCE) [12] is used. It is a robust technique and independent of the number of segments in each image. The GCE is defined as:

$$GCE(S_1, S_2) = \frac{1}{n} min\Big(\sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i)\Big) \tag{1}$$

To measure the same local refinement error E when changing the order of the reference image such that $E(S_1, S_2, p_i) = E(S_2, S_1, p_i)$, we take the minimum of both sums over all pixels in the GCE computation. In order to define E, we first denote the set difference of A and B as $A \setminus B$, and $|A|$ the cardinality of the set A. Let $R(S, p_i)$ be the set of pixels in the segmented image S that correspond to the region R containing pixels p_i, then the local refinement error E is defined as:

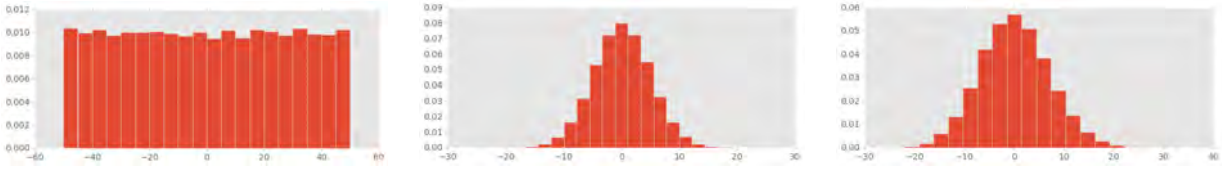
$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{|R(S_1, p_i)|} \tag{2}$$

4. Results

The GCE error is evaluated for all 26 test images for the range of 0 to 5000 segments, both for the reconstruction of the original images and of the noisy images. Figure 4 shows the original and the noisy images reconstructed with a different number of segments (compare with Figure 2b showing the ground truth).

As we can observe in Figure 4b, we expect the noise to introduce additional high frequencies to the original image and hence to result in smaller regions compared to the reconstruction of the original image (Figure 4a) for the same number of segments. For a bigger SNR, this effect should be less visible such as in Figure 4c and 4d.

Figure 5a shows the GCE evaluation of the reconstructed test images with an increasing number of segments. If one compares the GCE obtained from the original segmentation to the segmentation of the images with uniformly distributed noise (see Figure 5b), we observe that the GCE curves are



(a) Uniform noise (b) Gaussian noise (c) Poisson noise

Figure 1: Normalized noise distributions

(a) Original (105025.jpg)

(b) Ground truth

(c) Uniform noise

(d) Gaussian noise

(e) Poisson noise

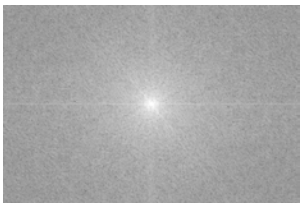
Figure 2: Image with overlaying noise

(a) 200 segments

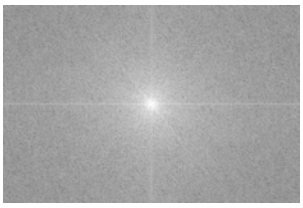
(b) 500 segments

(c) 1000 segments

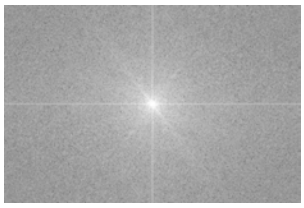
(d) 2000 segments



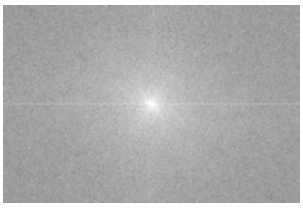
(e) 2D FFT of 200 segm.



(f) 2D FFT of 500 segm.



(g) 2D FFT of 1000 segm.



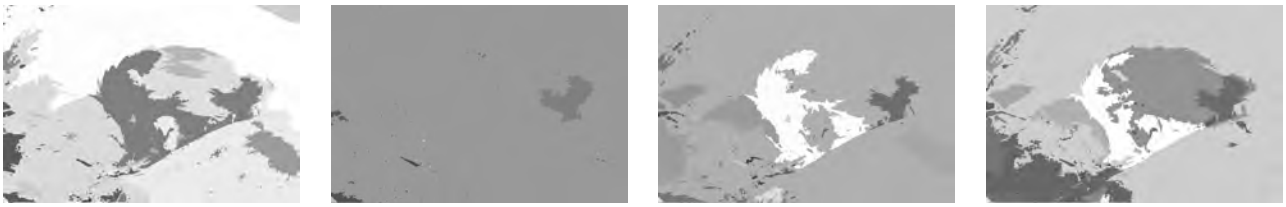
(h) 2D FFT of 2000 segm.

Figure 3: Reconstructions from the LBP pyramid (top) and their Fourier transforms (bottom).

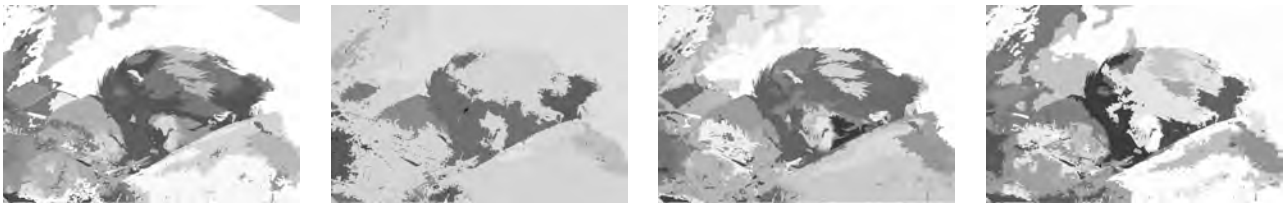
shifted to the right (towards increasing number of segments). The shift is around 200 to 500 segments for images with a low SNR and between 500 and 1500 segments for images with a better SNR. For Gaussian (see Figure 5c) and Poisson (see Figure 5d) distributed noise the shifts are more concentrated between 200 and 1000 segments and the GCE curve rises steeper.

However, this is not exactly the behavior that we were expecting. A shift of the GCE curve to the right means that for the same number of segments the GCE of the noisy image is lower than for the original image. This effect can be better observed when looking at the difference of the GCE from the reconstruction of the test images and noisy images in Figure 6, in the range of 200 to 1500 segments. In this figure, a positive value corresponds with a lower GCE than in the original image whereas a negative value corresponds with a greater GCE.

For uniformly distributed noise (see Figure 6a), we can differentiate between 2 types of images in the range below 1500 segments: one group with a lower GCE than the original images and the other group with a greater GCE. These image groups correlate with the amount of details in an image as

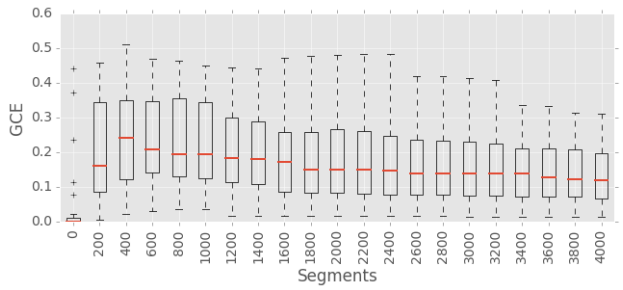


(a) Original 500 seg. (b) Uniform noise 500 seg. (c) Gaussian noise 500 seg. (d) Poisson noise 500 seg.

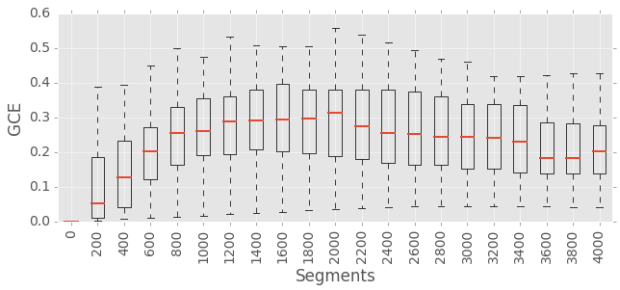


(e) Original 2000 seg. (f) Uniform noise 2000 seg. (g) Gaussian noise 2000 seg. (h) Poisson noise 2000 seg.

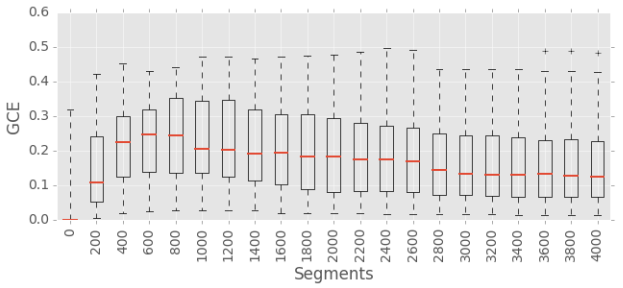
Figure 4: Reconstructed original and noisy images



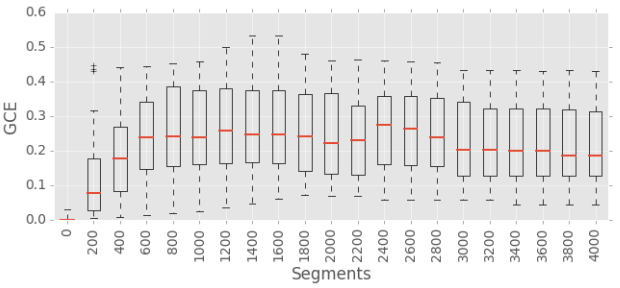
(a) Original images



(b) Images with uniform noise



(c) Images with Gaussian noise

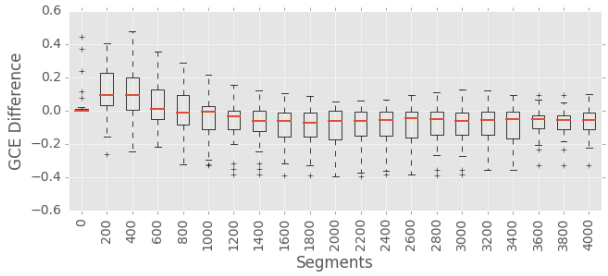


(d) Images with Poisson noise

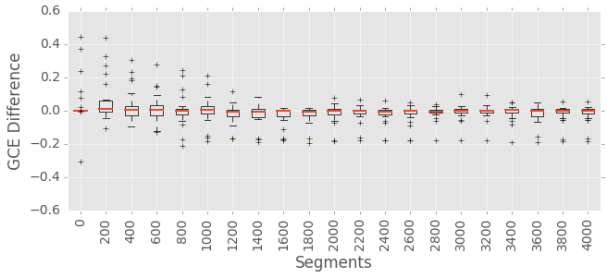
Figure 5: GCE segmentation error of the test images and noisy images

well as the SNR: high amount of details (low SNR) and average amount of details (higher SNR). The latter group of images corresponds with the previously expected behavior.

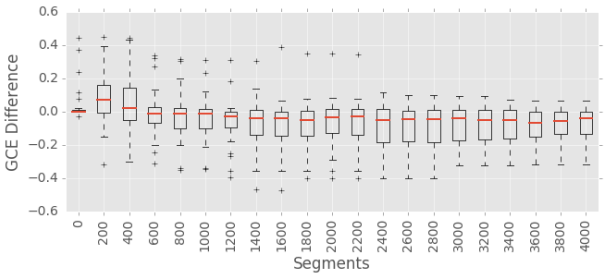
For images with Gaussian distributed noise (see Figure 6b), we conclude that the GCE is almost the same as for the original images with a few outliers at maximal difference of 0.2. The Poisson distributed noise (see Figure 6c) has a similar behavior to uniformly distributed noise but less outliers.



(a) Images with uniform noise



(b) Images with Gaussian noise



(c) Images with Poisson noise

Figure 6: Difference of the GCE segmentation error of original images and noisy images

5. Conclusion

The SCIS algorithm shows a maximal decrease of 0.2 for outliers in the GCE for images with Gaussian distributed noise for reconstructions using more than 200 segments (see Figure 6b). The GCE stays the same also for increasing σ and sometimes gets even lower for reconstructions using less than 1000 segments. Hence, it can be said that the SCIS is robust to Gaussian noise under the constraints of the testing environment.

For reconstructions using less than 1000 segments, the SCIS algorithm is very sensitive to uniform noise leading to both better and worst segmentation result strongly depending on the SNR and the amount of details in the image (see Figures 5b and 6a). Also for reconstructions using more than 1000 segments, the mean difference to the original GCE is around 0.075 with many outliers around 0.4. However, it should be noted that the tested noise amplitude was slightly higher compared to quantization noise in common digital sensors.

For Poisson distributed noise, the mean difference to original GCE values is less than 0.05 for reconstructions above 1000 segments with only a few outliers up to 0.5. For less segments, the behavior is similar to uniform distributed noise. Hence, we conclude that the SCIS algorithm is also robust to Poisson noise under the constraints of the testing environment.

We have shown that SCIS algorithm achieves good segmentation results for images with chromatic additive Gaussian and Poisson distributed noise and is sensitive to uniformly distributed noise. The experiments could be extended to also evaluate monochromatic noise and other relevant noise types e.g. Salt-and-pepper noise.

Acknowledgments

We thank Martin Cerman for assistance with the SCIS algorithm and helpful comments on the experiments and interpretation of the results.

References

[1] P. J. Burt. *Fast filter transform for image processing.* Computer graphics and image processing vol.16.1 pp. 20–51 (1981).

[2] P. J. Burt, E. H. Adelson. *The Laplacian pyramid as a compact image code.* IEEE Transactions on Communications. vol.31.4 pp. 532–540 (1983).

[3] M. Cerman: *Structurally Correct Image Segmentation using Local Binary Patterns and the Combinatorial Pyramid*. Technical Report 133[1], Vienna University of Technology, Pattern Recognition and Image Processing (PRIP) Group. (2015)

[4] M. Cerman; R. Gonzalez-Diaz; W. G. Kropatsch: *LBP and Irregular Graph Pyramids.* In the Proceedings of CAIP 2015, Part II, pp.687-699 (2015).

[5] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, W. Gao. *WLD: A robust local image descriptor*. IEEE TPAMI, vol. 32(9), pp. 1705–1720 (2010)

[6] S. Contassot-Vivier, G. L. Bosco, N. C. Dao. *Multiresolution approach for image processing.* Erasmus ICP-A-2007 (1996).

[7] M. Gerstmayer, Y. Haxhimusa, W. G. Kropatsch. *Hierarchical interactive image segmentation using irregular pyramids.* Graph-Based Representations in Pattern Recognition. Springer Berlin Heidelberg, pp.245–254 (2011).

[8] M. Heikkilä, M. Pietikäinen: *A texture-based method for modeling the background and detecting moving objects*. IEEE TPAMI, vol. 28(4), pp. 657–662 (2006).

[9] W. G. Kropatsch,Y. Haxhimusa, A. Ion *Multiresolution image segmentations in graph pyramids.* Applied Graph Theory in Computer Vision and Pattern Recognition, Springer Berlin Heidelberg, pp. 3–41 (2007).

[10] LBP'2014 Workshop on Computer Vision With Local Binary Pattern Variants [2]

[11] D. G. Lowe. *Distinctive image features from scale-invariant keypoints.* International journal of computer vision vol.60.2 pp.91–110 (2004).

[12] D. Martin; C. Fowlkes; C. Tal; J. Malik: *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.* In the Proceedings of ICCV 2001, pp.416-423 (2001).

[13] I.S. Oh, J. Lee, A. Majumder. *Multi-scale image segmentation using MSER.* In the Proceedings of CAIP 2013, Part II, pp. 201–208 (2013).

[14] T. Ojala, M. Pietikainen, D. Harwood. *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions.* Proceedings of ICPR 1994, pp. vol.1 582-585 (1994).

[15] G. Sharma, W. Wu, E. N. Dalal. *The CIEDE2000 Color-Dierence Formula: Implementation Notes*, Supplementary Test Data, and Mathematical Observations. Color Research and Rpplication, vol.30.1 pp.21–30 (2005).

[1] ftp://ftp.prip.tuwien.ac.at/pub/publications/trs/tr133.pdf

[2] https://sites.google.com/site/lbp2014ws/

WS 8: Task Planning

Controlling and Tracking an Unmanned Ground Vehicle with Ackermanndrive

Eugen Kaltenegger[1], Benjamin Binder[1], Markus Bader[2]

Institute of Computer Aided Automation
Vienna University of Technology, Austria

Abstract
This work presents a tracking and control mechanism for an UGV (Unmanned Ground Vehicle) and its integration into ROS (Robot Operating System). The overall goal of which this work is part, is the creation of a fleet of ackermann robots to conduct studies in the field of autonomous driving. In order to achieve this goal a 1:10 RC-race car model is equipped with an Arduino board to control the vehicles actuators and a Raspberry Pi to host the ROS server. In addition, a physics simulation is used to model this car for testing. The shown results support the used velocity motion model and the applicability of the developed interface to control both platforms.

1. Introduction

During the last years, many studies have been conducted in the field of autonomous driving [6, 1] and the automotive industries as well as companies like Google are showing great interest in this market. Up to 2007, competitions like the DARPA Grand and Urban Challenge pushed research towards autonomous cars with great success [4]. Nowadays, events like the Freescale Cup[1], the Carolo-Cup[2] and others are created to target young students by using RC-race car models which in terms of costs are very attractive. With this in mind, the Institute of Computer Aided Automation at the Technical University of Vienna is planning to create a fleet of autonomous ackermann robots to attract students and to at one point take part in such a competition.
This paper describes the creation of the first of these vehicles and its simulation while also introducing a common interface and a tracking system supporting them.
The robot is based on a RC-race car with an ackermann steering. The computation and controlling of the robot is achieved through a Raspberry Pi and an Arduino Uno equipped with a motor-shield. In addition the vehicle is simulated with Gazebo [8], an open source software for physical simulation based on ODE (Open Dynamics Engine). To keep the vehicle and the simulation compatible, the same interface is used, which is based upon the open source software ROS (Robot Operating System) [2]. The vehicle and the simulation are both using the same velocity motion model [4] which equals the prediction step of a Kalman filter for motion tracking [4, 7]. The velocity motion model introduced by Thrun is defined for differential drive robots, but given several changes, which will be further explained, it can also be used for ackermann robots. Since they are commonly used in robotics and provide enough information for an ackermann robots motion, differential drive commands are chosen as input. A ROS node transforms the differential drive commands to ackermann commands, which include a velocity and a steering angle. The robot and its simulation publish their estimated pose and its uncertainty into ROS topics. This allows for easy comparison of the trajectory driven by the

[1]Freescale Cup: https://community.freescale.com/docs/DOC-1284 (25.04.2016)
[2]Carolo-Cup: https://wiki.ifr.ing.tu-bs.de/carolocup/ (25.04.2016)

RC-race car, its simulation and the motion model.
This paper is structured as follows. At first, related research is introduced and the interface as well as the adapted velocity motion model are described. Based on this knowledge, the robot and its simulation are annotated and their basic structure is discussed. Additionally, the trajectories of the two vehicles are compared and the reasons why the trajectories and the motion model deviate from each other are explained. Finally, further improvements for the adapted motion model in use with the robot and the simulation are introduced.

2. Related Work

Autonomous driving is currently a research topic of both major automobile manufacturers like Volvo, Ford or Nissan and newcomers to the topic of automobiles like Google [6]. At the DARPA urban challenge, universities like the Massachusetts Institute of Technology and the Stanford University present their research accomplishments [3]. An example for research on autonomous vehicles with ackermann drive using ROS is Marvin, the autonomous car by the University of Texas at Austin. Members of the Marvin-Team ported the software of the autonomous car to ROS and shared it this way. The ackermann group represents a community developing open source ROS packages for such vehicles. For the project discussed within this paper, ROS is used because is allows to combine and enhance such packages for navigation and odometry. Twist messages[3] and ackermann messages[4] are used to control the robot, and odometry messages[5] are used for tracking. The structure of ROS allows to combine all these different messages contained in different packages into one interface.

3. Interface

The interface is created to ensure compatibility between the robots of the fleet and the simulation. For that reason, the interface converts twist messages to ackermann messages. It also converts these ROS messages into serial commands and vice versa for those vehicles unable to run ROS. The converting structure of the interface is shown in Figure 1.
Twist messages are commonly used as motion commands because the six parameters they hold provide enough information to define motions in a three dimensional space. In the further, twist messages holding only one linear velocity and one angular velocity are assumed, since they provide enough information for motions in a two dimensional space. Ackermann messages contain a velocity, a steering angle and information about the acceleration and the jerk. The last two are not used for this project.
The velocity of the ackermann messages equals the linear velocity of the twist messages. The steering angle of the ackermann messages can be calculated with the knowledge of the cars geometry. A curve radius of an imaginary third front wheel is calculated by dividing the rotational velocity of the twist message by the its linear velocity. With this radius, the knowledge of the wheelbase and the usage of trigonometric functions, the steering angle φ can be calculated. Based on the motion commands the car and its simulation receive, they return odometry messages containing the estimated pose and its uncertainty. This information is calculated based on the motion model.

[3]Twist Messages: http://wiki.ros.org/geometry_msgs (25.04.2016)
[4]Ackermann Messages: http://wiki.ros.org/ackermann_msgs (25.04.2016)
[5]Odometry Messages: http://wiki.ros.org/nav_msgs (25.04.2016)

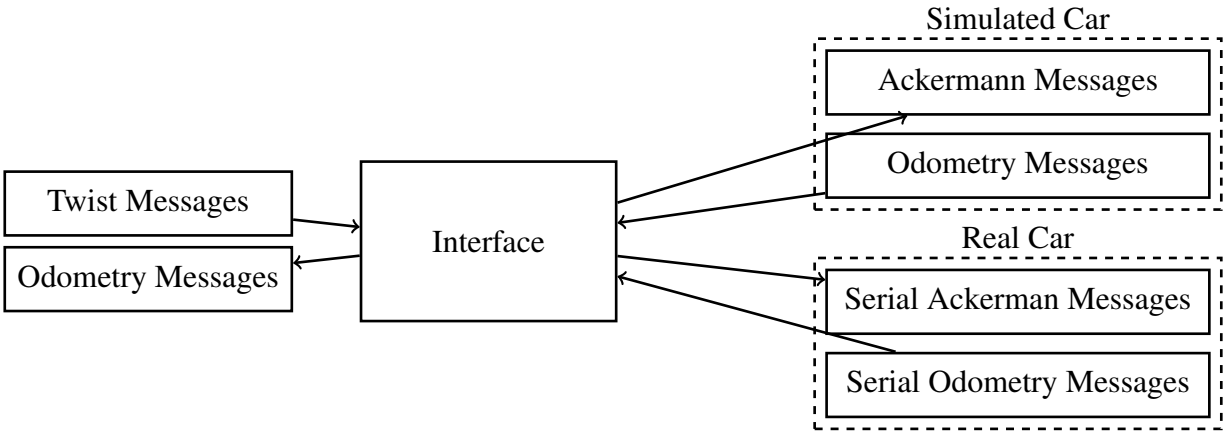


Figure 1: The interaction of the different messages of the interface.

4. Motion Model

Simple problems like wheel slips, bumps, and inaccuracies within the robot effect the robots motion [4, 7]. The tracking system for this project is based on the velocity motion model which considers these errors. Although the velocity motion model introduced by Thrun [4] is conceived for robots able to turn around their own axis, its simple structure allows its customization for ackermann drive robots. Based on the motion commands, the velocity motion model calculates the robots estimated pose and a matrix in which its uncertainty is contained. This is equal to the prediction step of a Kalman filter [4, 7].
Further, the motion model is applied to a flat space represented by x and y and the parameter θ which stands for the robots orientation as shown in Figure 2.

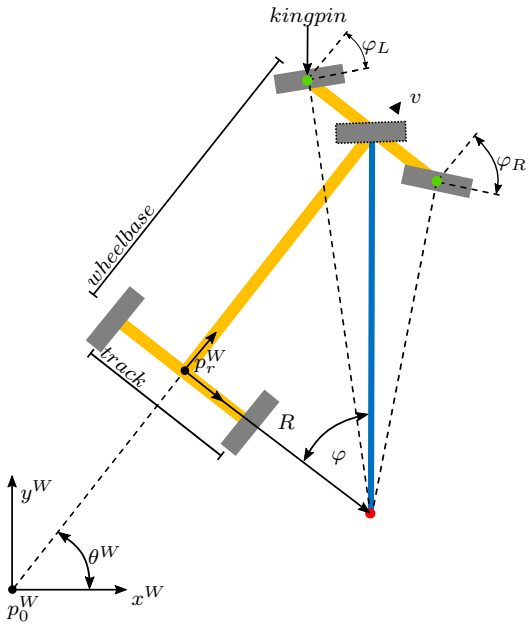


Figure 2: The geometry of the ackermann robot in the two dimensional space.

By adding the change of x, y and θ in one time step to the previous pose, the robots pose at any given time can be calculated recursively.

$$\mathbf{x}_t\left(\mathbf{x}_{t-1}, \mathbf{u}\right) = \begin{pmatrix} x_t = x_{t-1} + v \cdot \cos\left(\theta_{t-1}\right) \cdot \Delta t \\ y_t = y_{t-1} + v \cdot \sin\left(\theta_{t-1}\right) \cdot \Delta t \\ \theta_t = \theta_{t-1} + \frac{v \cdot \tan(\varphi)}{w_{wheelbase}} \cdot \Delta t \end{pmatrix} \tag{1}$$

The change of the pose is represented by the jacobian matrix $G\left(\mathbf{x}_{t-1}, u\right)$, which is the derivative of the pose $\mathbf{x}_{t-1}$ with respect to the pose x_{t-1}.

$$G = \frac{\partial \mathbf{x}_t\left(\mathbf{x}_{t-1}, \mathbf{u}\right)}{\partial \mathbf{x}_{t-1}} = \begin{pmatrix} 1 & 0 & -v \cdot \sin\left(\theta_{t-1}\right) \cdot \Delta t \\ 0 & 1 & v \cdot \cos\left(\theta_{t-1}\right) \cdot \Delta t \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

The jacobian matrix $V\left(\mathbf{x}_t, u\right)$ is the derivative of the pose $\mathbf{x}_{t-1}$ with respect to the motion command u. This equals the change of the motion.

$$V = \frac{\partial \mathbf{x}_t\left(\mathbf{x}_{t-1}, \mathbf{u}\right)}{\partial \mathbf{u}} = \begin{pmatrix} \cos\left(\theta_{t-1}\right) \cdot \Delta t & 0 \\ \sin\left(\theta_{t-1}\right) \cdot \Delta t & 0 \\ \frac{\tan(\theta_{t-1}) \cdot \Delta t}{w_{wheelbase}} & \frac{v \cdot \Delta t}{w_{wheelbase} \cdot \cos^2(\theta_{t-1})} \end{pmatrix} \tag{3}$$

The error matrix $M\left(u, \alpha\right)$ considers the effect of motion errors, while the parameter α considers their severity.

$$M = \begin{pmatrix} \alpha_1 v^2 + \alpha_2 \varphi^2 & 0 \\ 0 & \alpha_3 v^2 + \alpha_4 \varphi^2 \end{pmatrix} \tag{4}$$

The covariance matrix P_t contains the uncertainty of the pose.

$$P_t = G \cdot P_{t-1} \cdot G^T + V \cdot M \cdot V^T \tag{5}$$

The first term of calculation 5 represents the pose prediction and the second term the uncertainty in the accuracy of the motion. The covariance can be visualized by plotting the ellipse defined by the eigen-vectors and the eigen-values of this matrix. Without any correction, the covariance ellipse will grow whenever the robot moves. The growth rate of this ellipse is defined by α which depends on the robot and its environment.

(a) (b)

Figure 3: The (a) real robot and its (b) simulation.

5. Real Car

A Tamiya RC-race car in the scale $1{:}10$ is used as base frame for the ackermann robot, see Figure 3a. The vehicle is powered by a BLDC (Brushless Direct Current) motor, and a servo motor is used for steering. Since the position can be derived from internal hall sensors within the BLDC motor there is no need for additional encoders.
An Arduino Uno microcontroller is used because of its real time capability and its special hardware for such low level actors and sensors. Serial messages from the interface are the means of communication between the Arduino Uno and the Raspberry Pi. In this project, Raspbian is the operating system for the Raspberry Pi because it is based on Debian, which supports ROS. A W-LAN stick is installed on the Raspberry Pi to grant access from other workstations.
The Sensor Level CPU which is represented by the Arduino Uno is responsible for controlling the car, reading sensors and presenting the data in a useful way. A motion controller [5] is implemented for the BLDC motor. Three signals similar to sinus waves generated with pulse width modulation on the Arduino Uno are applied to the motor. The calculation of the pose and its covariance also takes place on the Arduino Uno based on the velocity motion model mentioned before. The calculation frequency is about $100Hz$ which results in an update rate of $0.01s$. The controlling structure of the vehicle is shown in Figure 4.

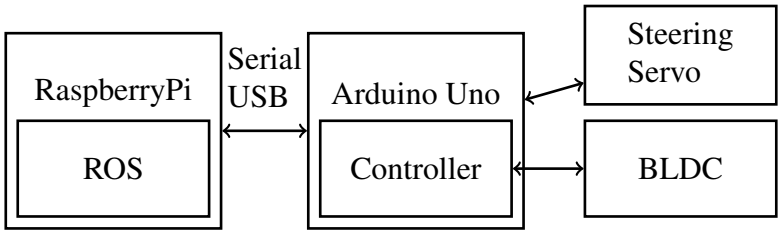


Figure 4: The control hierarchy from ROS to the cars actuators.

Since the steering appears to be the primary source of uncertainty, two improvements are considered. The first is to replace the unsteady steering with a more stable one. The second is to upgrade the car with an encoder for the steering.

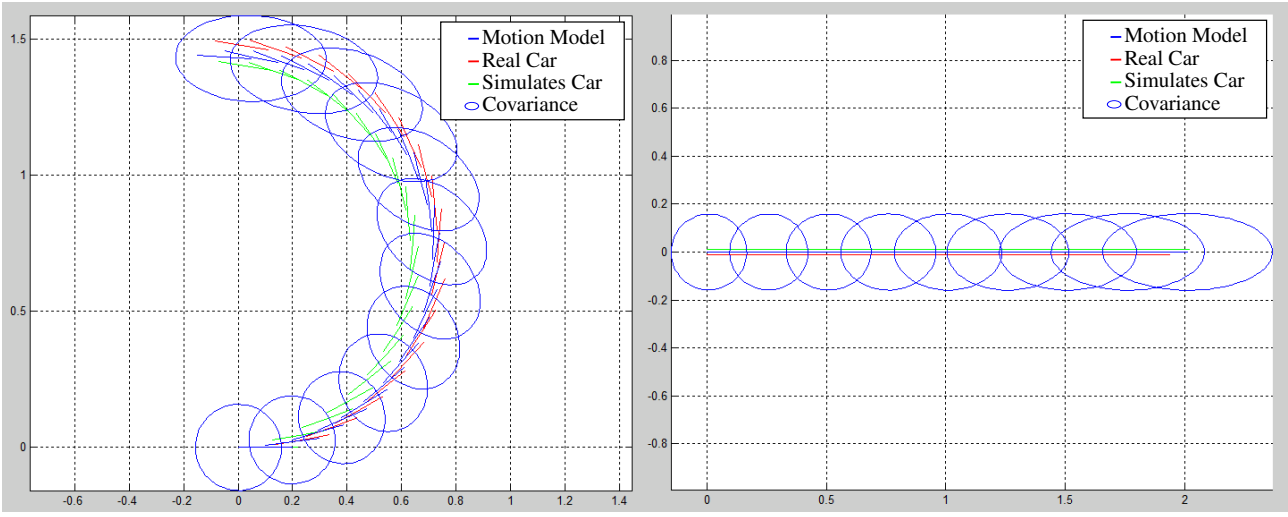
6. Simulated Car

Validation of systems and algorithms is an important task in mobile robotics. Thus, Gazebo is used for visualisation and physical simulation of the robot. The simulation contains the parts which are vital for the robots motion. They are imported to Gazebo with a URDF (Unified Robot Description Format) file, see Figure 3b. In the first attempt to simulate the ackermann drive robot, a link was created for each part of the steering and they were connected with joints. The parent-child structure of joints in URDF makes it impossible to create such a closed loop, so a workaround was needed. To get an ackermann steering like behavior, a ROS plug-in is used to control the kingpins, see Figure 2. The plug-in calculates the angles for both front wheels and adjusts the kingpins accordingly. For these calculations, the knowledge of the wheelbase and the track is required. The curve radius of the imaginary third front wheel has to be calculated. It has to be considered that the radii of the left and the right front wheel differ by a half track width from the previously calculated radius. Based on this, the steering angles φ_L and φ_R can be calculated, using the trigonometric functions.

To avoid unintended movements of the kingpin joints, the Gazebos real time update has to be $2000Hz$ and the maximum step size $0.0005s$.
The following three improvements would increase the accuracy of the simulation. Firstly, detailed measurements should be taken to replace the wheels approximated friction parameter. Secondly, damping should be added to the vehicle. Finally, the front wheels should be powered and equipped with a differential.

7. Results

Two tests are carried out to quantify the accuracy of the real and the simulated robots motion. For the first test, a semicircle with the maximum steering angle and a velocity of $0.1m/s$ was driven. The low speed used during the test allows for errors stemming from wheel slipping and centrifugal force to be ignored. The motion model represents the motion commands in this test, so it can be used as a reference. The radius of the semicircle driven by the real car is $5cm$ bigger than the reference. This is caused by the unsteady steering of the RC-race car. The simulated car drives a trajectory differing from a circle. During the whole test, the positions of the real and the simulated car are covered by the covariance ellipse. In Figure 5a the test results are shown.
For the second test, a straight line was driven with a velocity of $0.1m/s$, based on the motion models response. The real car stops $4.5cm$ before the reference, because of inaccuracies in the measurement of the wheel size. The simulated car stops $1.8cm$ behind the reference. The reason for this deviation is that unlike Gazebo, the motion model does not regard the kinetic energy of the vehicle. Again, the covariance ellipse covers the position of the real and the simulated vehicle. The test results are shown in Figure 5b.
To increase the accuracy of the motion model, two improvements can be made. Firstly, the number of updates can be increased to downsize the time steps. Secondly, the kinetic energy of the vehicle should be considered by the velocity motion model.



(a) (b)

Figure 5: Comparison of trajectory and visualisation of the (a) rotational and (b) the straight behavior of the covariance ellipse.

8. Conclusion

This paper presents the creation of an ackermann robot, its simulation and their common interface. Furthermore, the implementation of a velocity motion model for these vehicles was explained. For future work, the created software will be allocated to the robotics community. Adding sensors like an IMU (Internal Measurement Unit) to the vehicle to increase the accuracy of the motion tracking is planned. Therefore, the interface needs to be extended to handle the new data input. This platform will be expanded by adding self localisation, thus sensor input is required. Based on the knowledge gained with the robot, further vehicles will be built.

References

[1] S. A. Beiker. Einführungsszenarien für höhergradig automatisierte Straßenfahrzeuge. In *Autonomes Fahren*, pages 197–217, 2015.

[2] M. Quigley et al. Ros: an open-source robot operating system. In *IRCA Workshop on Open Source Software*, 2009.

[3] S. Thrun et al. Stanley: The robot that won the darpa grand challenge. *Journal of Robotic Systems - Special Issue on the DARPA Grand Challenge*, 23(9):661–692, 2006.

[4] S. Thrun W. Burgard D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[5] A. Kiruthika R. Agasthiya T. Ramesh. Speed control of a sensored brushless dc motor using flc. *International Journal of Engineering Research & Technology (IJERT)*, 3(4):1–4, 2014.

[6] P. Ross. Robot, you can drive my car. *IEEE Spectrum*, 51(6):60–90, 2014.

[7] R. Siegwart I. R. Nourbakhsh D. Scaramuzza. *Introduction to Autonomous Mobile Robots*. MIT Press, 2011.

[8] P. Castillo-Pizarro T. V. Arredondo M. Torres-Torriti. Introductory survey to open-source mobile robot simulation software. In *Robotics Symposium and Intelligent Robotic Meeting (LARS), 2010 Latin American*, pages 150–155, 2010.

Trajectory planning based on activity recognition and identification of low-level process deviations

Sriniwas Chowdhary Maddukuri[1], Gerald Fritz[1], Sharath Chandra Akkaladevi[1], Matthias Plasch[1], and Andreas Pichler[1]

[1]Department of Robotics and Assistive Systems
Profactor GmbH
{Sriniwas,Maddukuri}@profactor.at

Abstract
Improving work efficiency and ensuring safety of the human worker while the human worker and robot simultaneously perform the tasks in close proximity is one of the key research topics in human-robot cooperation. Given a process which contains a set of tasks or process steps performed within the shared human-robot workspace, a methodology for the robot's trajectory planning will be mentioned in this concept paper. The methodology will be based on activity recognition and identification of low-level process deviations. Here, the low-level process deviations which occur from the robot assistant side are mainly focussed.

1. Introduction

A key requirement in the field of human-robot cooperation is to realize the process execution in a safe and time-efficient manner. Here, process refers to a list of process steps/tasks performed simultaneously by the human worker and robot assistant. To achieve safe execution of shared human-robot tasks, a process monitoring component which identifies low-level process deviations is a pre-requisite. In the context of shared human-robot tasks, deviations are often classified into robot assistant side deviations and human worker side deviations. Robot assistant side deviations are defined as unexpected events like unreachable goal configuration, grasp failure reported by the robot's tool and high probabilistic existence of collision-prone trajectories with the nearby static or dynamic objects while the robot performs an object manipulation task in the shared workspace. Human worker side deviations are defined as expected events like performing spatial sequence of actions or activities and unexpected transition between the tasks or process steps. Process deviations from the human worker side are not considered within this work. The motivation behind this research work is to come up with a trajectory planning framework which can identify and handle low-level process deviations with respect to the simultaneous recognition of human activities and process steps/tasks. In this research work, the handling of process deviations will also be mentioned.

1.1. Related Work

Recent work which deals with trajectory planning is based on prediction of human actions and activities to achieve spatio-temporal synchronization in shared human-robot tasks. The

manipulation planning framework presented in [3], [9], [16], [5], and [1] considered the trajectory planning problem from the normal operation of a manipulation task. A time-series classification algorithm was presented in [3] to perform the online prediction of human reaching motion by applying a motion capture camera system. Partial segments of actual motion variables are compared with the subset of motion variables which represent the optimally time aligned human motion demonstrations. In [9], the predicted motion trajectories are represented as 3D voxels which infers the workspace occupancy information. Similar approaches were adopted in [16], [5] for human motion prediction. In [6], human-object interactions in combination with human motion trajectories were used to build temporal conditional random fields for anticipating human activities. In [1], a human worker's intent was estimated by computing the probabilistic representation of workspace segmented areas to which the human is heading.

Task and motion planners were integrated in [13] and [4] to identify and handle low-level process deviations such as collision-prone trajectories with the neighbouring objects. Here, the process addressed is a pick and place operation performed by a robot on a cluttered table and a payload carried by two robots respectively. During the process execution, the interface layer in between the task and motion planners determines the presence/absence of obstructions by identifying the collision-prone trajectories from the trajectory planner as low-level process deviations. Based on these deviations, the task planner is updated with a new state and sends a variation of the initial task plan to the trajectory planner. An alternative way to handle these kinds of deviations is to replace object grasping with multiple push-grasps in a cluttered environment [10]. With our work we intend to enhance the state of the art by cascading activity recognition and task recognition to identify low-level process deviations and perform task level trajectory planning. In this work, we also intend to realize activity recognition by estimating the skeletal joint positions with a higher sampling rate.

1.2. Paper Organization

Section 2 deals with the methodology proposed for trajectory planning based on activity recognition and identification of low-level process deviations. Section 3 will present the experimental setup including a static process plan where the human worker and robot performs process steps/tasks within their shared workspace. Section 4 will detail the expected contributions.

2. Methodology

In this section, the methodology behind the trajectory planning based on activity recognition and identification of low-level process deviations will be described along with the system architecture. Figure 2 depicts the system architecture which consists of 7 major building blocks 1) Object tracking 2) Skeletal joints estimation 3) Action recognition 4) Activity recognition 5) Task recognition 6) Trajectory planner and 7) High-level planner. The algorithms applied for object tracking and action recognition components have already been realized and evaluated in [14] and [15] respectively and will not be mentioned in this research work. Therefore, the methods required for the remaining major blocks will be mentioned here.

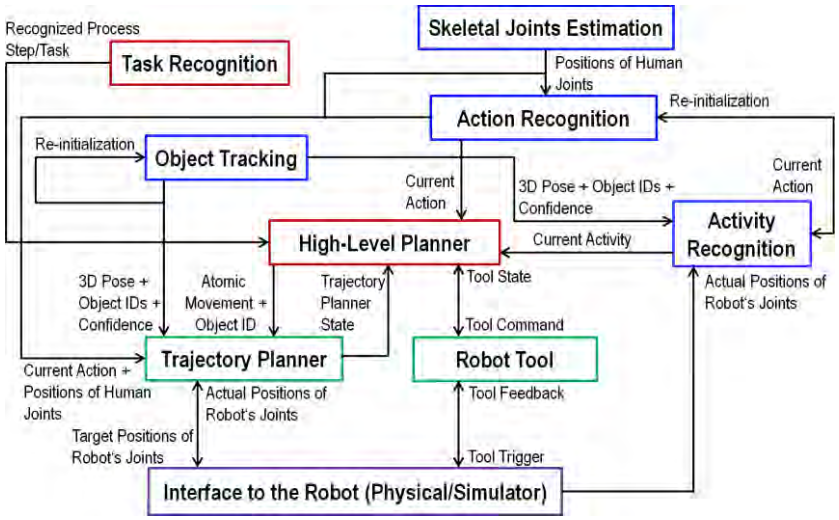


Figure 2: System architecture

2.1. Skeletal Joints Estimation

Estimation of skeletal joints is a crucial pre-requisite to overcome real-time data loss. The sampling rate of currently affordable RGB-D sensors is 30 fps. Recent works [3, Section 1.1], [9, Section 1.1] indicates that this sampling rate is not sufficient to recognize human activity in less than 1s. This leads to the motivation of estimating the skeletal joints data with a higher sampling rate. In the first stage, mathematical modelling of skeletal joints of left and right hands with respect to *Head*, *Neck* and *Spine Shoulder* skeletal joints will be performed in offline. In the second stage, the measured skeletal joints will be fed to a zero order hold (ZOH) component to provide the k^{th} sample at time instant $k*T_s$ with repeated values until the $k+1^{th}$ sample appears at time instant $(k+1)*T_s$. To overcome real-time data loss at time instant $k*T_s$, extrapolated values for skeletal joints of the left and right hands will be generated from the mathematical model. In the third stage, the samples with the higher sampling rate resulting from the ZOH and the extrapolated values resulting from the mathematical model will be used for estimating the desired skeletal joints positions. A forward Markov model describing the desired skeletal joints positions will be assumed and a stochastic subspace realization algorithm [8] will be applied to estimate the desired skeletal joint positions.

2.2. Activity and Task Recognition

Activity is defined as the sequence of actions or a single action performed by a human and his/her interactions with the objects of interest within an arbitrarily short time window. During the offline stage, probabilities of the recognized actions, human-object interactions and actual positions of robot's joints are considered as activity specific features and are collected with respect to M activity demonstrations by L individuals. Here, human-object interactions are represented by human motion trajectories and 3D position information, IDs and probability values of tracked objects. The recorded $M*L$ demonstrations are then fed to a classifier for activity classification. A Markov model will be adopted to represent the temporal relationship between human activities over time. During the online stage, partial segment of the activity specific features are used as inputs to compute the probability for states which represents human activities. The state with the highest probability will then be the recognized activity [12]. The activity recognition approach mentioned in this section will be extended for task recognition using a Hidden Markov model (HMM) to represent the process steps/task as its states. In the case of task recognition, the probability values of human

actions and his/her activities, robot's planned trajectories and positions of the robot's tool will be considered as task relevant features to model the states of the HMM [2].

2.3. Trajectory Planner

The trajectory planner considers the static workcell, actual skeletal joint positions, detected human activities and 3D locations of the objects of interest as an input and computes a collision-less trajectory for the robot. These activity dependent collisions-less trajectories will result in process-specific object manipulations like Grasp, Lift, Place, and Present. During the execution of the process, trajectory planner will send status updates about the object manipulations which will be requested by the high-level planner. Path planning algorithms which were applied in [11], [7] will be investigated to verify which one of them would be ideal for safe execution of the considered process.

2.4. High-Level Planner

High-Level Planner is an intermediate layer which receives the status updates continuously from major building blocks and robot's tool positions to monitor the process execution. The High-Level Planner will be included with the static description of sequential order of process steps/tasks involved within a process. During the execution of the process, the High-Level Planner will compare the actual state of the process with its desired state and identify the low-level process deviations from the robot assistant side. Based on these deviations, the trajectory planner will then compute a collision-free trajectory which will lead to successful completion of the previously failed process steps/tasks. Here, High-Level Planner will continuously send the same process step/task to the trajectory planner until the identified deviation vanishes.

3. Experimental Setup

The process of assembling a Steam cooker device using its individual objects is considered here. The individual objects of the steam cooker are present on the worktable as depicted below.

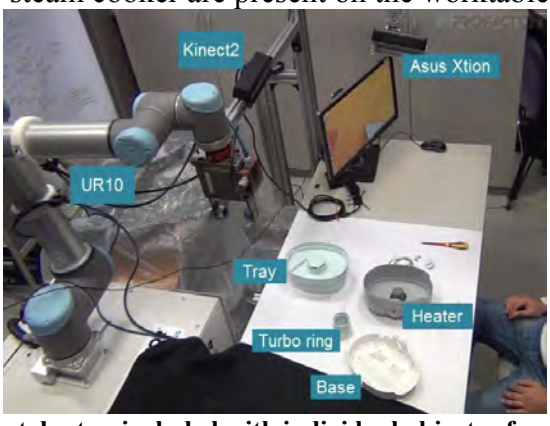


Figure 3: Experimental setup included with individual objects of a steam cooker device

In Figure 3, UR10 is the universal robot which is placed on a movable platform. This movable platform is clamped to the worktable where the human worker and ur10 robot will share the workspace. A Kinect v2 sensor is applied for the human action and human activity recognition and

an Asus Xtion sensor provides the scene data for the localization and tracking of objects of interest. The following static work plan related to the assembly process of a Steam cooker will be assumed.

- Step 1: Human worker picks the base object and robot grasps and lifts the heater object
- Step 2: Human worker holds the base object and robot shows the heater object to the human
- Step 3: Human worker attaches the base object to the heater object and inserts the timer cap on the side of heater object and performs the screwing

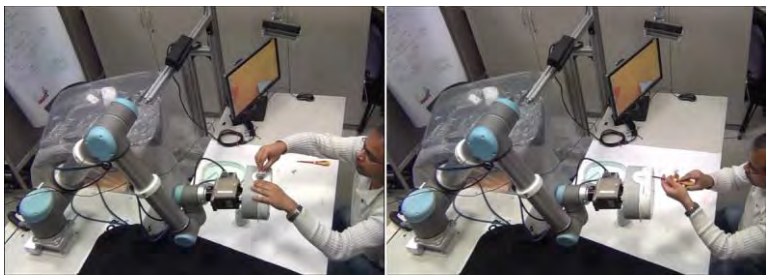


Figure 3.1.2: Human worker performing step 3

- Step 4: Robot lifts and places the compound object resulted from step 3
- Step 5: Human picks the turbo ring object and places it inside the compound object while the robot grasps and lifts the tray object
- Step 6: Robot presents and hands over the tray object to the human worker
- Step 7: Human worker inserts the tray object into the compound object resulted from step 5

Figure 3.1.3: Left image => step 6 and Right image =>step 7

4. Expected Contributions

The expected contributions resulting from this research work will be 1) Identification of low-level process deviations from the robot assistant side 2) task level trajectory planning based on simultaneous task and activity recognition to handle such process deviations 3) estimation of skeletal joints positions with a higher sampling rate.

5. Acknowledgements

This research is funded by the projects KoMoProd (Austrian Bundesministerium für Verkehr, Innovation und Technologie) and CompleteMe (FFG, 849441).

6. References

[1] Bascetta, L. and Ferretti, G. and Rocco, P. and Ardo, H. and Bruyninckx, H. and Demeester, E. and Di Lello, E., Towards safe human-robot interaction in robotic cells: An approach based on visual tracking and intention estimation, 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 2971–2978, Sept, 2011.

[2] Bastian Hartmann, Human Worker Activity Recognition in Industrial Environments, Published doctoral thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2011.

[3] Claudia Pérez D'Arpino and Shah, Julie A., Fast Target Prediction of Human Reaching Motion for Cooperative Human-Robot Manipulation Tasks Using Time Series Classification, 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015.

[4] E. Erdem, K. Haspalamutgil, C. Palaz, V. Patoglu and T. Uras, Combining high-level causal reasoning with low-level geometric reasoning and motion planning for robotic manipulation, 2011 IEEE International Conference on Robotics and Automation (ICRA), pp 4575–4581, 2011.

[5] Hawkins, K. P., Bansal, S., Vo, N. N., and Bobick, A. F, Anticipating Human Actions for Collaboration in the Presence of Task and Sensor Uncertainty, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp 2215–2222, May 31 – June 7, 2014.

[6] H. S. Koppula and A. Saxena, Anticipating human activities using object affordances for reactive robot response, Intelligent Robots and Systems (IROS), in Proceedings of Robotics: Science and Systems, 2012.

[7] Kalakrishnan, Mrinal and Chitta, S. and Theodorou, E. and Pastor, Peter and Schaal, S., STOMP: Stochastic trajectory optimization for motion planning, 2011 IEEE International Conference on Robotics and Automation (ICRA), pp 4569–4574, May, 2011.

[8] Katayama, T., Subspace methods for system identification, Springer Science \& Business Media, 2006.

[9] Mainprice, J. and Berenson, D., Human-robot collaborative manipulation planning using early prediction of human motion, 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 299–306, Nov, 2013.

[10] Mehmet Dogar and Siddhartha Srinivasa, A Framework for Push-Grasping in Clutter, Proceedings of Robotics: Science and Systems, June, 2011.

[11] Narayanan, V., Phillips, M., and Likhachev. M., Anytime Safe Interval Path Planning for dynamic environments, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 4708–4715, Oct, 2012.

[12] Roitberg, A., Perzylo, A., Somani, N., Giuliani, M., Rickert, M., and Knoll, A., Human activity recognition in the context of industrial human-robot interaction, 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Dec 1–10, 2014.

[13] S. Srivastava and E. Fang and L. Riano and R. Chitnis and S. Russell and P. Abbeel, Combined task and motion planning through an extensible planner-independent interface layer, 2014 IEEE International Conference on Robotics and Automation (ICRA), pp 639–646, 2014.

[14] Sharath Akkaladevi, Martin Ankerl, Christoph Heindl and Andreas Pichler, Tracking multiple rigid symmetric and non-symmetric objects in real-time using depth data, 2016 IEEE International Conference on Robotics and Automation (ICRA), May 16–21, 2016.

[15] Sharath Chandra Akkaladevi and Christoph Heindl, Action Recognition for human Robot Interaction in Industrial Applications, 2015 IEEE international Conference on Computer Graphics, Vision and information Security (CGVIS), Nov 2–3, 2015.

[16] Tanaka, Y. and Kinugawa, J. and Sugahara, Y. and Kosuge, K., Motion planning with worker's trajectory prediction for assembly task partner robot, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 1525–1532, Oct, 2012.

WS 9: Robotic Arm

Design, Modeling and Control of an Experimental Redundantly Actuated Parallel Platform*

Kyrill Krajoski[1], Andreas Müller[1], Hubert Gattringer[1] and Matthias Jörgl[1,2]

[1] Institute of Robotics
Johannes Kepler University Linz, Austria
{*kyrill.krajoski, a.mueller, hubert.gattringer, matthias.joergl*}*@jku.at*

[2] Trotec GmbH, Austria

Abstract
Actuation redundancy is a means to improve the dexterity, accuracy and reliability of parallel manipulators (PKMs). Over the last decade, various novel designs and control concepts have been developed and implemented in functional prototypes. In spite this extensive research several fundamental issues still remain to be addressed. This requires test benches allowing for flexible and modular setup of PKM prototypes. Aiming at agile light-weight PKMs, such a test bed should in particular enable to replace rigid by elastic links, and to implement model-based robust control concepts.
Such an experimental test platform is presented in this paper. The PKM under investigation is a 2-DOF planar PKM redundantly actuated by three actuators. Its mechanical design and actuation concepts together with the control system are presented. The dynamical model is presented as basis for the non-linear control. Fully parallel manipulators are characterized by repetitive use of identical modules connecting the moving and fixed platform. Therefore emphasize is given to the submodeling concept, which allows seamless integration of different modules (rigid vs. flexible links). Initial results are reported for the 2-PKM when controlled by an augmented PD scheme.

1. Introduction

The main purpose of the presented research is to create a modular parallel manipulator with actuation redundancy as a test platform. The dynamics modeling is carried out by means of subsystem modeling, see [1], [3] for details. The key for flexible and quick manufacturing is rapid prototyping. The prototype has links with low mass and inertia and are 3D printed.
The modularity allows for two, three, or four arms, connecting the moving platform, and thus gives rise to actuation redundancy. A model-based control scheme is used. This is based on a non-linear dynamic model. In this paper a computationally efficient formulation is used in terms of minimal as well as redundant coordinates [2], [4], [8], [7]. These dynamic models are the basis for the inverse dynamics and later for the augmented PD controller. Because of actuation redundancy, the inverse dynamic can be extended by a null space term, wich does not affect the manipulator's motion. It admits to increase the preload to annihilate backlash or manipulate the endeffector (EE) stiffness [9], [5], [6]. Finally, simulation results of an augmented PD controller are presented and analyzed.

*This work has been supported by the Austrian COMET-K2 program of the Linz Center of Mechatronics (LCM), and was funded by the Austrian federal government and the federal state of Upper Austria.

2. Platform Structure

The construction of the redundantly actuated parallel platform (see Fig. 1) is quite simple. It is a planar mechanism with $\delta = 2$ degrees of freedom. Redundantly actuated means, that the platform has more actuators ($m = 3$) than degrees of freedom. Altogether there are $n = 6$ joints.
However, the main dimension of the experimental test platform is 610 x 610 x 170 mm and the distance between the motors is $400\,\mathrm{mm}$. Each link of an arm is $l = 200\,\mathrm{mm}$ long. The general

Figure 1. Platform with three arms (motor, active link, passive link) in different colors

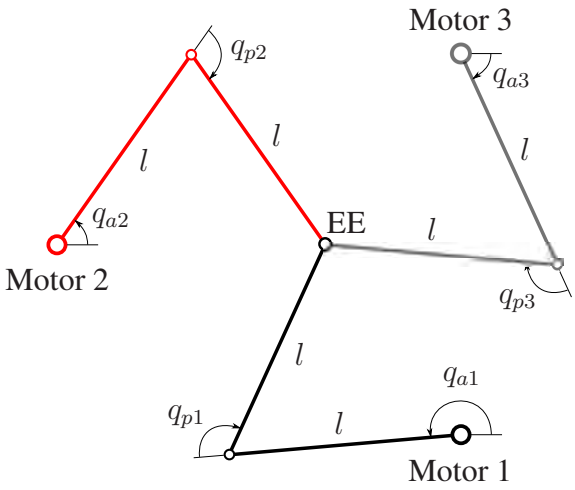


Figure 2. Joint coordinates of the redundant actuated parallel mechanism

purpose, the high modularity, can be noticed by many features of the platform. Base of this planar parallel manipulator is a four millimeter steel plate with twelve hole patterns. On these hole patterns motor sockets are centered and mounted. On the one hand, we are able to position the motor sockets on different location trying various motor constellations. On the other hand, a symmetrical disposal of two, three or four arms (each consisting of two links), is possible. The mounting concept, used for the platform, has the advantage to exchange arms (e.g. with flexible links instead of rigid ones) quickly. To distinguish arms from each other, they have got different colors (red, gray, black).
The arms are driven by brushless maxon motors EC-i 40 with a power of $100\,\mathrm{W}$.

3. Dynamic Modeling

The repetitive use of identical link combination (e.g. motor, active link, passive link) in parallel manipulators is a typical characterization. Therefore modeling of arms by means of subsystems is obvious.

3.1. Subsystem Modeling

The most important advantage of modeling a system by subsystems is the flexibility to add components such as additional actuated kinematic chains connecting the moving platform with the base platform. Furthermore it is easy to amend the model if in order to represent different phenomena such as elasticity of links, gear backlash or gear elasticities, which will be done in the near future.

The starting point is the Projection Equation of an entire arm as a kinematic chain

$$\sum_{b=1}^{N_j} \left[\left(\frac{\partial \mathbf{v}_s}{\partial \dot{\mathbf{q}}_j}\right)^T \quad \left(\frac{\partial \boldsymbol{\omega}_s}{\partial \dot{\mathbf{q}}_j}\right)^T \right]_b \left[\begin{array}{c} \dot{\mathbf{p}} + \widetilde{\boldsymbol{\omega}}_R \mathbf{p} - \mathbf{f}^e \\ \dot{\mathbf{L}} + \widetilde{\boldsymbol{\omega}}_R \mathbf{L} - \mathbf{M}^e \end{array} \right]_b \tag{1}$$

with index $j = 1, 2, 3$ for each arm. N_j is the number of bodies and $\dot{\mathbf{q}}_j = \begin{pmatrix} \dot{q}_{p,j} & \dot{q}_{a,j} \end{pmatrix}^T$ is describing velocity of each subsystem. Furthermore, $\mathbf{v}_s, \boldsymbol{\omega}_s$ are the absolute velocities of the center of gravity (CoG), $\boldsymbol{\omega}_R$ is the angular velocity of a chosen reference frame, $\mathbf{p}, \mathbf{L}$ are the linear and angular momenta, respectively, while $\mathbf{f}^e, \mathbf{M}^e$ are the applied forces of each body. Equation 1 leads to the motion equation of each arm modeled as a subsystem

$$\mathbf{M}_j \ddot{\mathbf{q}}_j + \mathbf{C}_j \dot{\mathbf{q}}_j - \mathbf{Q}_j = \mathbf{u}_j. \tag{2}$$

$\mathbf{M}_j$ is the mass matrix, $\mathbf{C}_j$ is the Coriolis and Centrifugal matrix, $\mathbf{Q}_j$ are the remaining forces and $\mathbf{u}_j = \begin{pmatrix} 0 & M_j \end{pmatrix}^T$ with the motor torque M_j describes the control forces of each arm. Furthermore, the equations of each arm (Eq. 1) can be assembled to the motion equation of the unconstrained system

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{Q}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{u}, \tag{3}$$

with $\mathbf{q}$ as the generalized coordinates written in an arbitrary sequence, f.e.

$$\mathbf{q} = \begin{pmatrix} q_{p,1} & q_{p,2} & q_{p,3} & q_{a,1} & q_{a,2} & q_{a,3} \end{pmatrix}^T. \tag{4}$$

Moreover $\mathbf{M}$ is the mass matrix , $\mathbf{C}$ is Coriolis and Centrifugal matrix , $\mathbf{Q}$ are the remaining and

$$\mathbf{u} = \begin{pmatrix} \mathbf{0} \\ \mathbf{c} \end{pmatrix}, \qquad \mathbf{u} \in \mathbb{R}^n, \qquad \mathbf{c} \in \mathbb{R}^m, \qquad \mathbf{c} = \begin{pmatrix} M_1 & M_2 & M_3 \end{pmatrix}^T. \tag{5}$$

are the control forces. Vector $\mathbf{c}$ contains the three motor torques.
Detailed calculations about dynamical modeling of subsystems can be found in [1], [3].

3.2. Subsystem Constraints

As described in the section before, the arms are modeled by means of subsystem modeling. Afterwards, these motion equations are assembled to an entire unconstrained system. Note that the sequence of joint coordinates $\mathbf{q}$ (Eq. 4) is arbitrary. In the unconstrained model the arms are not connected to the platform. Therefore, r geometric

$$\mathbf{h}(\mathbf{q}) = \mathbf{0}, \qquad \mathbf{h} \in \mathbb{R}^r \tag{6}$$

respectively kinematic constraints (with the Jacobian matrix $\mathbf{J}$)

$$\dot{\mathbf{h}}(\mathbf{q}) = \left(\frac{\partial \mathbf{h}}{\partial \mathbf{q}}\right)\dot{\mathbf{q}} = \mathbf{J}\dot{\mathbf{q}} = \mathbf{0}, \qquad \mathbf{J} \in \mathbb{R}^{r,n} \tag{7}$$

have to be built to connect them together. The geometrical constraints represents the linkage between the revolute joints and the EE. Thus, two independent loops, each with two independent constraints ($\Rightarrow r = 4$) can be located. Finally, after installing the constraint forces $\mathbf{J}^T(\mathbf{q})\boldsymbol{\lambda}$ into the motion equation of the unconstrained system, the entire model has a structure like

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{Q}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{J}^T(\mathbf{q})\boldsymbol{\lambda} = \mathbf{u} \tag{8}$$

$$\mathbf{J}\dot{\mathbf{q}} = \mathbf{0}. \tag{9}$$

Equation 8 is the Lagrangian motion equation of first kind.

3.3. Different Formulations of Motion Equations

Equations (8) and (9) are the point of departure, for many formulations. These formulations are necessary, because solving this system of equations (Eq. 8, 9), which is called a differential algebraic equation (DAE), is very complex. Moreover, it is not appropriate for the inverse dynamics. To reduce it to an ordinary differential equation (ODE), the constraint forces must be eliminated. This paper presents the minimal and redundant coordinates formulation [2], [4], [8], [7].

3.3.1. Minimal Coordinates Formulation

There are six independent joint angles, without the geometrical constraints. While introducing these four constraints, the number of independent angles will be reduced form six to two. Thus, the coordinates can be split in dependent $\mathbf{q}_d$ and independent $\mathbf{q}_i$ ones

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_d \\ \mathbf{q}_i \end{pmatrix}, \qquad \mathbf{q}_d \in \mathbb{R}^{n-\delta}, \quad \mathbf{q}_i \in \mathbb{R}^{\delta}. \tag{10}$$

Moreover, the kinematic constraints (Eq. 7) can be divided too, to express the dependent joint velocities explicitly

$$\mathbf{J}\dot{\mathbf{q}} = \mathbf{J}_{\mathbf{q}_d}\dot{\mathbf{q}}_d + \mathbf{J}_{\mathbf{q}_i}\dot{\mathbf{q}}_i = \mathbf{0}, \qquad \dot{\mathbf{q}} = \mathbf{F}\dot{\mathbf{q}}_i, \qquad \mathbf{F} = \begin{pmatrix} -\mathbf{J}_{\mathbf{q}_d}^{-1}\mathbf{J}_{\mathbf{q}_i} \\ \mathbf{I}_{\delta} \end{pmatrix}, \qquad \mathbf{F} \in \mathbb{R}^{n,\delta} \tag{11}$$

with the identity matrix $\mathbf{I}$. Matrix $\mathbf{F}$ is therefore an orthogonal complement of the Jacobian matrix $\mathbf{J}$, i.e. the product of both vanishes identically ($\mathbf{JF} \equiv \mathbf{0}$). Since the constraint forces vanish with matrix $\mathbf{F}$, it is an appropriate projector and leads to the minimal coordinates formulation

$$\overline{\mathbf{M}}(\mathbf{q})\ddot{\mathbf{q}}_i + \overline{\mathbf{C}}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}_i + \overline{\mathbf{Q}}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{A}^T(\mathbf{q})\mathbf{c} \tag{12}$$

with

$$\mathbf{F} = \begin{pmatrix} \mathbf{P} \\ \mathbf{A} \end{pmatrix}, \qquad \mathbf{A} \in \mathbb{R}^{m,\delta}, \qquad \mathbf{P} \in \mathbb{R}^{n-m,\delta} \tag{13}$$

$$\overline{\mathbf{M}} := \mathbf{F}^T\mathbf{M}\mathbf{F}, \qquad \overline{\mathbf{C}} := \mathbf{F}^T(\mathbf{C}\mathbf{F} + \mathbf{M}\dot{\mathbf{F}}), \qquad \overline{\mathbf{Q}} := \mathbf{F}^T\mathbf{Q}. \tag{14}$$

This formulation consists of δ independent equations. A drawback is the selection of two independent, local appropriate coordinates. Therefore parametrization singularities can occur. A method to avoid this is to switch between motion equations with different independent coordinates selection [4].

3.3.2. Redundant Coordinates Formulation

The problem of the latter formulation (Eq. 12) are the parametrization singularities, due to the choice of independent coordinates. There are two possibilities to avoid this. The first way, the switching method, has been mentioned before. The other way is to use another formulation without any coordinates selection, by means of a null-space projector

$$\mathbf{N}_{\mathbf{J},\mathbf{M}} := \mathbf{I}_n - \mathbf{J}_{\mathbf{M}}^{+}\mathbf{J}, \qquad \mathbf{N}_{\mathbf{J},\mathbf{M}} \in \mathbb{R}_n^n \tag{15}$$

with the right pseudoinverse

$$\mathbf{J}_{\mathbf{M}}^{+} = \mathbf{M}^{-1}\mathbf{J}^T(\mathbf{J}\mathbf{M}^{-1}\mathbf{J}^T)^{-1}. \tag{16}$$

Since $\mathbf{J}\mathbf{N}_{\mathbf{J},\mathbf{M}} \equiv \mathbf{0}$, the transformation leads to the redundant coordinates formulation

$$\widetilde{\mathbf{M}}(\mathbf{q})\ddot{\mathbf{q}} + \widetilde{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\dot{\mathbf{q}} + \widetilde{\mathbf{Q}}(\mathbf{q},\dot{\mathbf{q}}) = \widetilde{\mathbf{A}}^T(\mathbf{q})\,\mathbf{c} \tag{17}$$

with

$$\mathbf{N}_{\mathbf{J},\mathbf{M}} = \begin{pmatrix} \widetilde{\mathbf{P}} \\ \widetilde{\mathbf{A}} \end{pmatrix}, \qquad \widetilde{\mathbf{A}} \in \mathbb{R}^{m,n}, \qquad \widetilde{\mathbf{P}} \in \mathbb{R}^{n-m,n} \tag{18}$$

$$\widetilde{\mathbf{M}} := \mathbf{N}_{\mathbf{J},\mathbf{M}}^T \mathbf{M} \mathbf{N}_{\mathbf{J},\mathbf{M}}, \qquad \widetilde{\mathbf{C}} := \mathbf{N}_{\mathbf{J},\mathbf{M}}^T (\mathbf{C}\mathbf{N}_{\mathbf{J},\mathbf{M}} + \mathbf{M}\dot{\mathbf{N}}_{\mathbf{J},\mathbf{M}}), \qquad \widetilde{\mathbf{Q}} := \mathbf{N}_{\mathbf{J},\mathbf{M}}^T \mathbf{Q}. \tag{19}$$

Unlike before, this formulation consists of n equations, where δ ones are independent.

4. Model-Based Control with an Augmented PD-Controller

Model-based control is very important for parallel mechanisms with actuation redundancy, because of the antagonistic forces. As the name, redundant actuation, implies, there are more driving forces than degrees of freedom $m > \delta_{loc}$ to control the mechanism.
However, with this feature it is possible to increase the internal preload and thus, e.g. to annihilate backlash due to manufacturing or manipulate the EE stiffness [9], [5], [6].
The generalized force of an augmented PD Controller consists of three parts. The first part is a feed forward term calculated with the inverse dynamics, which releases the feedback controller. Thus, the joint angle error is much smaller. The second one is a feedback term, with weighted error position and velocity of the joint angles. And finally the third one has no dynamic effect i.e. it increases the internal forces. For further information, see [2], [4], [8], [7].

4.1. Inverse Dynamics

The inverse dynamics solution is the basis for an augmented PD or a computed torque controller.

4.1.1. Minimal Coordinates Formulation

The solution of the inverse dynamics is given by minimization of $(\mathbf{c}-\mathbf{c}^0)^T\mathbf{W}(\mathbf{c}-\mathbf{c}^0)$ as

$$\mathbf{c} = \underbrace{\left(\mathbf{A}^T(\mathbf{q})\right)_{\mathbf{W}}^{+} \left(\overline{\mathbf{M}}(\mathbf{q})\,\ddot{\mathbf{q}}_i + \overline{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}}_i + \overline{\mathbf{Q}}(\mathbf{q},\dot{\mathbf{q}})\right)}_{1} + \underbrace{\mathbf{N}_{\mathbf{A}^T,\mathbf{W}}(\mathbf{q})\,\mathbf{c}^0}_{3} \tag{20}$$

with the weighting matrix $\mathbf{W}$ and an arbitrary preload parameter vector $\mathbf{c}^0$. Furthermore, $(\mathbf{A}^T)_{\mathbf{W}}^{+} = \mathbf{W}^{-1}\mathbf{A}(\mathbf{A}^T\mathbf{W}^{-1}\mathbf{A})^{-1}$ is the right pseudoinverse and $\mathbf{N}_{\mathbf{A}^T,\mathbf{W}} = \mathbf{I}_m - (\mathbf{A}^T)_{\mathbf{W}}^{+}\mathbf{A}^T$ is a null space projector of matrix $\mathbf{A}^T$.

4.1.2. Redundant Coordinates Formulation

The number of equations of the redundant coordinates formulation is higher, than the number of free parameters $\mathbf{c} \in \mathbb{R}^m$ ($n < m$). Furthermore, there are δ independent equations, i.e. only δ columns of $\widetilde{\mathbf{A}}^T$ are linear independent. Therefore Eq. 17 must be rewritten as

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{A}}^T\mathbf{c} = \widetilde{\mathbf{A}}_1^T\mathbf{c}_1 + \widetilde{\mathbf{A}}_2^T\mathbf{c}_2, \qquad \mathbf{c}_1 \in \mathbb{R}^{\delta}, \quad \mathbf{c}_2 \in \mathbb{R}^{m-\delta}, \tag{21}$$

$$\mathbf{c}_1 = \left(\widetilde{\mathbf{A}}_1^T\right)^{+}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{A}}_2^T\mathbf{c}_2\right), \qquad \left(\widetilde{\mathbf{A}}_1^T\right)^{+} = \left(\widetilde{\mathbf{A}}_1\widetilde{\mathbf{A}}_1^T\right)^{-1}\widetilde{\mathbf{A}}_1, \tag{22}$$

with the modified optimization problem

$$\left\{ \begin{array}{l} \|\mathbf{c}\| = \|\mathbf{c}_1\| + \|\mathbf{c}_2\| \rightarrow \min \\ \mathbf{c}_1 = \left(\widetilde{\mathbf{A}}_1^T\right)^+ \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{A}}_2^T \mathbf{c}_2\right) \end{array} \right\}. \tag{23}$$

The solution structure is equivalent to

$$\mathbf{c} = \underbrace{\left(\widetilde{\mathbf{A}}^T(\mathbf{q})\right)^+ \left(\widetilde{\mathbf{M}}(\mathbf{q})\,\ddot{\mathbf{q}} + \widetilde{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}} + \widetilde{\mathbf{Q}}(\mathbf{q},\dot{\mathbf{q}})\right)}_{1} + \underbrace{\mathbf{N}_{\widetilde{\mathbf{A}}^T}(\mathbf{q})\,\mathbf{c}^0}_{3}, \tag{24}$$

with $\mathbf{N}_{\widetilde{\mathbf{A}}^T} = \mathbf{I}_m - (\widetilde{\mathbf{A}}^T)^+ \widetilde{\mathbf{A}}^T$, but unlike before

$$\left(\widetilde{\mathbf{A}}^T\right)^+ = \left(\begin{array}{c} \left(\widetilde{\mathbf{A}}_1^T\right)^+ \left(\mathbf{I}_n - \widetilde{\mathbf{A}}_2^T \left(\mathbf{I}_{m-\delta} + \mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T \left(\widetilde{\mathbf{A}}_1^T\right)^+\right) \\ \left(\mathbf{I}_{m-\delta} + \mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}^T \left(\widetilde{\mathbf{A}}_1^T\right)^+ \end{array} \right), \qquad \mathbf{B} = \left(\widetilde{\mathbf{A}}_1^T\right)^+ \widetilde{\mathbf{A}}_2^T \tag{25}$$

is not the right pseudoinverse.

4.2. Augmented PD Controller

The solution of the inverse dynamics is only a control scheme without the feedback term by weighted joint errors. Therefore, such a term has to be added to Eq. 20

$$\mathbf{c} = \underbrace{\left(\mathbf{A}^T(\mathbf{q})\right)_{\mathbf{W}}^+ \left(\overline{\mathbf{M}}(\mathbf{q})\,\ddot{\mathbf{q}}_i^d + \overline{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}}_i^d + \overline{\mathbf{Q}}(\mathbf{q},\dot{\mathbf{q}})\right)}_{1} - \underbrace{\left(\mathbf{A}^T(\mathbf{q})\right)_{\mathbf{W}}^+ \left(\overline{\mathbf{K}}_P \mathbf{e}_i + \overline{\mathbf{K}}_D \dot{\mathbf{e}}_i\right)}_{2} + \underbrace{\mathbf{N}_{\mathbf{A}^T,\mathbf{W}}(\mathbf{q})\,\mathbf{c}^0}_{3}, \tag{26}$$

for the control torques in minimal formulation and

$$\mathbf{c} = \underbrace{\left(\widetilde{\mathbf{A}}^T(\mathbf{q})\right)^+ \left(\widetilde{\mathbf{M}}(\mathbf{q})\,\ddot{\mathbf{q}}^d + \widetilde{\mathbf{C}}(\mathbf{q},\dot{\mathbf{q}})\,\dot{\mathbf{q}}^d + \widetilde{\mathbf{Q}}(\mathbf{q},\dot{\mathbf{q}})\right)}_{1} - \underbrace{\left(\widetilde{\mathbf{A}}^T(\mathbf{q})\right)^+ \left(\widetilde{\mathbf{K}}_P \mathbf{e} + \widetilde{\mathbf{K}}_D \dot{\mathbf{e}}\right)}_{2} + \underbrace{\mathbf{N}_{\widetilde{\mathbf{A}}^T}(\mathbf{q})\,\mathbf{c}^0}_{3}. \tag{27}$$

for the redundant formulation (Eq. 24). The superscript d indicates the desired values. The variables $\mathbf{e}_i = \mathbf{q}_i - \mathbf{q}_i^d$, respectively $\mathbf{e} = \mathbf{q} - \mathbf{q}^d$ are the error coordinates and $\overline{\mathbf{K}}_P, \overline{\mathbf{K}}_D, \widetilde{\mathbf{K}}_P, \widetilde{\mathbf{K}}_D$ are the weighting matrices for the PD controller.

4.3. Simulation Results

The simulations are realized with an augmented PD controller in both formulations. Furthermore, quantization effects of encoders are implemented. The weighting matrices of the PD controller are chosen by two issues. Firstly, the joint error shall be as small as possible. And secondly, the torque of the motors shall not be too high and aggressive.
The path is a simple point to point motion with a few arbitrary points and the acceleration is realized by $\sin^2$ profiles.
Furthermore, the unknown dynamics parameters are found with a CAD program. Table 1 shows an overview. J is the mass inertia around the CoG and relevant axis, except the entry of the Motor/active link - combination. Instead it is along the motor axis. m is the mass and l_s is the distance between

Component	J in $\mathrm{kg\,m^2}$	m in kg	l_s in m
Motor/active link - combination	4.22×10^{-4}	–	–
Black passive link	3.97×10^{-4}	5.41×10^{-2}	1.034×10^{-1}
Red passive link	2.90×10^{-4}	4.48×10^{-2}	8.32×10^{-2}
Gray passive link	3.61×10^{-4}	5.04×10^{-2}	9.64×10^{-2}

Table 1. Overview of the dynamic parameters

the previous joint and the CoG of each component. A comparison of both formulations implies, that behavior of the joint error, while controlled with an augmented PD controller in minimal coordinates formulation, is at least $e_{i,max} = 0.014\,\mathrm{rad}$ (see Fig. 3). Additionally, there is a discontinuity in the motor torques at $t = 0.6\,\mathrm{s}$, while the selection of independent coordinates is changed. From the simulation point of view, the redundant coordinates formulation is a quite better choice (c.f. Fig. 4). There are no discontinuities in the motor torques and joint errors are quite smaller ($e_{i,max} = 0.005\,\mathrm{rad}$).

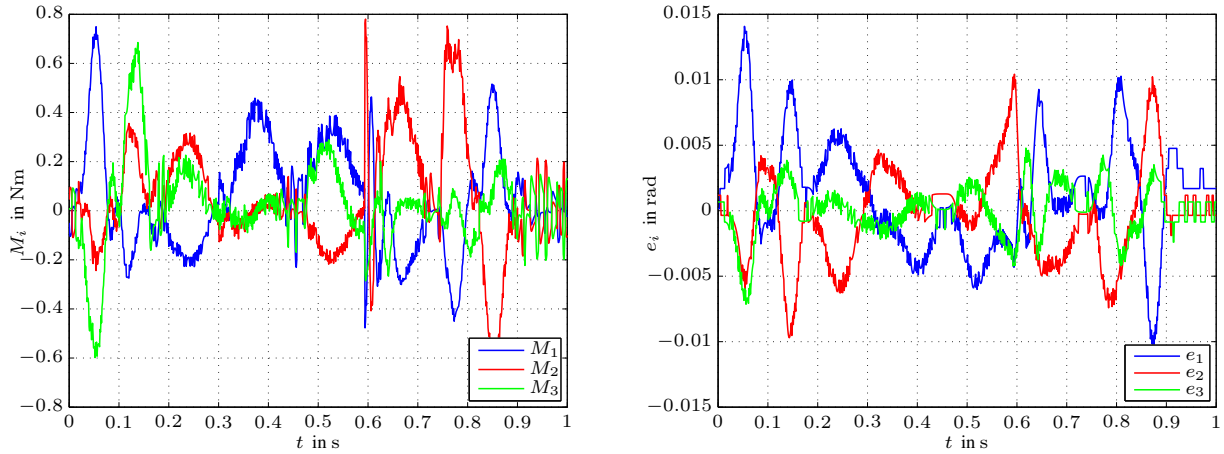


Figure 3. Simulation results with an augmented PD-controller in minimal coordinates formulation

5. Conclusion

In this paper a proposal for a experimental test platform with modular setup has been given. Because of the usual reuse of similar modules in parallel manipulators, a method to model entire systems with subsystems has been demonstrated. Different model formulations for designing a model-based controller have been given and simulation results with an augmented PD controller in both formulations have been presented.
Shortly, a geometric and dynamic calibration must be done and thus, the simulation results has to be validated with experimental results. Furthermore, in the near future a replacement of these arms with elastic ones is proposed.

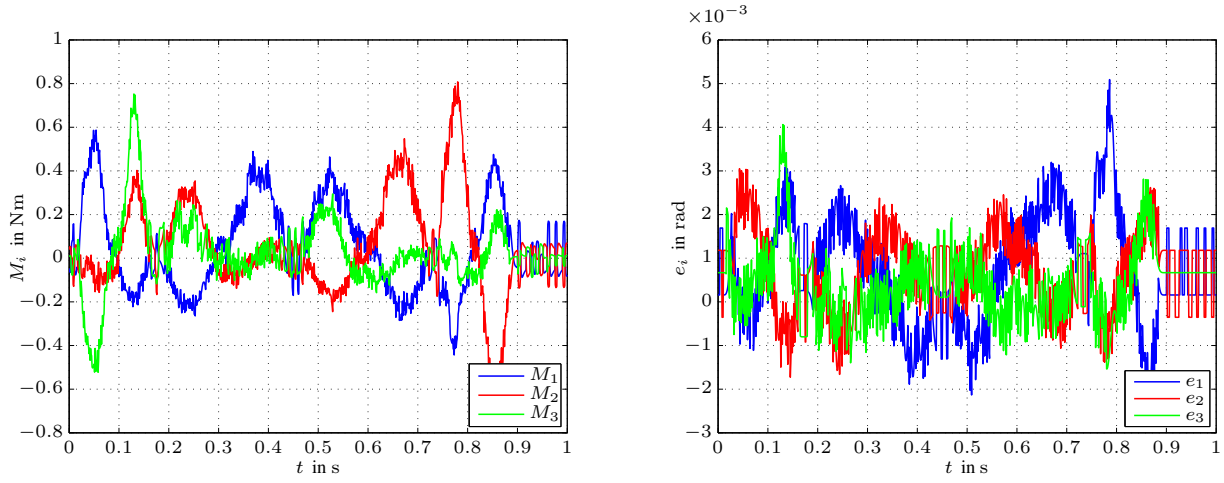


Figure 4. Simulation results with an augmented PD-controller in redundant coordinates formulation

References

[1] H. Bremer. *Elastic multibody dynamics: A direct Ritz approach*. Springer-Verlag, 2008.

[2] H. Cheng, Y. K. Yiu, and Z. Li. Dynamics and control of redundantly actuated parallel manipulators. *Mechatronics, IEEE/ASME Transactions on*, 8(4):483–491, 2003.

[3] H. Gattringer. *Starr-elastische Robotersysteme: Theorie und Anwendungen*. Springer-Verlag, 2011.

[4] T. Hufnagel and A. Müller. A realtime coordinate switching method for model-based control of PKM. In *Multibody Dynamics 2011, ECCOMAS Thematic Conf. Brussels, Belgium, 4-7 July*, 2011.

[5] A. Müller. Internal preload control of redundantly actuated parallel manipulators – its application to backlash avoiding control. *Robotics, IEEE Transactions on*, 21(4):668–677, 2005.

[6] A. Müller. Stiffness control of redundantly actuated parallel manipulators. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 1153–1158. IEEE, 2006.

[7] A. Müller. A robust inverse dynamics formulation for redundantly actuated PKM. In *13th World Congress in Mechanism and Machine Science, Guanajuato, Mexico*, pages 19–25, 2011.

[8] A. Müller and T. Hufnagel. Model-based control of redundantly actuated parallel manipulators in redundant coordinates. *Robotics and Autonomous Systems*, 60(4):563–571, 2012.

[9] B. J. Yi, R. A. Freeman, and D. Tesar. Open-loop stiffness control of overconstrained mechanisms/robotic linkage systems. In *Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on*, pages 1340–1345. IEEE, 1989.

Energy Optimal Control of an Industrial Robot by using the Adjoint Method

Thomas Lauss, Peter Leitner, Stefan Oberpeilsteiner, and Wolfgang Steiner

University of Applied Science Upper Austria
School of Engineering and Environmental Sciences
Stelzhamerstraße 23, 4600 Wels, Austria
{*thomas.lauss, wolfgang.steiner*}*@fh-wels.at*

Abstract
The main goal of this contribution is to determine the excitation of an industrial robot, such that the energy consumption becomes a minimum during the manipulation of the tool center point (TCP) from a start position to a given end point within a predefined time. Such tasks can be restated as optimization problems where the functional to be minimized consists of the endpoint error and a measure for the energy. The gradient of this functional can be calculated by solving a linear differential equation, called the adjoint system. On the one hand the minimum of the cost functional can be achieved by the method of steepest descent where a proper step size has to be found or on the other hand by a Quasi-Newton algorithm where the Hessian can be appreciated. The theory is applied to a six-axis robot and the identification leads to a reduction of 47% of the signal energy.

Keywords: *optimal control, multibody dynamics, adjoint system, optimization, calculus of variation.*

1. Introduction

In this contribution an approach to such inverse dynamical problems is presented. It starts from an optimal control formulation of the problem by introducing a cost functional which has to be minimized subject to a system of differential equations (c.f. [1, 2]). The gradient computation of the cost functional is based on the so called adjoint method. Due to better convergence a Quasi-Newton method is used instead of the simple gradient method. Therefore, the Hessian matrix is approximated by using the BFGS-algorithm.

The adjoint method is already used in a wide range of optimization problems in engineering sciences. Especially, in the field of multibody systems, the computation of the gradient of the cost function is often the bottleneck for computational efficiency and the adjoint method serves as the most efficient strategy in this case. The basic idea of the adjoint method is the introduction of additional *adjoint* variables determined by a set of adjoint differential equations from which the gradient can be computed straightforward. This main idea directly corresponds to the gradient technique for trajectory optimization pioneered by Bryson and Ho [3].

Various authors have utilized the adjoint method in the sensitivity analysis of multibody system, as e.g., [4, 5]. Bottasso et al. [6] presented a combined indirect approach of the adjoint method in multibody dynamics for solving inverse dynamics and trajectory optimization problems, also similar to the ideas presented in [7].

For a signal energy optimal manipulation of the robot a cost functional is introduced, which consists of the quadratic input signals in every time step and of a so-called Scrap-function which defines the end point deviation.

The identified movements were tested on a PUMA six axis robot. With the measured control variables the required energy was evaluated. Based on this test data a considerably energy reduction was detected.

2. Problem definition

At first, let us consider a nonlinear dynamical system

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{v} \\ \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{v}} &= \tilde{\boldsymbol{f}}(\boldsymbol{q}, \boldsymbol{v}, \boldsymbol{u}, t), \end{aligned} \tag{1}$$

where $\boldsymbol{q} \in \mathbb{R}^n$ is the vector of generalized coordinates and $\boldsymbol{v} \in \mathbb{R}^n$ is the vector of generalized velocities. In addition, $\boldsymbol{M}$ is the $n \times n$ mass matrix and $\tilde{\boldsymbol{f}} \in \mathbb{R}^n$ the force vector. The vector $\boldsymbol{u}$ indicates the control variables in an opened or enclosed region $\Gamma \subseteq \mathbb{R}^m$. By introducing the vector of state variables $\boldsymbol{x}^\mathsf{T} = (\boldsymbol{q}^\mathsf{T}\ \boldsymbol{v}^\mathsf{T})$ we may rewrite Equation (1) by

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0. \tag{2}$$

In general the force vector $\boldsymbol{f}$ is a continuous vector field which depends on the states $\boldsymbol{x}$, controls $\boldsymbol{u}$ and on time t. In robotics, the position and velocity of the tool center point (TCP) will be of particular interest instead of the joint angles and angular velocities. Hence, the system output $\boldsymbol{y} \in \mathbb{R}^l$ is given by

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}).$$

In order to meet a predefined end point we have to satisfy the boundary condition

$$\boldsymbol{g}(\boldsymbol{x}(t_f)) = \bar{\boldsymbol{y}}. \tag{3}$$

However, we substitute the boundary condition of Equation (3) by the optimal control problem

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \\ J &= \int_{t_0}^{t_f} h(\boldsymbol{x}, \boldsymbol{u}, t)\ \mathrm{d}t + S(t_f, \boldsymbol{x}(t_f)) \longrightarrow \text{Min.} \end{aligned} \tag{4}$$

where the integral describes the energy consumption and the *Scrap-function* S includes the end point error. If the closed region Γ is not empty the solution of the *optimal control* problem of Equation (4) leads to an energy optimal manipulation of the dynamical system of Equation (2).

3. Gradient computation

To determine the gradient of the cost functional (4) we first add zero terms to it:

$$J = \int_{t_0}^{t_f} h(\boldsymbol{x}, \boldsymbol{u}, t) + \boldsymbol{p}^\mathsf{T} \underbrace{[\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) - \dot{\boldsymbol{x}}]}_{=0 \text{ Eq. (2)}}\, \mathrm{d}t + S(t_f, \boldsymbol{x}(t_f)) \tag{5}$$

The Lagrange-multipliers $\boldsymbol{p}$ are denoted as adjoint variables and are arbitrary at this point. Integration by parts of the term $\int \boldsymbol{p}\dot{\boldsymbol{x}}\ \mathrm{d}t$ leads to

$$J = \int_{t_0}^{t_f} \left(H + \dot{\boldsymbol{p}}^\mathsf{T}\boldsymbol{x}\right) \mathrm{d}t + S(t_f, \boldsymbol{x}(t_f)) - \boldsymbol{p}^\mathsf{T}\boldsymbol{x}\Big|_{t_0}^{t_f}, \tag{6}$$

where the *Hamiltonian* $H(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{p}, t) = h(\boldsymbol{x}, \boldsymbol{u}, t) + \boldsymbol{p}^\mathsf{T}\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)$ is introduced. In order to find a minimum of the cost functional J with respect to $\boldsymbol{u}$ we consider the variation of J according to a small change $\delta\boldsymbol{u}$ which is given by

$$\delta J = \int_{t_0}^{t_f} \left[\left(H_{\boldsymbol{x}} + \dot{\boldsymbol{p}}^\mathsf{T}\right)\delta\boldsymbol{x} + H_{\boldsymbol{u}}\delta\boldsymbol{u}\right] \mathrm{d}t + \left[S_{\boldsymbol{x}}(t_f, \boldsymbol{x}(t_f)) - \boldsymbol{p}^\mathsf{T}(t_f)\right]\delta\boldsymbol{x}(t_f) + \boldsymbol{p}^\mathsf{T}(t_0)\delta\boldsymbol{x}(t_0). \tag{7}$$

Due to the fact that no variation of the states at $t = t_0$ is allowed, the term $\boldsymbol{p}(t_0)\delta\boldsymbol{x}(t_0)$ is zero. If the adjoint variables are defined, such that

$$\dot{\boldsymbol{p}}^\mathsf{T} = -H_{\boldsymbol{x}} \quad \text{and} \quad \boldsymbol{p}^\mathsf{T}(t_f) = S_{\boldsymbol{x}}(t_f, \boldsymbol{x}(t_f)), \tag{8}$$

the complex relations between $\delta\boldsymbol{x}$ and $\delta\boldsymbol{u}$ need not to be computed and the variation of J according to Equation (7) is reduced to

$$\delta J = \int_{t_0}^{t_f} H_{\boldsymbol{u}}\delta\boldsymbol{u}\ \mathrm{d}t. \tag{9}$$

Equation (8) is a linear and time-variant system of differential equations which have to be solved backwards in time starting at $t = t_f$. Hence, the largest possible increase of δJ is obtained, if $\delta\boldsymbol{u}(t)$ is chosen in the direction of $H_{\boldsymbol{u}}^\mathsf{T}$. For that reason $H_{\boldsymbol{u}}^\mathsf{T}$ may be considered as the gradient of the cost functional $J(\boldsymbol{u})$.

4. Numerical determination of the optimal control

Based on the adjoint gradient computation, outlined in the previous section, we may now search for a control $\boldsymbol{u}$ which minimizes the objective functional J. First of all, the method of steepest descent is described, where we always walk a certain distance along the negative gradient until we end up in a local minimum of J. Due to the costly line search step during every iteration and the slow convergence the gradient method is extended to a Quasi-Newton method. Therefore, we have to solve the problem of finding $\boldsymbol{u}$ such that the gradient becomes zero.

4.1. The Method of Steepest Descent

The method of steepest descent tries to find a minimum of a function or, subsequently, of a functional by walking always along the direction of its negative gradient. This concept has first been developed to optimal control problems by H.J. Kelley [8] and A.E. Bryson [9].

The gradient is already derived from the adjoint system which is shown in Section 3. Now we use $H_{\boldsymbol{u}}^\mathsf{T}$ and simply walk a short distance along the negative gradient of J. By reason of numerics the continuous functions are discretized. Hence, the cost functional reads

$$J(\boldsymbol{u}) \approx \hat{J}(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N) \tag{10}$$

where $\boldsymbol{u}_i = \boldsymbol{u}(t_i)$ and $t_1, \ldots, t_N$ is a sequence of consecutive time steps in the interval $[t_0, t_f]$. A variation of the controls $\boldsymbol{u}_i$ leads to a variation of the cost functional

$$\delta \hat{J} = \sum_{i=1}^{N} \frac{\partial \hat{J}}{\partial \boldsymbol{u}_i} \delta \boldsymbol{u}_i.$$

On the other hand, the variation $\delta \hat{J}$ can be expressed by Equation (9) which, after discretisation, results in

$$\delta \hat{J} = \sum_{i=1}^{N} H_{\boldsymbol{u},i} \Delta t_i \delta \boldsymbol{u}_i$$

where $H_{\boldsymbol{u},i}$ is the evaluation of $H_{\boldsymbol{u}}$ at $t = t_i$. Hence, the gradient of the discretised functional may be identified as

$$\frac{\partial \hat{J}}{\partial \boldsymbol{u}_i} = H_{\boldsymbol{u},i} \Delta t_i$$

in which $\Delta t_i = t_i - t_{i-1}$. For walking in the direction of the negative gradient a small number $\kappa > 0$ has to be chosen to get the increment

$$\delta \boldsymbol{u}_i = -\kappa H_{\boldsymbol{u},i}^{\mathsf{T}} \Delta t_i. \tag{11}$$

If κ is sufficiently small, the updated control $\boldsymbol{u}_i + \delta \boldsymbol{u}_i$ will always reduce the cost functional J. However, finding the number κ such that J is reduced may require several simulations of the system equations. For that purpose, the increments given by Equation (11) are considered as functions of κ. After solving the equations of motion with $\boldsymbol{u} + \delta \boldsymbol{u}$ as inputs also the objective function J becomes, ultimately, a function of κ. By means of a line search algorithm one may find a number κ in a predefined interval $[0, \kappa_{\max}]$ which minimizes J.

4.2. Application of a Quasi-Newton Method

It is well known that the convergence of the gradient method is rather slow, especially near the optimal solution. Hence, a Newton method provides an alternative approach to find the minimum of the cost functional J. The basic idea is the following one: If $\hat{\boldsymbol{u}} = (\boldsymbol{u}_1^{\mathsf{T}}, \boldsymbol{u}_2^{\mathsf{T}}, \ldots, \boldsymbol{u}_N^{\mathsf{T}})^{\mathsf{T}}$ is defined by a zero gradient, i.e. by the equations

$$\nabla \hat{J} = \left[\frac{\partial \hat{J}}{\partial \boldsymbol{u}_1}, \cdots, \frac{\partial \hat{J}}{\partial \boldsymbol{u}_N} \right]^{\mathsf{T}} = \boldsymbol{0}$$

which can be solved for $\hat{\boldsymbol{u}}$ by Newton's method. However, the Hessian $\boldsymbol{H}$ is required for that purpose. To avoid the full computation of $\boldsymbol{H}$, which would be extremely time consuming, several quasi-Newton methods have been developed. They all approximate the Hessian by using the gradients of successive Newton-iterations. For example, the Hessian can be estimated efficiently by the well known *Broyden-Fletcher-Goldfarb-Shanno* (BFGS)-Algorithm (c.f. [10]). Even its inverse can be efficiently obtained by applying the *Sherman-Morrison formula* (c.f. [11]).

We compute an approximation $\tilde{\boldsymbol{H}}^{-1}$ of the inverse of the Hessian from the BFGS-algorithm. Then, an increment $\delta \hat{\boldsymbol{u}}$ of the discretized control signal is given by

$$\begin{pmatrix} \delta \boldsymbol{u}_1 \\ \delta \boldsymbol{u}_2 \\ \vdots \\ \delta \boldsymbol{u}_N \end{pmatrix} = -\tilde{\boldsymbol{H}}^{-1} \nabla \hat{J} \tag{12}$$

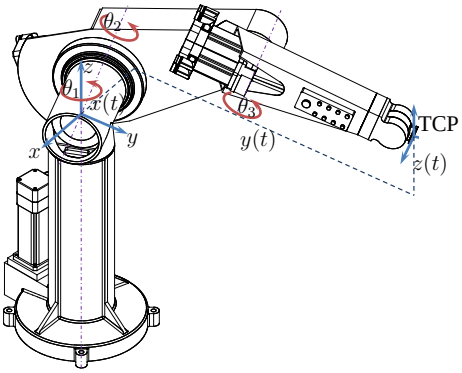


Figure 1. Schematics of the six-axis PUMA robot

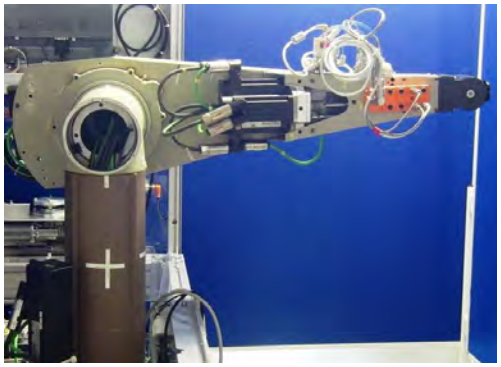


Figure 2. Image of the six-axis PUMA robot

Note, that it is strongly recommended to use a quasi-Newton method which directly approximates the inverse of the Hessian. Otherwise, if the original Hessian is computed, a very large and dense matrix must be inverted, since the number of components of J might become large.

The inverse of the Hessian after $k+1$ iterations is given by

$$\tilde{\boldsymbol{H}}_{k+1}^{-1} = \left(\boldsymbol{I} - \frac{\boldsymbol{p}_k \boldsymbol{q}_k^{\mathsf{T}}}{\boldsymbol{q}_k^{\mathsf{T}} \boldsymbol{p}_k}\right) \tilde{\boldsymbol{H}}_k^{-1} \left(\boldsymbol{I} - \frac{\boldsymbol{q}_k \boldsymbol{p}_k^{\mathsf{T}}}{\boldsymbol{q}_k^{\mathsf{T}} \boldsymbol{p}_k}\right) + \frac{\boldsymbol{p}_k \boldsymbol{p}_k^{\mathsf{T}}}{\boldsymbol{q}_k^{\mathsf{T}} \boldsymbol{p}_k} \tag{13}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{p}_k$ is the gradient direction of the k^{th}-iteration and $\boldsymbol{q}_k$ is the change of the gradient during the last iteration.

5. Application to the six-axis-robot

The presented method is used to minimize the signal energy consumption of the robot which is depicted in Figure 1. The reason why we have chosen this robot is that a lot of different parameters are available which are necessary for the evaluation and verification of the results. Afterwards, the simulation results are verified at a real six-axis-robot which is shown in Figure 2.

5.1. Problem definition

The system consists of three degrees of freedom, θ_1, θ_2 and θ_3 which denote the relative rotation angles of the joints. Due to the complicated structure of the equations of motion and the minor influence on the energy consumption the three wrist joints are fixed. First of all the equations of motion are derived and have the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)$ with the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and where $\boldsymbol{u} = [M_1, M_2, M_3]^{\mathsf{T}}$ contains the torques of the motors and $\boldsymbol{x} = [\theta_1, \theta_2, \theta_3, \dot{\theta}_1, \dot{\theta}_2, \dot{\theta}_3]^{\mathsf{T}}$ is the vector of states of the dynamical system. The system output $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ is a nonlinear function which depends on the states and describes the coordinates of the tool center point $\boldsymbol{y} = [x(t), y(t), z(t)]^{\mathsf{T}}$.

For the energy optimal manipulation of the robot from a start-point $\boldsymbol{x}_0$ to a given end-point $\overline{\boldsymbol{y}}$, $\dot{\overline{\boldsymbol{y}}}$ (c.f. Table 1) within a predefined time t_f we define the cost functional in the form

$$J = \underbrace{\int_{t_0}^{t_f} \boldsymbol{u}^{\mathsf{T}} \boldsymbol{u} \, \mathrm{d}t}_{\text{signal-energy}} + S(t_f, \boldsymbol{x}(t_f)). \tag{14}$$

Table 1. Start and end position of the robot

	start position	final position	start velocity	final velocity
θ_1	$0\,^\circ$	$-90\,^\circ$	$0\,\mathrm{rad/s}$	$0\,\mathrm{rad/s}$
θ_2	$0\,^\circ$	$-10\,^\circ$	$0\,\mathrm{rad/s}$	$0\,\mathrm{rad/s}$
θ_3	$0\,^\circ$	$45\,^\circ$	$0\,\mathrm{rad/s}$	$0\,\mathrm{rad/s}$
x_{TCP}	$-0.15320\,\mathrm{m}$	$0.81441\,\mathrm{m}$	$0\,\mathrm{m/s}$	$0\,\mathrm{m/s}$
y_{TCP}	$0.92112\,\mathrm{m}$	$-0.15320\,\mathrm{m}$	$0\,\mathrm{m/s}$	$0\,\mathrm{m/s}$
z_{TCP}	$0.02032\,\mathrm{m}$	$0.22233\,\mathrm{m}$	$0\,\mathrm{m/s}$	$0\,\mathrm{m/s}$

which contains the quadratical signal energy to be minimized. The scrap-function S of Equation (14) describes the endpoint error and is specified by

$$S(\boldsymbol{x},t) = \alpha \left\{ \beta \left[\boldsymbol{y}(\boldsymbol{x}) - \overline{\boldsymbol{y}} \right]^2 + \left[\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{q}} \dot{\boldsymbol{q}} - \dot{\overline{\boldsymbol{y}}} \right]^2 \right\} \tag{15}$$

where α and β are proper weighting factors and $\overline{\boldsymbol{y}}$, $\dot{\overline{\boldsymbol{y}}}$ contains the position and velocity of the endpoint in coordinates of the system output.

5.2. Results

The identification process of the signal energy optimal manipulation was started with the standard motion which is given from the robot controller. The results were verified on a real six-axis robot at the home institution. Hence, the data of the experiment and the simulation results are summarized in Figure 3. On the vertical axis the signal energy consumption is plotted over the time. It can be seen, that the standard manipulation wastes a lot of energy at the beginning and at the end of the motion due to the abrupt acceleration of the bodies. However, the signal energy optimal manipulation starts with a smooth movement of the heavy bodies. Therefore, the maximal speed of the axis have to be higher in comparison to the standard manipulation to reach the endpoint in the same period of time. As a result the reduction of the signal energy after the optimization process is about 47% with respect

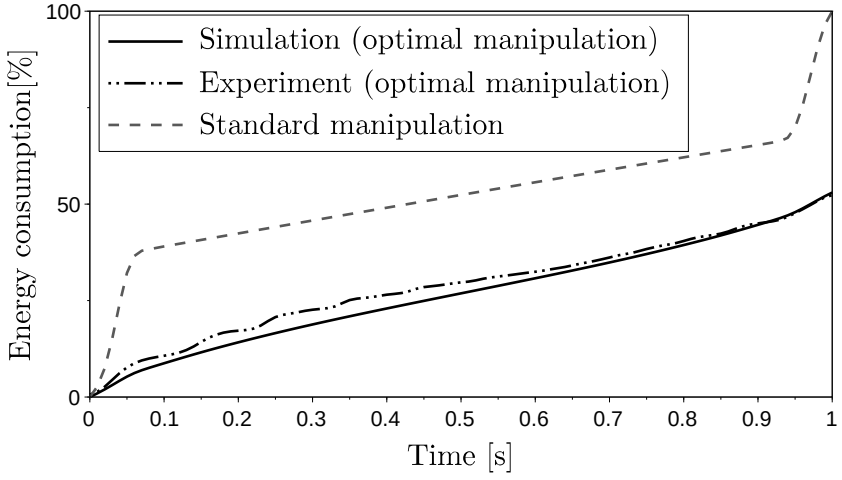


Figure 3. Build-up of the mechanical energy consumption

to the standard manipulation of the robot control.

In the upper part of Figure 4 the joint angles of the signal energy optimal manipulation in comparison to the standard manipulation of the robot are plotted over time. Obviously, the smooth characteristic of the optimal solution, which corresponds to the dashed line can be seen. However, the standard manipulation, which corresponds to the solid line, shows the commonly used standard motion calculated by the robot controller. In the lower part of Figure 4 the torques are depicted over the time. Here, the smooth characteristic of the optimized solution can be seen clearly.

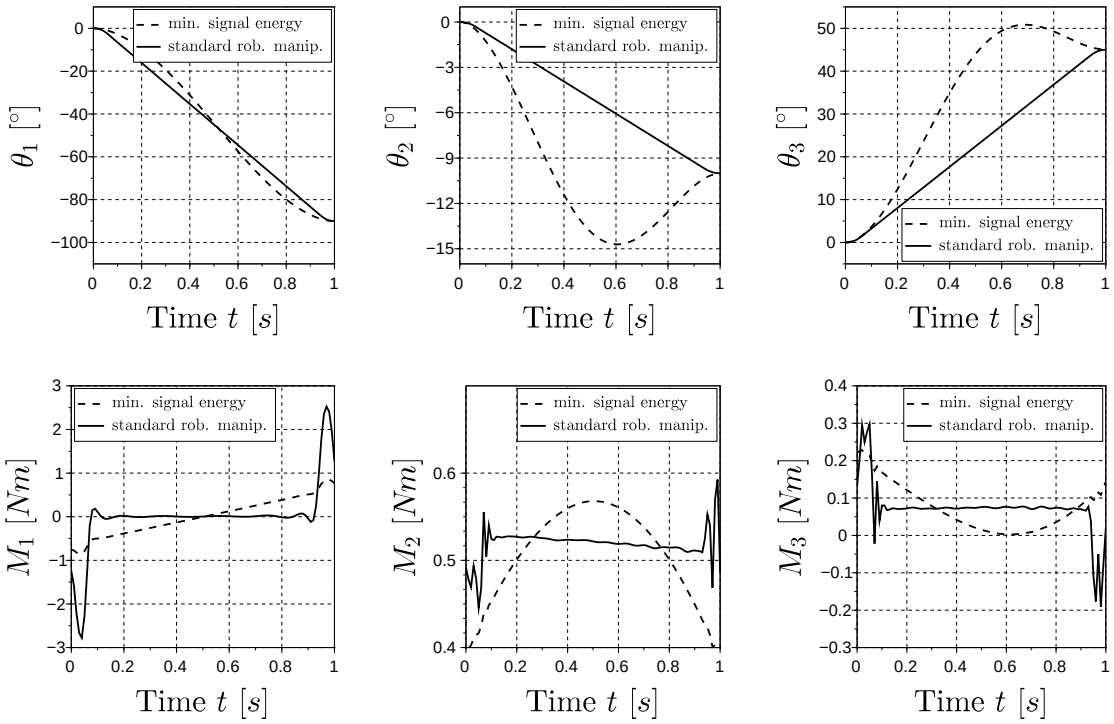


Figure 4. Trajectory of the states and torques in the axis

6. Conclusions and outlook

To reach a desired endpoint within a predefined time, the definition of a Scrap-function is required only. In addition, various requests to the system behavior can be considered in the integral part of the cost functional, such as the signal energy of an industrial robot.

This paper should reveal that the trajectory with minimal signal energy does not lead automatically to the mechanical energy optimal manipulation of the robot. Nevertheless, in practice such quadratic input terms are often used because this leads to less stress of the components. In simply terms you can say that the electrical parts are protected against overheating and the operation life span is increased additionally if the torques remain small and smooth over the manipulations.

For the results in Section 5.2. we neglected the three degrees of freedom of the wrist and fixed them to keep the equations of motion and the necessary matrices simple. However, if we consider this joint angles in the system equations it is possible to reach a predefined endpoint in different ways. This means that more than one final configuration of the robot exists which meet the end point in the coordinates of the tool center point.

Furthermore, the proposed identification can be done during operation. Instead of the forward simulation the measures of the previous manipulation can be used to solve the adjoint system and calculate the gradient. Hence, the defined cost functional, and therefore the signal energy, decreases during the manipulation of the robot. A big advantage is that it is not necessary to exchange any part of the robot, only an update of the robot control is required.

Acknowledgment

This project was supported by the program ”Regionale Wettbewerbsfähigkeit OÖ 2010-2013”, which is financed by the European Regional Development Fund and the government of Upper Austria.

References

[1] S. Reichl, W. Steiner. The Optimal Control Approach to Dynamical Inverse Problems. *Journal of Dynamic Systems, Measurement, and Control*. Vol. 134, Is. 2, 2012

[2] K. Nachbagauer, S. Oberpeilsteiner, K. Sherif, W. Steiner. The Use of the Adjoint Method for Solving Typical Optimization Problems in Multibody Dynamics. *Journal of Computational and Nonlinear Dynamics*. 2014

[3] Bryson, A., Ho, Y.: Applied Optimal Control. Hemisphere, Washington, DC (1975)

[4] Eberhard, P.: Adjoint Variable Method for Sensitivity Analysis of Multibody Systems Interpreted as a Continuous, Hybrid Form of Automatic Differentiation. In: Proc. of the 2nd Int. Workshop on Computational Differentiation, Santa Fe. Philadelphia, pp. 319–328 (1996)

[5] Haug, E., Wehage, R., Mani, N.: Design Sensitivity Analysis of Large-Scaled Constrained Dynamic Mechanical Systems. Journal of Mechanisms, Transmissions, and Automation in Design **106**(2), 156–162 (1984)

[6] Bottasso, C., Croce, A., Ghezzi, L., Faure, P.: On the Solution of Inverse Dynamics and Trajectory Optimization Problems for Multibody Systems. Multibody System Dynamics **11**(1), 1–22 (2004)

[7] Bertolazzi, E., Biral, F., Lio, M.D.: Symbolic-Numeric Indirect Method for Solving Optimal Control Problems for Large Multibody Systems. Multibody System Dynamics **13**, 233–252 (2005)

[8] H.J. Kelley. *Method of Gradients, Optimization techniques with applications to aerospace systems* Mathematics in Science and Engineering, Elsevier Science, 1952

[9] A.E. Bryson. Optimal Programming Problems with Inequality Constraints II: Solution by Steepest-Ascent *AIAA Journal*, (1964), 25-34.

[10] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, C.A. Sagastizábal. *Numerical Optimization - Theoretical and Practical Aspects.* Springer Berlin Heidelberg, 2006

[11] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.*, 21(1):124–127, 03 1950.

Design of an Industrial Robot with Six Degrees of Freedom for Educational Purposes

René Schweidler, Mohamed Aburaia and Corinna Engelhardt-Nowitzki

Department of Advanced Engineering Technologies
University of Applied Sciences Technikum Vienna, Austria
{*rene.schweidler, aburaia, corinna.engelhardt*}*@technikum-wien.at*

Abstract
In state of the art production and assembly lines industrial robots with six axes are widely used to manipulate production goods in all six degrees of freedom in space. Hence, mechatronics and robotics students have to achieve an in-depth comprehension regarding the configuration and adaptation of industrial robots from different manufacturers for applications such as welding, milling, assembling or the handling of components. However, these industrial robots typically cannot be disassembled to explore their internal structure and functionality due to, e.g., warranty reasons. Thus, educational facilities have to use auxiliary means, such as simulation in respective teaching units. To solve that problem, this paper describes the dimensioning and design of an industrial robot with six degrees of freedom for educational purposes, produced by the use of additive manufacturing techniques. Its main strengths are its low costs despite full functionality, its sound maintainability, and the fact that it can be disassembled multiple times by students in the course of, e.g., mechanics, electronics or software development projects. Besides, the proposed educational robot platform has been designed safe-to-use and aesthetically pleasing. Further mechanical structure optimization, the synthesis of the mathematical and kinematic model and control system configuration have to be done in future projects.

1. Introduction

Mechatronics and robotics students have to achieve an in-depth comprehension regarding the configuration and adaptation of industrial robots with six axes to manipulate production goods in all six degrees of freedom in space. Therefore, for educational purposes it is especially important to work with robots that may be disassembled in order to explore their internal structure and functionality. This normally rules out the use of commercial robots beyond manufacturer's designated robot functionalities – usually pure end effector choice, parameter configuration and programming. The use of auxiliary means like 3D-simulation programs is possible, but pedagogically disadvantageous, as the learning experience is impaired by the fictitiousness of virtual robot behavior [Tocháček et al., 2016]. Additionally, educational institutions will typically be subject to strict financial restrictions. Hence, they rely on either loans (i.e., industrial robots that cannot be disassembled) or low-price facilities, eventually self-constructed. Therefore the objective of this paper is to construct a robot platform that suits the aforementioned educational purposes and context requirements.

The technical challenge of this project was to create a robot with full and accessible functionality at a fraction of the costs of a commercial product by using additive manufacturing techniques. Critical attributes to be met in the course of the design process were the educational robot's ability to be

disassembled multiple times and to be maintained easily, with as little effort as possible. Besides, the proposed educational robot platform had to be designed safe-to-use and aesthetically pleasing.

Additionally this contribution pays respect to the fact that some authors even state, that the practice of introducing robotics into the academic process is still in an initial development stage (cp. e.g., [Ospennikova et al., 2015]) – at least in specific sectors: the majority of contributions within the current body of literature refers to school education below university level (e.g., [Eguchi, 2010]). Other major developments and respective projects and publications in the field of educational robotics are driven either by major industry players, i.e., robot manufacturers and similar companies (cp. e.g., [Yoo, 2015]) with the disadvantage that disassembly for a deeper understanding is prohibited. Another field of huge activities is the topic of robot competitions (cp. e.g., [Eguchi, 2016]), primarily focusing on robot performance optimization, but rarely on the teaching of advanced robot functional and structural principles at the level of robot engineering master courses within university education.

The remainder of the paper is as follows: section 2 provides a short overview on the field and explains general design principles with regard to the current endeavor. Subsequently sections 3-5 describe design details of the educational robot that has been developed in the course of this project. Finally section 6 draws a brief conclusion with regard to the achieved results and provides an outlook towards future activities.

2. Design principles for educational robotic experiences

The abilities of collegiate robotic and computational thinking and sufficient ways to facilitate the achievement of respective learning objectives within educational programs have been widely discussed in the literature (see e.g., [Miller et al., 2008], [Wing, 2008], [Eguchi, 2010], [Lee et al., 2011] or [Khanlari, 2013]). Although it is not the aim of this paper to provide an exhaustive literature overview, it can be said that the field of juvenile and undergraduate education is well elaborated in particular regarding elementary robot handling, control and programming. However the topic of advanced engineering and mechatronics education has to face further issues. As Alessandri and Paciaroni [Alessandri, 2012] conclude with reference to neuroscience (in particular, cp. [Varela et al., 1995]), an educational robotic experience has to allow for a shift from (more or less passive) observation of a device towards a deep immersion into the system in action. Transferred to the learning target of gaining an in-depth understanding not only from an industrial robots behavior and control, but as well from its functional principles and structures with regard to mechanics, electronics and software development, this leads to the conclusion that advanced robotics and mechatronics students must be provided with the opportunity to construct, simulate, assemble *and* disable a robotic system alternating with physical system-behavior experiments. Thus, before-after explorations could be done, e.g., after having improved mechanical components like a gripper or a joint, after having modified the electronic circuits, after having changed software code or parameters, or even after having totally disassembled and reassembled the whole robot for either maintenance, repair or experimental purposes.

Moreover, this practical education approach supports a further aspect that gains more and more importance in modern engineering disciplines, and especially in the field of mechatronics and robotics: teaching mechanical, electronic and informatics-related skills is a well-known issue. However, the interdisciplinary integration of these (and as needed also further) fields requires greater emphasis, and the same applies for system integration abilities [Gómez et al., 2014]. The educational robot, developed in the course of the current project was also designed for the purpose of strengthening

an integrative system engineering approach in theory (robot design and dimensioning) and practical application (robot programming, control and optimization).

Supporting factors to enable robotics learning experiences according to the aforementioned criteria are the ongoing price decline of progressively powerful sensors, motors and micro-controllers together with increasingly widespread additive manufacturing abilities in order to create adequate mechanical parts and actuator components. 3D-printing of synthetic materials has not only become financially affordable. Moreover, meanwhile even basic knowledge of production techniques like e.g., fused deposition modeling (FDM) allows for a rapid design, construction and fabrication of customized robot components with low mass (for further details of filament fabrication refer to e.g., [Allen, 2015]). There is, however, one possible disadvantage to be taken account, when using 3D-printed components: due to expectable inexactness of the printed parts, each robot prototype might have slightly differing attributes. As educational application will scarcely have the need of producing high quantities of identical machines, this can yet be considered as a minor constraint.

Altogether the mentioned developments have enabled the current project. Concretely, a six-axis robot was designed using CAD software that can handle payloads of up to $500g$ mass within a working envelope of $700mm$ in diameter. The belt driven joints were designed to hide all contained drive belts inside the robot in order to ensure safety and to achieve an aesthetically pleasing design. The goals of operator-friendliness and maintainability were obtained by assembling all components in enclosed modular sub-assemblies in order to be able to change each module easily or to adapt just one specific part of the robot. In contrast to commercial industrial robots, there is no need to strictly separate the working range of the robot from human reaching areas by means of a closed assembly cell. Figure 1 illustrates the naming conventions and the structure of the robot.

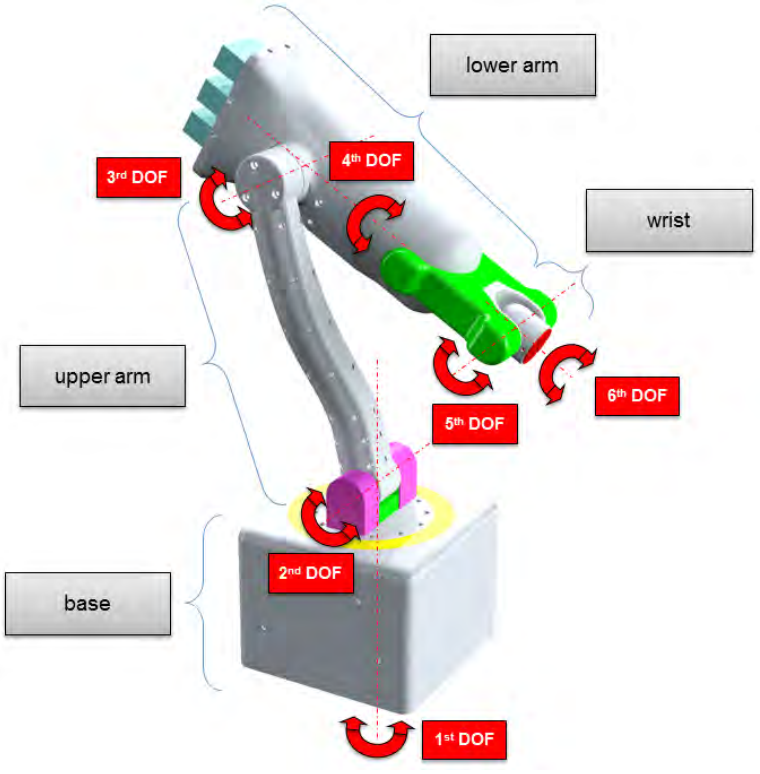


Figure 1. Naming convention and structure of the robot

In this paper the different assemblies are referred to as 'wrist', 'lower arm', 'upper arm'and 'base'. The following metaphor is applied to alleviate the comprehension of the system: The robot can be seen like a human arm. The 'shoulder joint'is mounted to the surface of the table and will be referred to as the 'base'. Next comes the 'upper arm', which connects the 'shoulder'to the 'elbow joint'. In this project the assemblies that are equal to the 'upper arm'and the 'elbow joint'are called 'upper arm'and 'third joint'. Finally, the 'lower arm assembly'is connected between the 'elbow joint'and the 'wrist'.

According to this rough concept, a six-axis robot was designed using CAD software that can handle payloads of up to $500g$ mass within a working envelope of $700mm$ in diameter. In order to provide a save to use platform the robot was designed to hide all moving parts such as belts, pulleys and shafts inside the robot. This at the same time allows for an likable design. In contrast to commercial industrial robots, there is no need to strictly separate the working range of the robot from human reaching areas by means of a closed assembly cell. The goals of operator-friendliness and maintainability were obtained by assembling all components in enclosed, modular assemblies in order to be able to change each module easily or to adapt just one specific part of the robot. A further important design requirement was to provide a platform that is decomposable multiple times without relevant part defects as a consequence of the dis- and re-assembly process. Even after multiple assembly loops, the robot must ensure a sufficient level of precision. This was achieved by introducing index pins in order to prevent from inaccurate re-assembly. Comparably, centering pins are used in order to be able to re-establish the exact coaxial position of every joint after reassembly. All parts were designed and optimized for the use of additive manufacturing techniques in order to enable the re-manufacturing of any part quickly, easily and cost-efficient. A further major design objective was to preferably use standard parts instead of manufacturing customized items. This helps to decrease manufacturing-time and -effort and at the same time makes use of the granted precision provided by supplier-dependent tolerances. In order to cut the maintenance effort to a minimum, only encapsulated bearings were used (no greasing or cleaning).

3. Design of the Base

Due to limitations regarding the maximum size of the 3D-manufactured objects, the base was split into two parts which were screwed together to provide a single solid base. Educational institutions that have access to more advanced equipment or are willing to deviate from the pure 3D-printing approach, could easily design their own robot concepts by means of using a one-piece manufactured base as an alternative. The base has two hollow chambers in order to hold all electronic components and the controller boards shielded and space-saving. As these openings contain all electronic and controller components, the robot can be used in stand alone mode as well as connected to a computer. A further compartment on the bottom of the base hides the drive belt of the first axis and its motor. This additional opening could as well be used to append extra weight (e.g., heavy steel plates) to prevent the base from moving while the robot is used in a stand-alone mode. Another possible use of the bottom compartment is to hold batteries, in case the robot shall be used in locations without electrical power supply, e.g., on fairs or exhibitions. Besides, the batteries take effect as additional weight. As shown in figure 2, the base consists of a cylindrical tube, also containing the motor-holder for the motors of the second and third axis. The first axis is moved by simply shifting the whole cylindrical tube together with the rest of the robot. The motors for the second and third axis are connected to the fork by means of two timing belts. The fork consists of a hollow shaft which is directly powered by the timing belt assigned to the motor of the second axis. The timing belt on the

other side of the hollow shaft drives a second pulley inside the shaft.

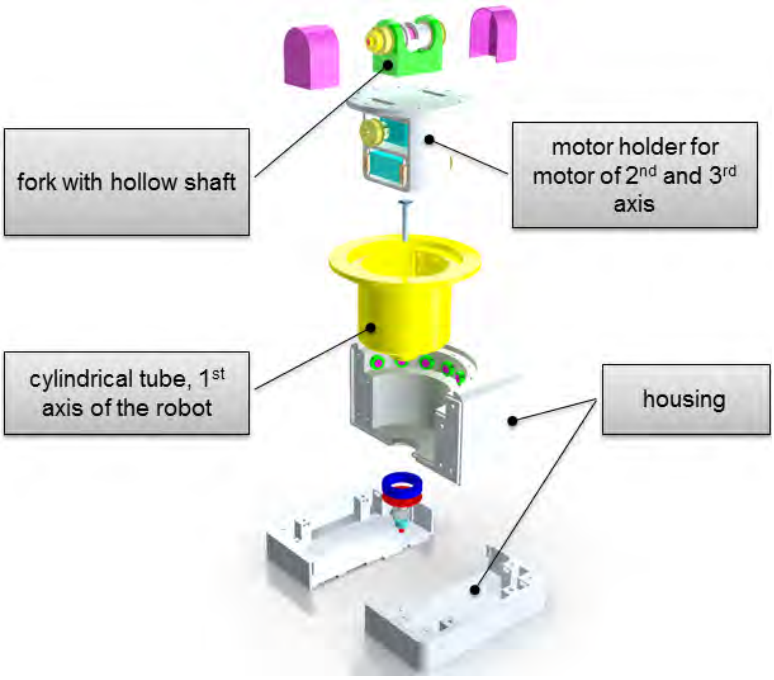


Figure 2. Exploded view of the base

4. Design of the Upper Arm

To reduce the weight of the arm and to maximize the manipulable payload, the motor of the third axis was relocated from the elbow joint into the base. To be able to transmit the torque of the third axis from the base to the elbow joint an arrangement of rods and cardan joints was used inside the upper arm assembly. The upper arm assembly is composed of two main housing parts. It consists of a timing wheel inside the upper arm assembly, which in turn is connected to the timing wheel inside the hollow fork shaft. Finally the torque is transmitted by a bevel gear to the rod assembly and further to the lower arm assembly.

5. Design of the Lower Arm

Another huge challenge was the design of the lower arm and the TCP-gearhead due to heavy restrictions regarding size and weight. Every additional gram that the robot weights implies one gram less that can be manipulated by the later robot. Therefore the construction maxim was to use as few components as possible and within as small space as possible. To provide a counterweight to the handled payload and to alleviate the drive train design, the motors for the fourth, fifth and sixth axis were located at the third joint. The challenge was to place three motors next to each other, but still have their drive shafts positioned coaxial since a parallel position of two or more axes would lock at least one of the three axes. This problem was solved by using a spur gear drive with a various amount of gears combined with a hollow shaft which holds another shaft inside. The spur gears bridge the two dimensional displacement of the motor shafts from the coaxial position. Each of the two coincide shafts inside the upper arm assembly drives a bevel gear drive inside the lower arm which subsequently drives the fifth and sixth axis of the robot using a timing belt connection and another bevel gear box inside the TCP head. The assembly can be seen in figure 3. For the purpose of adjusting

the belt tension of the fifth and sixth axis it was necessary to split the lower arm into two pieces, as to adjust the belt tension by a longitudinal displacement between lower arm and wrist assembly. The wrist is accordingly hold in the position by the cover that encloses the lower arm assembly, together with the exposed belt drives.

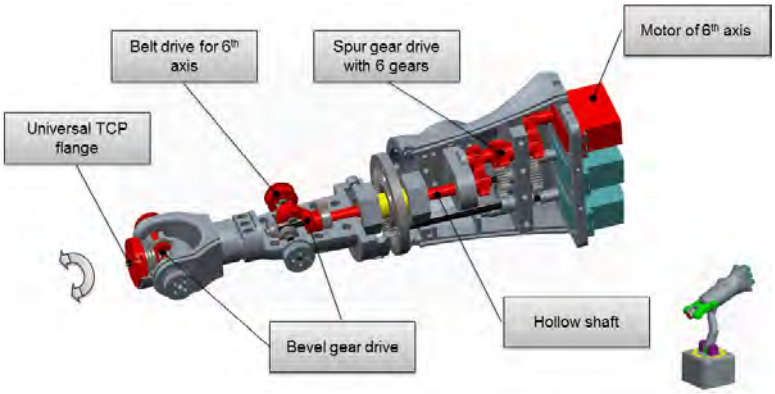


Figure 3. Lower arm assembly with visible torque path of 6^{th} axis

6. Conclusion and Outlook

Concluding, the robot was designed and can be used within future projects. The platform was split into several enclosed, modular sub-assemblies to be able to adapt or change only parts of the robot and to alleviate the final assembly of the robot. Furthermore the robot was designed in the most cost efficient way that was possible with the given resources. Future projects will have to further improve the electronics, the control system and the software to be able to further program the robot. Another possible future project could implement am additional force feedback system by measuring and controlling the current that flows to the motors. This would enable the user to teach the robot by dragging the manipulator in the desired pose and teaching its joint positions. Another big advantage of measuring and controlling the motor currents would be the ability of preventing damage to the mechanical structure or the motors itself by mechanical overload of each joints. Besides this technical improvements, practical evidence within educational context will show the practicability and usefulness of the learning experiences, the robot is able to offer.

References

[Alessandri, 2012] Giuseppe, Alessandri; Martina, Paciaroni (2012): Educational Robotics Between Narration and Simulation. In The World Conference on Design, Arts and Education (DAE-2012), May 1-3 2012, Antalya, Turkey 51, pp. 104-109.

[Allen, 2015] Allen, Robert J.A.; Trask, Richard S. (2015): An experimental demonstration of effective Curved Layer Fused Filament Fabrication utilising a parallel deposition robot. In Additive Manufacturing 8, pp. 78?87.

[Eguchi, 2010] Amy Eguchi (2010): What is Educational Robotics? Theories behind it and practical implementation. In David Gibson, Bernie Dodge (Eds.): Proceedings of Society for Information Technology & Teacher Education International Conference 2010. San Diego, CA, USA: Association for the Advancement of Computing in Education (AACE), pp. 4006-4014.

[Eguchi, 2016] Eguchi, Amy (2016): RoboCupJunior for promoting STEM education, 21st century skills, and technological advancement through robotics competition. In Robotics and Autonomous Systems 75, Part B, pp. 692-699.

[Gómez et al., 2014] Gómez-Espinosa, A.; Lafuente-Ramón, P. D.; Rebollar-Huerta, C.; Hernández-Maldonado, M. A.; Olguín-Callejas, E. H.; Jiménez-Hernández, H. et al. (2014): Design and Construction of a Didactic 3-DOF Parallel Links Robot Station with a 1-DOF Gripper. In Journal of Applied Research and Technology 12 (3), pp. 435-443.

[Khanlari, 2013] Ahmad Khanlari (2013): Effects of Robotics on 21st Century Skills. In European Scientific Journal 9 (27), pp. 26-36.

[Lee et al., 2011] Lee, Irene; Martin, Fred; Denner, Jill; Coulter, Bob; Allan, Walter; Erickson, Jeri et al. (2011): Computational Thinking for Youth in Practice. In ACM Inroads 2 (1), pp. 32-37.

[Miller et al., 2008] Miller, David P.; Nourbakhsh, Illah R.; Siegwart, Roland (2008): Robots for Education. In Bruno Siciliano, Oussama Khatib (Eds.): Springer Handbook of Robotics. Springer Berlin Heidelberg, pp. 1283-1301.

[Ospennikova et al., 2015] Elena Ospennikova; Michael Ershov; Ivan Iljin (2015): Educational Robotics as an Inovative Educational Technology. In Procedia - Social and Behavioral Sciences 214, pp. 18-26.

[Tocháček et al., 2016] Tocháček, Daniel; Lapeš , Jakub; Fuglík, Viktor (2016): Developing Technological Knowledge and Programming Skills of Secondary Schools Students through the Educational Robotics Projects. In Future Academy Multidisciplinary Conference "ICEEPSY & CPSYC & icPSIRS & BE-ci"13-17 October 2015 Istanbul 217, pp. 377-381.

[Varela et al., 1995] Varela, Francisco J.; Thompson, Evan; Rosch, Eleanor (1995): Embodied mind. Cognitive science and human experience. 4. print. Cambridge, Mass.: MIT Press.

[Wing, 2008] Wing, Jeannette M. (2008): Computational thinking and thinking about computing. In Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 366 (1881), pp. 3717-3725.

[Yoo, 2015] Yoo, Juyoung (2015): Results and Outlooks of Robot Education in Republic of Korea. In International Educational Technology Conference, IETC 2014, 3-5 September 2014, Chicago, IL, USA 176, pp. 251-254.