Vision-based Human Detection from Mobile Machinery in Industrial Environments

Örebro Studies in Technology 68

RAFAEL MOSBERGER

Vision-based Human Detection from Mobile Machinery in Industrial Environments

Title: Vision-based Human Detection from Mobile Machinery
in Industrial Environments
Publisher: Örebro University 2016
www.oru.se/publikationer-avhandlingar

Print: Örebro University, Repro 03/2016

ISSN 1650-8580
ISBN 978-91-7529-126-0

Abstract

Rafael Mosberger (2016): Vision-based Human Detection from Mobile Machinery in Industrial Environments. Örebro Studies in Technology 68.

The problem addressed in this thesis is the detection, localisation and tracking of human workers from mobile industrial machinery using a customised vision system developed at Örebro University. Coined the *RefleX Vision System,* its hardware configuration and computer vision algorithms were specifically designed for real-world industrial scenarios where workers are required to wear protective high-visibility garments with retro-reflective markers. The demand for robust industry-purpose human sensing methods originates from the fact that many industrial environments represent work spaces that are shared between humans and mobile machinery. Typical examples of such environments include construction sites, surface and underground mines, storage yards and warehouses. Here, accidents involving mobile equipment and human workers frequently result in serious injuries and fatalities. Robust sensor-based detection of humans in the surrounding of mobile equipment is therefore an active research topic and represents a crucial requirement for safe vehicle operation and accident prevention in increasingly automated production sites. Addressing the described safety issue, this thesis presents a collection of papers which introduce, analyse and evaluate a novel vision-based method for detecting humans equipped with protective high-visibility garments in the neighbourhood of manned or unmanned industrial vehicles. The thesis provides a comprehensive discussion of the numerous aspects regarding the design of the hardware and the computer vision algorithms that constitute the vision system. An active near-infrared camera setup that is customised for the robust perception of retro-reflective markers builds the basis for the sensing method. Using its specific input, a set of computer vision and machine learning algorithms then perform extraction, analysis, classification and localisation of the observed reflective patterns, and eventually detection and tracking of workers with protective garments. Multiple real-world challenges, which existing methods frequently struggle to cope with, are discussed throughout the thesis, including varying ambient lighting conditions and human body pose variation. The presented work has been carried out with a strong focus on industrial applicability, and therefore includes an extensive experimental evaluation in a number of different real-world indoor and outdoor work environments.

Keywords: Industrial Safety, Mobile Machinery, Human Detection, Computer Vision, Machine Learning, Infrared Vision, High-visibility Clothing, Reflective Markers

Rafael Mosberger, School of Science and Technology
Örebro University, SE-701 82 Örebro, Sweden, rafael.mosberger@oru.se

Acknowledgements

First, I would like to express my gratitude to my advisors Achim J. Lilienthal and Henrik Andreasson who both have been tremendously supportive and encouraging mentors during my studies. I am grateful to have been given the opportunity of working at the Mobile Robotics and Olfaction Lab at the Centre for Applied Autonomous Sensor Systems (AASS) at Örebro University.

My gratitude also goes to all my colleagues who, through their support, fruitful thinking, and enthusiasm, make AASS a stimulating and productive work environment. Special thanks further go to Per Sporrong and Bo-Lennart Silfverdal for their technical support during my hours spent in the lab with soldering, calibrating cameras, or setting up experiment sessions.

A further big thank-you goes to Erik Schaffernicht, Stephanie Lowry, Henrik Andreasson and Achim J. Lilienthal who contributed valuable comments on how to improve the thesis text.

It was an honor for me to work with Bastian Leibe and his Computer Vision Group at RWTH Aachen during my stay as a guest researcher, where I received a lot of support to tackle the problems addressed in this thesis.

I would like to acknowledge the big technical and practical help of Admir Ribic from Linde Material Handling, Torbjörn Martinsson from Volvo CE, and Johan Larsson from Atlas Copco during the experimental evaluation of our vision system on various industrial machinery.

My gratitude further goes to Charlotta Nordenberg and Nils-Olof Gunnarsson from External Relations at Örebro University for raising the public awareness of our research, and for their support in protecting our intellectual property rights. I also express my gratitude to Peter Stany from Robotdalen, who offered substantial support for the realisation of the industrial prototype of our camera system that is depicted on the cover page.

Finally, my sincere gratitude goes to my family who provided me with their encouragement and support during all the years of study. I am also very grateful for the charming presence of Alla Rybina who efficiently managed to keep me away from work during many weekends.

So much for the formalities. Now, as my colleague and friend Todor recently pointed out to me, the acknowledgements section is likely to be the only part of this PhD thesis that most among you who will get hold of this book will ever actually read. I consider that reason enough to give it a personal touch by adding a pinch of hopefully entertaining information. Also, this is the section where I feel entirely comfortable with applying modifications at will after handing in the manuscript for revision by my supervisors.

After having spent several years in a robotics lab, I can now confidently say that I have learned a lot. Many things appear in a clearer light than they used to in the beginning. However, there are still questions within the robotics community that leave me completely puzzled at times, and to which I probably will never find an answer. For example: What on earth is this ridiculous obsession with Star Wars movies!? I simply do not get it. Or, an equally persistent issue: Who the bloody heck is Sheldon!?

On a completely different topic, did you know I have a couch in my office? I know some of you do, not all for the same reason, though. Anyway, if you do not have a couch in your office, you most likely have one at home. And if you further happen to be a researcher you have probably found yourself in the situation that, on a rainy Sunday afternoon, you planned to read an important research paper. Surely, after reading a paragraph or two while sitting on a chair, you thought it was more comfortable to read the rest of the paper lying on the couch. I bet that was the last thing you were consciously thinking for quite a while that day and you finally ended up reading the paper in your office the day after. As a conclusion, I really think sofas were not designed for reading research papers on them. They are simply too comfortable.

By the way, have you ever tried to quickly type the word *acknowledgements* on your keyboard and managed to get it right? Me neither. It's virtually impossible! It is an irritating word, deliberately and maliciously designed to annoy everybody who attempts to use it. Even if I type it slowly, I start with something like acknolegements, correct it to acknoledgements, then try acknowlegments before figuring that acknowledgements somehow looks most familiar but without being entirely sure if it is correct. So I look it up again.

With this said, I wish you all the best for whatever you are up to today! If you think that the topic I have been working with for writing this thesis is interesting, you may want to glance through the book! There are a lot of illustrative figures that show what my work is about and you don't need to be an engineer to understand them! If, instead, you think that the topic is boring and you really don't know what to do right now, you can browse to page 57 and try to find something that does not belong in the pictures.

List of Publications

This thesis is a compilation of scientific publications. The articles are listed in the chronological order that they were written and are referenced throughout the text using the indicated labels.

Paper I

Rafael Mosberger and Henrik Andreasson. *Estimating the 3D Position of Humans wearing a Reflective Vest using a Single Camera System.* Field and Service Robotics, Springer Tracts in Advanced Robotics, Springer Berlin Heidelberg, 92, pages 143–157, 2014.

Paper II

Rafael Mosberger and Henrik Andreasson. *An Inexpensive Monocular Vision System for Tracking Humans in Industrial Environments.* Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pages 5850–5857, 2013.

Paper III

Rafael Mosberger, Henrik Andreasson and Achim J. Lilienthal. *Multi-human Tracking using High-visibility Clothing for Industrial Safety.* Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 638–644, 2013.

Paper IV

Rafael Mosberger, Henrik Andreasson and Achim J. Lilienthal. *A Customized Vision System for Tracking Humans Wearing Reflective Safety Clothing from Industrial Vehicles and Machinery.* Sensors MDPI 2014, 14:10, 2014, pages 17952–17980, 2014.

Paper V

Rafael Mosberger, Bastian Leibe, Henrik Andreasson and Achim J. Lilienthal. *Multi-band Hough Forests for Detecting Humans with Reflective Safety Clothing from Mobile Machinery.* Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pages 697–703, 2015.

Further publications of the author which are not part of this PhD thesis but make use of the proposed camera system include:

Robert Krug, Todor Stoyanov, Vinicio Tincani, Henrik Andreasson, Rafael Mosberger, Gualtiero Fantoni, Antonio Bicchi and Achim J. Lilienthal. *On Using Optimization-based Control instead of Path-Planning for Robot Grasp Motion Generation.* IEEE International Conference on Robotics and Automation (ICRA), Workshop on Robotic Hands, Grasping, and Manipulation, 2015.

Robert Krug, Todor Stoyanov, Vinicio Tincani, Henrik Andreasson, Rafael Mosberger, Gualtiero Fantoni and Achim J. Lilienthal. *The Next Step in Robot Commissioning: Autonomous Picking & Palletizing.* IEEE Robotics and Automation Letters (RA-L), 1(1), pages 1–8, 2016.

Contents

1 **Introduction** 1
1.1 Motivation . . . 2
1.2 Problem Statement . . . 4
1.3 Contributions . . . 5
1.4 Thesis Outline . . . 6

2 **Background and Related Work** 7
2.1 High-visibility Clothing in Industry . . . 9
2.2 Sensor Modalities for Human Detection . . . 9
2.3 Pedestrian Detection in Road Traffic Scenes . . . 12
2.4 Human Detection in Industrial Environments . . . 13

3 **Sensors** 19
3.1 Near-infrared (NIR) Sensing . . . 21
3.2 Visible-light RGB Sensing . . . 22

4 **Models and Methods** 27
4.1 Reflector Extraction . . . 27
4.2 Reflector Classification . . . 28
4.2.1 Randomised Classification and Regression Forests . . . 29
4.3 Human Appearance Model . . . 30
4.3.1 Hough Forests . . . 32
4.3.2 Multi-band Hough Forests . . . 33
4.3.3 Model Learning . . . 34
4.3.4 Model Inference . . . 38

5 **Systems and Applications** 43
5.1 Human Detection and Tracking . . . 44
5.1.1 Monocular NIR Vision . . . 45
5.1.2 Stereoscopic NIR Vision . . . 48
5.1.3 Combined NIR and RGB Vision . . . 50

5.1.4 The Role of the Protective Garments 53
5.2 Alternative Applications 55
5.2.1 Vehicle Pose Tracking 56

6 Conclusion and Future Work 59

Chapter 1
Introduction

The problem addressed in this thesis is vision-based detection and tracking of human workers from manned or unmanned mobile industrial machinery. The proposed novel approach uses the reflective properties of conventional protective workwear as the key feature to achieve robust detection performance in harsh industrial conditions. Its application aims at increasing occupational health and safety in a broad range of industrial environments.

Work environments in industrial sectors such as logistics, construction or mining are frequently organised as shared workspaces which have no physical separation between pedestrian routes and vehicle operation areas. In many scenarios humans must carry out tasks in close proximity to machines and vehicles, or directly interact with them. This exposes human workers to a number of constant safety risks such as getting struck or rolled over by a moving vehicle or getting caught between vehicles and stationary objects. Due to the significant dimensions and mass of common industrial machinery, such accidents often result in serious injuries and death.

To minimise the number of accidents involving human workers and mobile machinery, the industry has seen an increasing trend towards using intelligent sensor technology that monitors the surroundings of a vehicle and provides information about the presence and location of objects and persons. The availability of such sensory information is important for a multitude of applications. It can provide the input to advanced driver assistance systems for human-driven vehicles, and is even indispensable when building autonomous machinery. Here, the robotic vehicle is entirely dependent on such sensory input for safe path planning and collision avoidance. Regarding the technologies in use, cameras are among the most widely used sensors, due to both the rich scene information they provide as well as their low cost.

This thesis focuses on a novel vision-based approach for human detection from mobile machinery, such as but not limited to forklifts, loaders, dump trucks or mining vehicles. The method was specifically designed for applications in industrial environments where human workers are equipped with protective clothing with retro-reflective markers. The proposed vision system, coined *RefleX Vision System*, consists of a customised active camera setup and a set of computer vision algorithms that in combination detect and locate retro-reflective markers and use this capability to specifically track human workers over time. A particular challenge in that process is that the sensor system might also frequently encounter other reflective objects than the workers' safety garments. This makes a closer analysis and classification of the observed reflective patterns necessary for robust system performance.

Strong focus has been given to the industrial applicability of the proposed approach. The thesis therefore discusses the specific challenges and requirements of a human detection system for industrial applications in construction, logistics or mining, and presents methods for coping with these challenges. A thorough experimental evaluation has been carried out on a series of video sequences that were acquired in real-world industrial indoor and outdoor environments. A selection of environments in which the system has been tested is shown in Figure 1.1.

This thesis is a collection of five scientific articles. It offers a summary and synthesis of the work carried out over the course of several years of research within the field of human workforce detection for the industrial sector. Building on a common core concept, several variations of the RefleX vision system have been presented in the articles. This thesis reviews the underlying sensors, models and algorithms, offers a comparison of the proposed system configurations, and discusses their respective advantages and limitations.

1.1 Motivation

The underlying research is important for multiple reasons. A great deal of research studied pedestrian detection for road traffic scenes which has led to the deployment of advanced driver assistance systems that are now wide-spread among new generations of cars. At the same time, the industry of mobile machinery has not yet seen the same progress. As it will be discussed in Chapter 2, little research has focused on the specific requirements of human detection modules for industrial applications. While some commercial systems with human detection capabilities are available on the market, their performance is far from satisfactory.

Furthermore, wearing protective high-visibility clothing with retro-reflective markers is either a legal requirement or mandatory by employer's regulations in many countries. The initial idea of safety garments with reflectors was to ensure that human drivers see people when they are illuminated with a light source on the vehicle. This effect can also be exploited by a sensor, however to the best of the authors' knowledge this idea has not yet been investigated. It is therefore an important research contribution to the field of vision-based human detection to determine the extent to which reflective safety garments can support the human detection task.

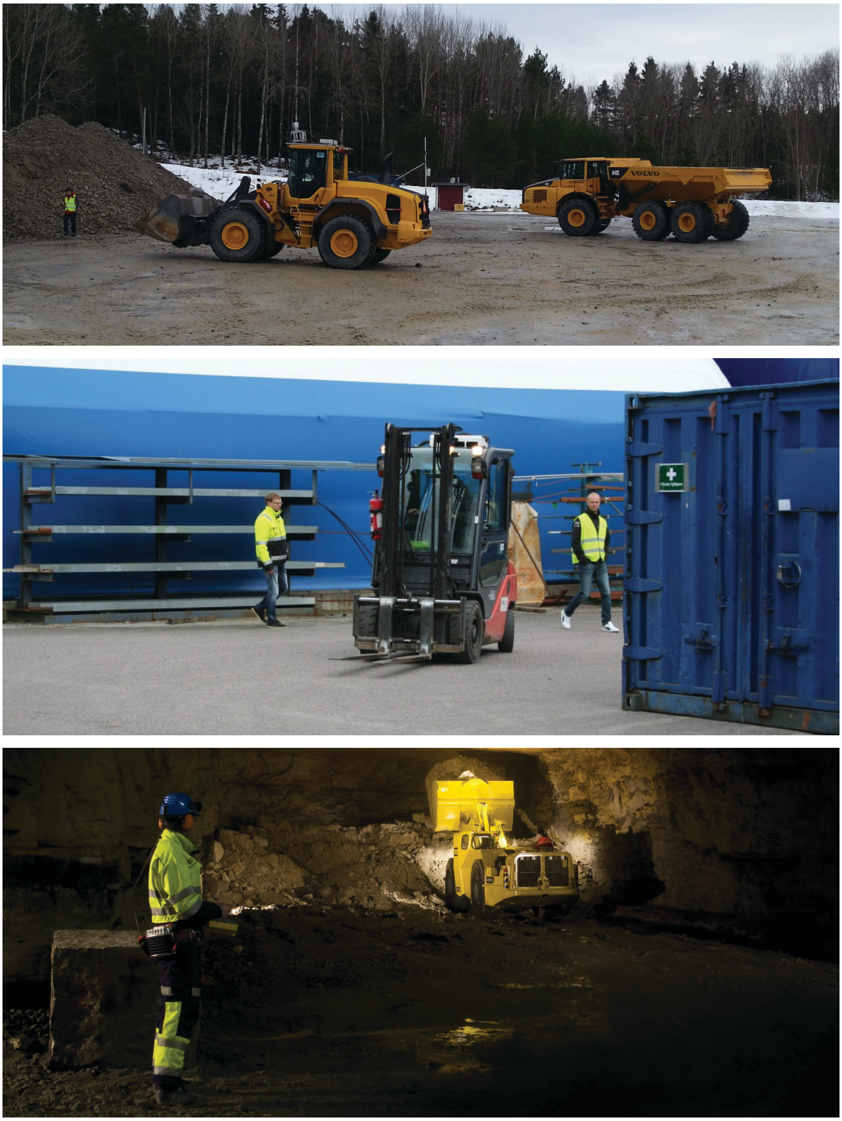

Figure 1.1: Examples of industrial environments in which the proposed vision system has been tested on different machinery: (top) wheel loader and dump truck in a gravel pit, (middle) forklift in an outdoor storage yard, and (bottom) load-haul-dump truck in an underground mine (Image: Atlas Copco).

1.2 Problem Statement

The overall problem addressed by the work underlying this thesis is how to increase and ensure the safety of human workforce around mobile industrial equipment with a novel low-cost sensor system that exploits the domain-specific conditions of industrial work environments. Particular focus is placed on the industrial applicability of the approach, by addressing the specific challenges and requirements imposed by the industry sector. The proposed solution is required to be an on-board system, implying that it is to be physically located on the vehicle, and perform robust human detection, localisation and tracking.

A further requirement is robustness to a wide range of lighting conditions typically met in industrial environments, ranging from broad daylight with direct sun exposure to nighttime conditions with little or no ambient illumination. Therefore, if not stated differently within specific parts of the work, no assumptions are made regarding the prevailing lighting and illumination conditions the vehicle finds itself in. The method further has to be applicable to indoor and outdoor environments without re-adjusting parameters. In view of the potential operation on rough and uneven terrain, we make no assumptions on the planarity of the ground the vehicle is moving on, which stands in contrast to the case of road traffic scenarios.

Given the targeted application area, we make the following assumptions. All human workers in the surrounding of the host vehicle are equipped with protective high-visibility work clothing with several retro-reflective markers. Garments worn by industrial workers may include conventional high-visibility workwear such as vests, jackets or trousers. Furthermore, it can be assumed that there exists a line of sight between the sensor and at least one of the retro-reflective markers on the garment of a worker to be detected.

For defining the precise entities of information the system is supposed to extract, we employ the taxonomy proposed in [71] and list the following four spatio-temporal properties of interest:

- **Presence**: Is there a person present?
- **Count**: How many persons are present?
- **Location**: Where are the persons located with respect to the sensor?
- **Track**: How does a person's location change over time?

It is important to mention that the list explicitly excludes the fifth and last property defined in [71], which is the identity of a person. The objective of the underlying research is increasing occupational safety at industrial work sites, so it is considered essential that a person in the neighbourhood of a vehicle is detected, localised and tracked. However, knowing the identity of the person is not considered necessary in a safety context.

1.3 Contributions

The research that underlies this thesis addresses the vision-based detection of human workers from mobile machinery operating in real-world industrial work environments. To the best of the authors' knowledge, the work presents the first human detection approach that exploits the reflective properties of conventional protective workwear in order to facilitate the detection task and make the resulting system more robust and computationally efficient. The specific contributions of this thesis are:

- Design of a customised low-cost hardware setup aimed at perceiving retro-reflective markers. The thesis proposes a tailored infrared vision system with spectral filtering and active illumination that allows for a distinctive separation of retro-reflective markers from the image background, and thereby a significant complexity reduction of the subsequent image processing chain (Paper I, Paper II).

- Design of an algorithm that robustly extracts reflective markers from successive pairs of infrared images, acquired with and without active illumination. The approach builds on the input from the specialised infrared camera and specifically copes with challenging lighting conditions such as direct sun exposure (Paper I, Paper II).

- Implementation of a supervised learning based classification algorithm for distinguishing safety garments from other reflective objects, as well as a regression algorithm that estimates the distance between the camera and an observed reflective garment from monocular vision input (Paper I).

- Implementation of an algorithm for tracking multiple industrial workers in 3D space. The algorithm assigns observed reflectors to individually tracked persons by a applying a measurement model taking the uncertainty of the distance estimates into account (Paper II, Paper III).

- Design and implementation of an algorithm for learning and inference of a human appearance model which fuses multiple spectral bands by incorporating features from NIR and RGB images (Paper V). The model learns the spatial distribution of image patches of particular appearance with respect to a defined object centre.

- Collection of a set of video sequences[1] acquired by the hardware configuration deployed in this work (Paper I–Paper V). The sequences are recorded in a range of indoor and outdoor environments, contain both NIR and RGB image data, and show persons with reflective garments in a variety of body poses. No such data sets were found publicly available.

[1] Parts of the data set are publicly available under www.mrolab.eu/datasets.html, while portions that are subject to corporate privacy regulations are only available upon request.

1.4 Thesis Outline

The remaining chapters of this thesis are structured as follows:

Chapter 2: Background and Related Work

The chapter gives an overview of sensors and methods commonly used in human detection from mobile vehicles. It is shown how the problem of pedestrian detection has been addressed within the context of road traffic safety, and the similarities to and differences from human detection from industrial machinery.

Chapter 3: Sensors

This chapter introduces the sensor modalities comprising the RefleX vision system, namely NIR and RGB vision. Particular focus is given to the customised configuration of an active NIR camera for sensing retro-reflective markers, which lays the foundation for the efficient human detection approach presented in this thesis.

Chapter 4: Models and Methods

The chapter introduces the underlying models and methods that form the building blocks for the design of several variations of the RefleX vision system as presented in Chapter 5. The discussion includes the robust extraction, description and classification of reflective interest regions as well as the representation and learning of a single or multi-spectral human appearance model.

Chapter 5: Systems and Applications

The chapter revisits the different variations of the RefleX vision system proposed throughout the scientific articles, and compares advantages and drawbacks of the different versions. Monocular versus stereoscopic vision as well as the fusion of multiple spectral bands using NIR in combination with RGB vision are discussed. Furthermore, the chapter gives an insight into applications of the sensor technology other than human detection.

Chapter 6: Conclusion and Future Work

The chapter summarises the contributions and achievements made with the proposed vision system. It further discusses the limitations of the presented approach and gives an outlook on potential future research directions.

Chapter 2
Background and Related Work

Occupational safety ranks among the key areas of activity defined in the social policy of the European Union (EU) and considerable efforts have been taken in recent years to increase safety standards at work sites. The European project ESAW (European Statistics on Accidents at Work) was launched in 1990 with the aim of collecting union-wide statistical data on work-related accidents, including their causes and circumstances. Despite a significant decreasing trend in accidents at work in the EU, occupational safety is far from being achieved and remains a primary concern. According to Eurostat, the statistical office of the European Union, 5 million employees suffer serious work-related accidents each year, while around 5000 occupational fatalities are reported in Europe on a yearly basis. In its report *Causes and circumstances of accidents at work in the European Union*[1] the European Commission presents an assessment of the statistical data with regard to the specific occupation and the activity of victims with the aim to develop more appropriate prevention policies. The investigation revealed that incidents involving human workers getting struck by or colliding with an object in motion account for 35% of all fatal and 18.1% of all non-fatal work related accidents.

In a comparison of accident rates in the EU-15 countries between 1995 and 2005, construction followed by agriculture and transportation are singled out as the three sectors with the highest risk of accidents. A particularly high occurrence, when compared to the other sectors, is registered for fatal accidents. Eurostat further reveals that within the construction sector, every third fatal accident at work involves mobile equipment. Such accidents include persons falling from vehicles, persons getting struck by objects falling from vehicles, death or injury through overturning vehicles, or persons getting struck or run over by the vehicle.

[1] European Commission, DG Employment, Social Affairs and Inclusion. *Causes and circumstances of accidents at work in the European Union.* Office of Official Publications of the European Communities, Luxembourg, 2009.

According to a report from the *European Agency for Safety and Health at Work*[2], the most common cause for occupational fatalities involving vehicles on construction sites are workers being struck or rolled over by mobile equipment. The main reasons for these incidents include poor visibility, inadequate brakes, and untrained drivers. A particular increase in the likelihood for vehicle accidents is observed in the presence of difficult weather conditions, during operation on rough and uneven grounds, and in crowded workplaces where employees work under time pressure.

Similar observations can be made with regard to other sectors in which mobile equipment is heavily utilised, including warehouse facilities, storage yards, manufacturing sites or surface and underground mines. A broad range of mobile machinery is employed in these sectors that constantly expose human workers to a considerable safety risk. The outlined figures regarding occupational accidents clearly illustrate the need for further accident prevention methods. Advanced technological solutions in the form of intelligent sensor systems can thereby play an important role for the implementation of higher safety standards.

There is also an increasing trend towards deploying autonomous mobile machinery for different industrial applications. Examples include the automation of modern warehouse facilities with automated guided vehicles (AGVs) [60, 66], the use of robotic machinery in the construction industry [74, 69], or the deployment of autonomous mining vehicles [23]. Here, robust object and human detection modules are crucial to guarantee the safety of workers around the autonomous equipment. In contrast to the market of driver assistance systems, full autonomy signifies the complete absence of any human being in the control loop that could potentially compensate for a missed detection by the sensor system.

The category of accidents that is addressed with the sensor system discussed in this thesis are human workers that are getting struck or rolled over by industrial vehicles. The purpose of the proposed system is the acquisition of information regarding the presence and location of human workers in a defined neighbourhood of an industrial vehicle. The acquired information may then be used by vehicle manufacturers to design advanced driver assistance systems for human operated vehicles, or navigation and collision avoidance modules for autonomous machinery. The underlying work is a contribution that in combination with other technical measures allows for deployment of new vehicle safety technology and finally contribute to increased industrial safety and reduced accident rates.

[2] EU-OSHA: European Agency for Safety and Health at Work, *E-fact 2: Preventing Vehicle Accidents in Construction*, Office for Official Publications of the European Communities, Luxembourg, 2004

2.1 High-visibility Clothing in Industry

High-visibility clothing is a type of personal protective equipment and comprises any variety of garments with an easily distinguishable, often fluorescent colour and a certain coverage of highly retro-reflective material. The main objective of the garments is to increase the conspicuity of the wearer, or in other words, to make the wearer more easily discernible from any background. Frequent users of high-visibility clothing include road and railroad workers, police officers, firefighters, emergency services, airport personnel, construction workers, and in general human workforce that is frequently engaged in dark areas or in the neighbourhood of moving vehicles. According to the European standards for high-visibility clothing EN 471 and the later EN ISO 20471, an employer is obliged to provide any high-visibility clothing needed for a respective work activity free of charge to any employees who may be exposed to significant risks to their personal safety. In road traffic, high-visibility garments are occasionally used by cyclists and runners, but rather rarely by pedestrians.

The retro-reflective material that covers parts of the high-visibility garments is designed to reflect light backwards in the direction of its source with a minimum of scattering. The principal purpose of this behaviour is to reflect the light emitted by a light source on a vehicle, such as the headlights of a truck, and thereby enhance the visibility of the wearer of the reflective garment in nighttime or low-light conditions.

The principal novelty of the method presented in this thesis consists in the exploitation of the retro-reflective properties of high-visibility garments for the purpose of robustly detecting human workers from mobile industrial machinery. Even though the primary intention behind equipping industrial workwear with reflective markers was to increase the visibility of workers in night-time conditions, it is demonstrated that the approach offers a convenient way of detecting human workforce with an infrared imaging device in both day and night time applications.

2.2 Sensor Modalities for Human Detection

Human detection is a broad area of research where numerous sensor modalities have been employed to address the problem in various contexts and applications. A comprehensive review of the different technologies is therefore beyond the scope of this thesis and the reader is referred to the extensive survey by Teixeira *et al.* [71]. This section instead focuses on a compact discussion of the sensor technologies that have been predominantly employed in literature when addressing the problem of human detection from mobile platforms, that is, when not only the observed target but also the observing sensor might be in motion. Table 2.1 presents a structured overview of the different families of sensor technologies and gives a selection of recent related work.

Sensor Technology	Categories	Related Work
1.) Range Finders		
Lidar	active, uninstrumented	Gidel *et al.* [41], Kidono *et al.* [50], Sato *et al.* [67], Häselich *et al.* [44]
Radar	active, uninstrumented	Ritter *et al.* [64], Chang *et al.* [22], Heuel *et al.* [46], Heuer *et al.* [47]
Sonar	active, uninstrumented	Moebus *et al.* [58], Blumrosen *et al.* [9]
2.) Cameras		
Visible Spectrum (VS)	passive, uninstrumented	Dalal *et al.* [26], Montabone *et al.* [59], Yan *et al.* [76], Milanés *et al.* [56]
Near-infrared (NIR)	active, uninstrumented	Andreone *et al.* [4], Broggi *et al.* [15], Ge *et al.* [39], Luo *et al.* [55]
Thermal Infrared (TIR)	passive, uninstrumented	Suard *et al.* [70], Bertozzi *et al.* [7], Fernández *et al.* [34], Besbes *et al.* [8]
3.) Device-to-Device Ranging		
Radio Frequency (RF)	active, instrumented	Ruff *et al.* [65], Rasshofer *et al.* [63], Koch *et al.* [51], Fackelmeier *et al.* [32]
Magnetic Field	active, instrumented	Schiffbauer [68], Carr *et al.* [21], Jobes *et al.* [49], Teizer *et al.* [72]

Table 2.1: Main families of sensor technologies employed for human sensing from mobile platforms, with a selection of recent literature describing respective single-modality approaches.

For a categorisation of the sensing approaches, the taxonomy suggested in [71] is adopted. Human detection methods may be classified into *instrumented* and *uninstrumented* solutions. While the former class requires each person to carry a device on them, the latter does not depend on any wearable technical equipment. Sensors are further grouped into an *active* and a *passive* category. Passive sensing involves sensing signals that are available in the environment, while active sensing implies that signals are emitted before their responses are measured. Finally, a subdivision into single-modality and sensor fusion approaches has been suggested.

A popular family of active sensors that was studied in the scope of human detection are different versions of range finders. Depending on the medium they use, they are subdivided into sonar (ultrasound), lidar (visible or infrared light) and radar (radio waves). A major advantage of range finders is that they, as their name indicates, directly deliver range measurements without any additional computational effort. Range is thereby most commonly obtained by measuring the timing or energy of the response signal. In multi-transmitter configurations, the precision of range measurements can be increased through techniques such as triangulation. While the obtained range measurements are precise in open space, a considerable noise component is added in cluttered indoor environments as a result of multi-path and scattering effects [71]. This makes robust detection of people based on shape information alone still a challenging task and range finders are therefore frequently employed in combination with vision systems

The most broadly used family of sensors are various types of cameras. In literature, they are divided into several groups according to the spectral range they are sensitive to, namely visible light spectrum (VS, 0.4–0.7µm), near-infrared (NIR, 0.75–1.4µm), and thermal infrared (TIR, 8–15µm) imagers. Visible light imaging especially represents a mature and low-cost technology, allowing for the acquisition of high-resolution data with rich information about the environment. However, extracting the relevant portion of information from an image is often a complex endeavour that can require computationally expensive computer vision and image processing methods. A further difficulty is that the image content is highly affected by several uncontrollable factors including lighting, illumination and weather conditions.

TIR and active NIR vision systems have been widely studied, especially for operation under low light and night time conditions. It is observed that these sensors offer a lower sensitivity to ambient lighting but also to varying textures, colours, and shadows when compared to visible light cameras [8]. Thermal cameras offer the advantage that humans appear in the image as distinct isolated high intensity regions, given that the background has a significantly lower and uniform temperature distribution. However, it is observed that the clothing has a strong influence on the observed thermal structure of a human, and especially thick and highly isolating winter garments can hinder successful detection. Furthermore, no scientific work has systematically addressed the problem of detecting humans with a thermal camera under the frequent presence of heat-emitting objects such as machinery and various electrical facilities that disturb the thermal profile of a human.

Instrumented human sensing approaches, where persons are equipped with wearable devices, are frequently described under the term device-to-device ranging. The core idea is that a wearable device announces its presence by transmitting a signal to a receiver located on a vehicle. The principle has been frequently used for tracking items and supplies in industrial scenarios and is often referred to as proximity detection. Such systems achieve close to perfect detection performance, and can directly deliver the number of people if the wearable tags contain a unique identifier, as it is the case in radio frequency identification (RFID). However, localisation of detected people is not straight-forward and remains an active research topic. Furthermore, the entire personnel of a work environment needs to be equipped with active devices whose maintenance can prove cumbersome.

With respect to the adopted sensor taxonomy, the approach proposed in this thesis can be classified as active and semi-instrumented. The method is clearly active because of the emitted infrared signal. It can be interpreted as uninstrumented because it does not require persons to wear any powered device, or as instrumented because of the requirement that workers wear high-visibility clothing with retro-reflective markers. Nevertheless, the European policy for occupational safety requires employers to provide personnel around vehicles with high-visibility clothing, and the garments can therefore not be seen as part of the sensor solution, but rather as a part of the environmental preconditions in which the sensor system is placed.

2.3 Pedestrian Detection in Road Traffic Scenes

Advanced driver assistance systems (ADASs) and in particular their sub-category pedestrian protection systems (PPSs) have become active and widely studied research areas in the context of road traffic safety. The major purpose of a PPS is the on-board detection of both static and moving pedestrians in order to provide the driver of a vehicle with situational information and if necessary perform evasive braking or steering actions in order to avoid accidents. Although this definition does not specifically exclude vehicles operating at industrial workplaces, the vast majority of research carried out in the field has heavily focused on pedestrian detection in urban traffic scenes.

Considerable advances in the research of PPSs have resulted in the development of the first generations of commercially available pedestrian detection systems. Mobileye[3] offered the first vision-based pedestrian protection system to automotive manufacturers to allow them to integrate collision warning and auto braking systems into their cars. Today, several car manufacturers already offer pedestrian detection warning systems while others plan to integrate them into their vehicles in the near future.

Several comprehensive surveys document the research on pedestrian detection for advanced driver assistance in road traffic scenes. In a broad survey on pedestrian detection methods, Gandhi *et al.* [37] review approaches with different types of active and passive sensors and discuss ways for collision risk assessment. Enzweiler *et al.* [31] survey work on vision-based pedestrian detection, focusing on monocular camera systems, and suggest approaches for the methodological analysis and experimental evaluation of systems. Geronimo *et al.* [40] give an overview on how to incorporate pedestrian detectors into full pedestrian protection systems. The authors offer a review of the state-of-the-art sensors, suggest a general module-based system architecture for PPSs,

[3] http://www.mobileye.com

and discuss different approaches for the individual modules defined in the architecture. Dollár *et al.* [29] perform an extensive evaluation of 16 state-of-the-art pedestrian detection methods focusing on individual monocular images instead of video input.

Two assumptions are commonly made in pedestrian detection that restrict the search space of the problem at hand. People are assumed to be on foot, hence the term *pedestrian*. Furthermore, vehicles are assumed to move on flat road. The first assumption is manifested by limiting certain geometrical variables of pedestrians such as their height and aspect ratio in the image. The flat-road assumption on the other hand is often incorporated in the form of spatial constraints prescribing that pedestrians have to stand on a ground plane. To allow for small deviations from this assumption, the flat-road constraint can be relaxed with a certain tolerance on the pitch angle [38]. More advanced approaches further try to continuously estimate the 3D camera pose in order to take road slope variability and the vehicle dynamics into account [61].

Dollár *et al.* [29] name several directions within the field of pedestrian detection that need further research to cope with more challenging scenarios. These include the detection of pedestrians at smaller scales and under partial occlusion, the use of motion features, and more extensive studies on temporal information integration. Furthermore, the authors suggest utilizing extended context information from road traffic scenes to replace the often employed simple ground plane assumption.

2.4 Human Detection in Industrial Environments

Pedestrian detection from cars in road traffic scenes and industrial purpose human detection from mobile machinery share many similarities. Both aim at robustly detecting humans for the sake of preventing potential collisions that might entail injuries and fatalities. Both applications require to discriminate humans from static objects, as the prevention of collisions with humans is given the highest importance. At the same time there exist a number of significant differences between the two areas which should be taken into account when designing intelligent sensor solutions for the industrial sector.

In the context of road traffic safety and advanced driver assistance, research explicitly focuses on pedestrian detection. A pedestrian is by definition a person travelling on foot. Human detection instead, as the term says, refers to detecting people regardless of body position. When comparing image material from industrial sites and road traffic scenes, a clearly higher body pose variation is observed for industrial workers than for pedestrians in urban scenes. This difference is due to the fact that pedestrians mainly stand or walk, while working in an industrial environment may involve a broad range of work tasks that frequently require bending over, squatting, kneeling, or, albeit less frequently, lying on the floor. The assumption that humans are always on foot and more or less upright standing is not valid in an industrial context and a direct application of pedestrian detectors is therefore not recommended if safety is to be ensured on a broad basis.

Similarly problematic is to maintain the flat-floor assumption and restrict detections to be located directly above ground level. Even if a vehicle actually is moving on flat ground, such as a forklift in a warehouse, it is still not advisable to spatially constrain detections to be located directly on the ground level. A

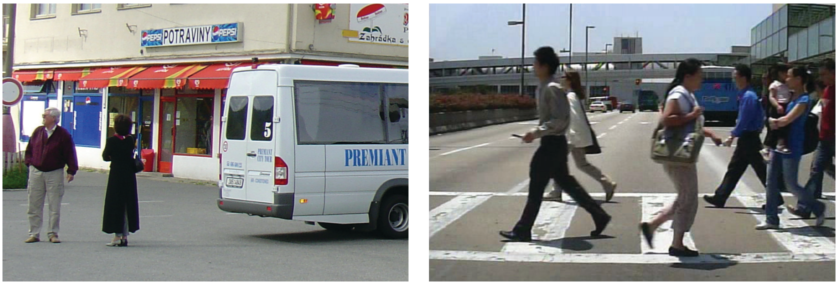

Figure 2.1: Example frames from the INRIA [26] and the Caltech [28] pedestrian detection datasets, showing humans on foot in typical road traffic scenarios.

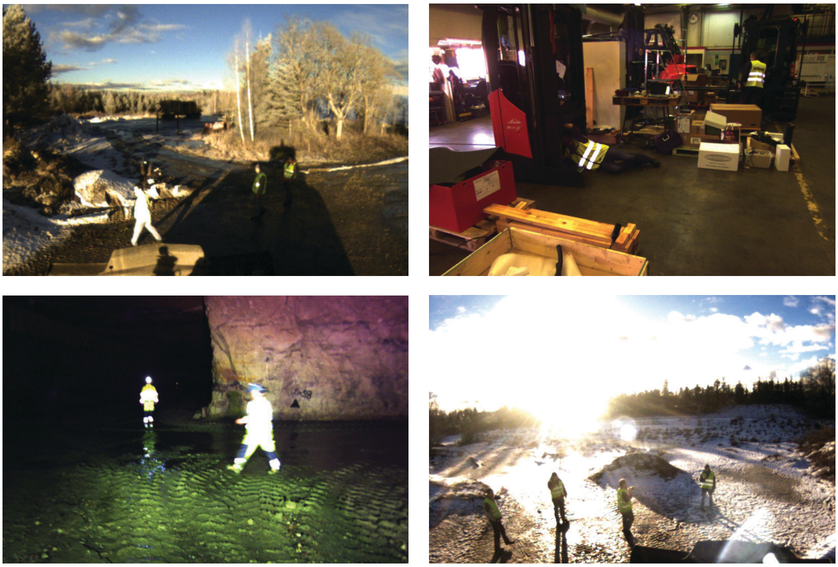

Figure 2.2: Example frames from the proprietary data sets recorded in the scope of this work. The images were acquired from mobile machinery in various industrial environments and show person occurrences under strongly varying lighting conditions.

worker who is climbing up a ladder to pick an object from a shelf should be equally well detected as somebody standing on the floor. Moreover, there exist a number of industrial sites in areas such as construction and mining where a flat floor assumption becomes invalid because vehicles are operating in rough terrain between mounds and cavities.

A further difference concerns the degree to which the environment can be controlled. In road traffic, the appearance of pedestrians is strongly influenced by their clothing and cannot be controlled. Large data sets have been established, such as the Caltech Pedestrian Dataset [28], which allow vision systems to learn the large variability in pedestrian appearance. In contrast, industrial work sites are more controlled environments where the employer can impose rules regarding work clothing and equipment. This implies that instrumented detection approaches can be employed that require workers to be equipped with wearable devices as part of a safety solution. In certain industrial environments such as warehouses or manufacturing sites, employers further have the possibility to install static cameras in addition to on-board safety systems.

Figures 2.1 and 2.2 partially illustrate the described differences between industrial sites and road traffic scenes, and show several challenging example images contained in the data sets acquired and evaluated in the scope of this thesis. A further factor to be taken into account, which is not visualised in the figures, are the motion patterns of cars and industrial machines. Cars regularly move forwards and most proposed sensor systems are therefore forward facing and observing a relatively narrow cone. In contrast, industry purpose vehicles are often involved in loading and unloading scenarios which includes frequent acceleration and braking, sharp turns and reversing. Blind spots and risk zones for accidents are heavily dependent on the vehicle layout but generally include frontal, lateral and rear areas.

Building an industrial purpose human detection system is therefore a complex task. In addition to coping with the described challenges, it has to be mechanically robust and withstand harsh industrial conditions such as vibrations, shocks, and in the case of outdoor operation, the exposure to a range of weather conditions. Table 2.2 presents an overview of research contributions that addressed the specific field of human detection from mobile industrial machinery and that feature an evaluation in industrial work environments. Similarly to the case of road traffic applications, vision sensors represent the most popular family of sensing devices. Even though the authors specifically address industrial environments, the majority do not specifically make use of any particular features in the appearance of industrial workers. Two exceptions to this observation can be named, however. Park *et al.* [62] learn specific colour histograms that incorporate the fluorescent colours of high-visibility vests, while Yang *et al.* [19] perform detection of underground coal miners by means of detecting their helmets which were found to have a more distinctive appearance than the worker's clothing.

Year	Authors	Sensors	Approach
2001	Ruff *et al.* [65]	RF Sensing	Collision avoidance system for haulage equipment in surface/underground mines
2002	Schiffbauer [68]	MF Sensing	Proximity warning system for surface and underground mining applications
2010	Teizer *et al.* [73]	RF Sensing	Proximity alert system that warns both vehicle operators and workers
	Heimonen *et al.* [45]	Stereoscopic VIS Vision	Modular framework for fusion of several pedestrian detector responses
	Carr *et al.* [21]	MF Sensing	Worker proximity detection for mobile underground mining equipment
2011	Dickens *et al.* [27]	TIR Vision + TOF Vision	Human detection using TIR vision and localisition using TOF vision
	Yang *et al.* [19]	VIS Vision	Detection of miners in underground coal mines by detecting their helmets
2012	Park *et al.* [62]	VIS Vision	Detection of construction workers wearing fluorescent safety vests
	Yang *et al.* [77]	Stereoscopic VIS Vision	Omni-directional human detection for a robot tractor
2013	Bui *et al.* [16]	VIS Vision	Human detection in fish-eye images with enhance distortion handling
	Borges *et al.* [11]	VIS Vision	Worker detection and collision prediction using on- and off-board cameras
2014	Bödecker etal. [10]	VIS Vision	Construction worker and equipment detection using optical flow
	Bui *et al.* [18, 17]	VIS Vision + Lidar	Multi-sensor construction worker detection using deformable part models
2015	Costea *et al.* [25]	VIS Vision	Omni-directional stereo vision for obstacle detection in warehouses
	Miseikis *et al.* [57]	VIS Vision	Off and on-board camera fusion for worker detection in industrial scenarios
	Teizer *et al.* [72]	MF Sensing	Proximity alert system that warns both vehicle operators and workers

Table 2.2: Related work in human detection for mobile industrial machinery.
VIS: Visible Spectrum, **TIR:** Thermal Infrared, **TOF:** Time-of-flight,
RF: Radio Frequency, **MF**: Magnetic Field

Approaches which perform information fusion from on- and off-board cameras [11, 57] were shown to yield robust performance over longer evaluation periods. An improvement results from the fact that the scene is observed from different angles using multiple cameras with communication capabilities. Such methods offer the advantage that they can detect humans which are not necessarily in the line of sight of the sensor system on-board the vehicle. However, the necessity of installing static cameras and establishing a central communication system makes their application more cumbersome than pure on-board solutions. Furthermore, as the static cameras maintain a background model of the scene, the system runs the risk of classifying workers as background objects if they are standing still for extended periods of time [11].

Even though some authors investigated vision-based human detection in industrial scenarios, none of the referenced works addresses the particular challenge of detecting non-upright humans. Furthermore, the authors commonly avoid exposing their test systems to the most challenging conditions, such as scenarios with heavy under- or over-illumination. It is therefore difficult to assess the extent to which the proposed methods would cope with challenging real-world conditions.

Several authors also proposed sensor fusion approaches which aim at combining the advantages of different sensor modalities [27, 18, 17]. A popular approach is to perform initial detection on camera data and use the range measurements from sensors such as lidar [18, 17] or time-of-flight cameras [27] to localise detected persons in space. The authors show that a performance increase can be yielded if sensors with complementary characteristics are combined. However, from a commercial point-of-view it is of high interest to limit the number of sensor modalities and with it the manufacturing cost of a sensor system.

In summary, it can be concluded that the operation of mobile machinery at industrial work sites still exposes human workers to a considerable safety risk, and that improving safe working conditions is a major concern of the industry. Relatively little research has been carried out with focus on investigating the use of sensor systems that can contribute to increased safety levels. The material presented in this thesis is therefore an important contribution to the field of industrial safety, because it analyses and highlights an important problem and proposes a novel and low-cost sensor system to address it.

Chapter 3
Sensors

This chapter describes multiple variations of a customised camera-based sensor unit for the specific task of detecting human workers wearing protective garments with retro-reflective markers. The proposed hardware configurations address a concrete safety requirement in the industrial sector, namely monitoring the neighbourhood of heavy mobile machinery with intelligent sensor systems and detecting the presence and location of human workers entering a defined risk zone. For broad industrial applicability, sensor systems have to be suitable for indoor and outdoor use as well as day and night time operation. This requires a high robustness towards illumination conditions that can range from over-exposure to bright sunlight to poorly illuminated or even completely dark working areas.

All sensor setups presented in this chapter are composed of imaging sensors, optical components such as filters and lenses, and electronic circuitry for active illumination of the observed scene. Their purpose is the acquisition of images from mobile industrial machinery which capture and depict the characteristic key features of the appearance of industrial workers, in particular the reflectivity and fluorescent colours of their protective garments.

Different variations of camera-based sensor devices have been studied in the scope of this research. All setups feature at least one near-infrared (NIR) camera, customised as detailed in Section 3.1, that is dedicated to the acquisition of monochrome images in which retro-reflective markers appear as distinct high-intensity regions of interest. More established hardware pieces were further equipped with RGB camera which senses complementary appearance information such as colour and texture. Figure 3.1 depicts the different hardware devices assembled in the process and used during the experimental evaluation. The monocular NIR camera in Figure 3.1a has been employed for the work presented in Paper I and Paper II, and for parts of Paper IV. The multi-camera rig shown in Figure 3.1c was used for Paper III, Paper IV and Paper V. Further testing and evaluation as discussed in Chapter 5 has been carried out on the ba-

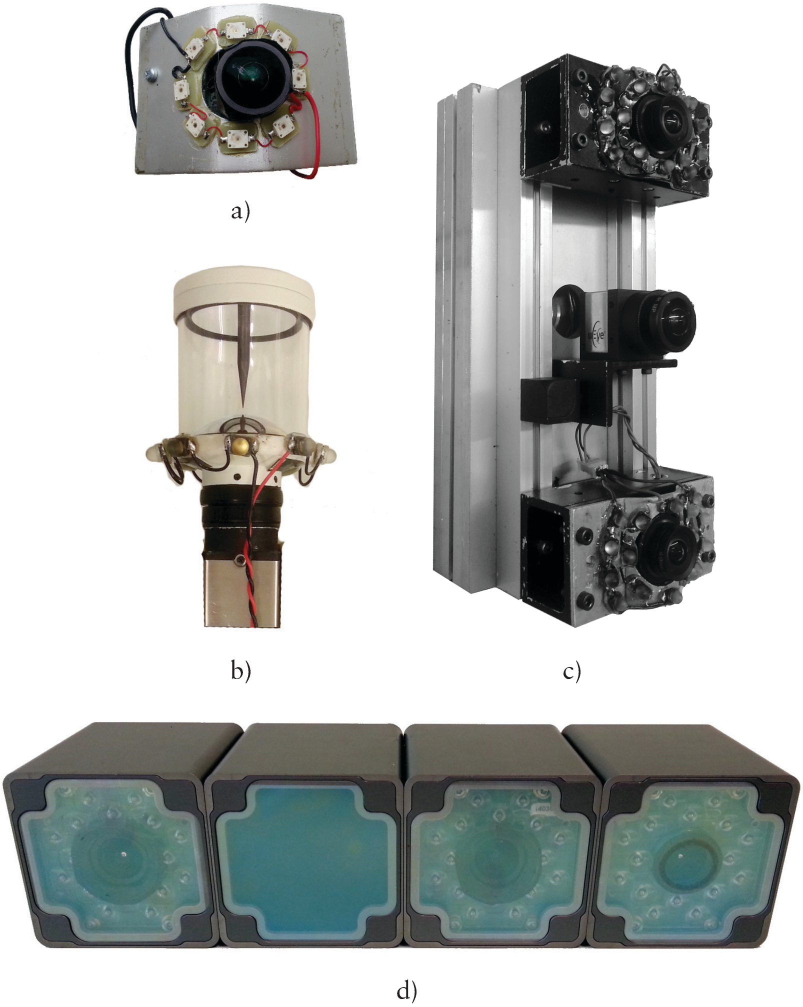

Figure 3.1: The figure shows the different camera configurations designed for the underlying research: (a) Monocular NIR camera used in Paper I, II and IV, (b) its omnidirectional variant, (c) the multi-camera rig with two NIR and one RGB camera used in Paper III–V, and (d) its more robust and industrialised version.

sis of the omni-directional NIR camera device according to Figure 3.1b and an industrialised version of the RGB and NIR camera module (Figure 3.1d).

3.1 Near-infrared (NIR) Sensing

The human detection approach presented in this thesis uses the retro-reflective markers attached to industrial workwear as the key feature to trigger the detection pipeline discussed in the two subsequent chapters. This requires robust and efficient detection and extraction of the reflectors from the acquired image material. Consequently, it is essential to separate reflective interest regions from the non-reflective image background on an early sensory level, and thus decrease the complexity of the subsequent image processing methods.

The desired separation is achieved using a combination of monochrome image sensor, optical band-bass filter, and active light source. The interplay between these three principal components pursues two goals regarding the acquired images: 1) depict retro-reflective markers as bright as possible, and 2) depict everything else as dark as possible. Figure 3.2 shows the schematic setup of the proposed sensor while Figure 3.3 describes the spectral characteristics of its individual components.

The role of the band-pass filter is to suppress the influence of any secondary light source to the extent possible, and make objects with low reflectivity appear dark in the image. On the other hand, short pulse-wise illumination from an NIR light source takes the role of saturating the retro-reflective markers in the acquired images. The key parameters in the design of the proposed device are:

- **Filter Bandwidth.** Ideally, the filter suppresses all incoming light that was not emitted by the sensors' own light source. This can be achieved by using a narrow filter band which coincides as much as possible with the spectral emission curve of the light source. A filter band with fullwidth at half maximum (FWHM) of 10 nm has proven effective for this purpose.

- **Centre Wavelength.** Especially under the influence of sunlight during outdoor operation, the centre wavelength of both the illumination unit and the bandpass filter are preferably matching a negative peak in the radiation spectrum of the sun. As illustrated in Figure 3.3, several negative peaks can be distinguished in the spectrum, due to atmospheric gas absorption. A centre wavelength of 940 nm has proven appropriate to limit the effect of background illumination as illustrated in Figure 3.5 (middle) and further discussed in Chapter 4.

- **Illumination Intensity.** The intensity of the light source has to be strong enough to achieve a clear separation of the retro-reflective markers from the background in the acquired images. The parameter depends on the exposure time and the desired sensor range, as the amount of reflected light decreases quadratically with increasing distance from the sensor.

- **Exposure Time.** Images are acquired using a relatively short exposure time. This avoids motion blur and in combination with the optical bandpass filter suppresses to a large extent the illumination of objects that are not highly reflective. A value between 1 ms and 3 ms has been found appropriate to cover a detection range up to 20 m distance.

- **Light Source Location.** It is crucial that the light source is located as close as possible to the lens. This is due to the fact that retro-reflective markers only reflect light back in the direction of its source with a minimum of scattering. As illustrated in Figure 3.1, a ring of infrared LEDs has therefore been placed closely around the lens.

Ideally, if no secondary light source coincides with the filter band in use, the acquired images resemble the example given in Figure 3.4 (middle) in case no active illumination is used, and Figure 3.4 (bottom) if the scene is illuminated with a flash pulse from the sensors' own light source. Contrastingly, Figure 3.5 (middle and bottom) show the characteristic appearance of the acquired images under the presence of a secondary light source that contains wavelenghts that are transmitted by the filter. The combined use of images taken with and without active illumination for robust extraction of reflective interest regions will be discussed in Chapter 4.

It is important to state that a similar sensor could be built using a visible light image sensor. However, infrared sensing is strongly motivated for two reasons. First, the emitted NIR light pulses will not be detectable for the human eye and thus not disturb the human workforce. Second, the solar radiation power is lower in the infrared domain than in the visible light domain. The problem of background illumination through secondary light sources as illustrated in Figure 3.5 (middle) and further discussed in Chapter 4 is therefore significantly reduced.

3.2 Visible-light RGB Sensing

While active NIR vision has been used to capture the characteristic reflectivity of protective garments, the additional use of an RGB camera has proven useful for several reasons. Most notably, the RGB offers a highly complementary source of information compared to the customised NIR sensor described in the previous section. Instead of focusing on reflectivity, the RGB camera captures the scene in its entirety, and its images offer a much richer source of information comprising structure and colour. In particular, the RGB data allows the observation of the typical colours of the worker's safety garments and the characteristic human silhouette distinguishing them from the image background. Furthermore, the additional RGB input was of great practical help throughout the research presented here. It was heavily used for data interpretation, annotation and labelling, as well as during the experimental evaluation and the visualisation of the results. The principal drawback of using RGB input is that the input image strongly varies with the illumination conditions.

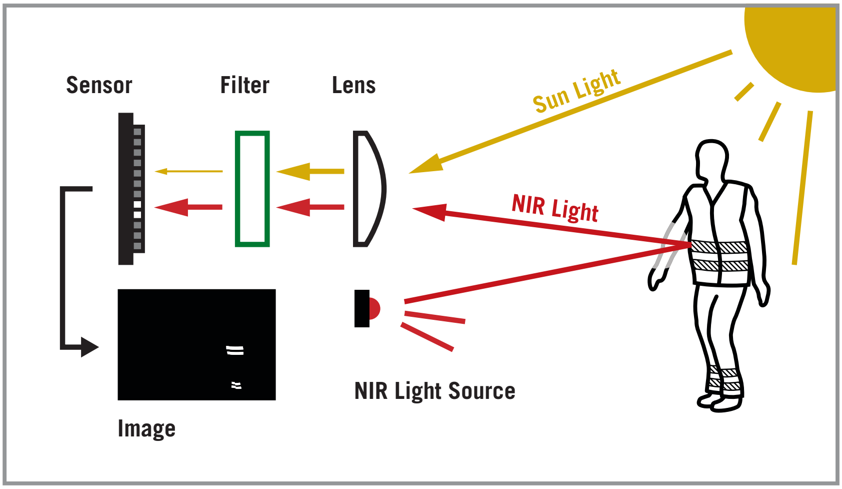

Figure 3.2: The figure shows the schematic structure of the NIR sensor designed to acquire images that discriminate objects with high reflectivity from objects with low reflectivity. Sunlight (yellow) as well as light from other secondary light sources is filtered to a high extent by the optical bandpass filter (green), leading to a dark image background. The NIR light emitted by the sensor's own light source (red) corresponds to wavelengths transmitted by the filter, leading to reflectors being depicted white in the image.

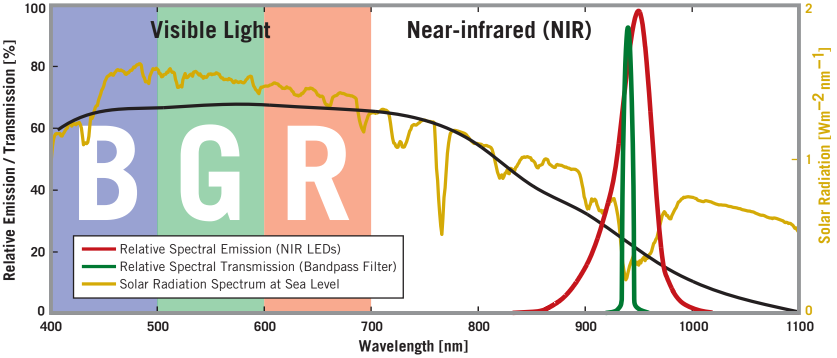

Figure 3.3: The figure shows the relative spectral characteristics of the the bandpass filter (green) and the LEDs used for active illumination (red). The yellow curve represents the solar irradiation spectrum at sea level (Source: ASTM [35]). The operation wavelength of 940 nm is chosen to exploit the negative peak in the sun spectrum. Especially in outdoor applications, this allows us to considerably reduce the undesired background illumination.

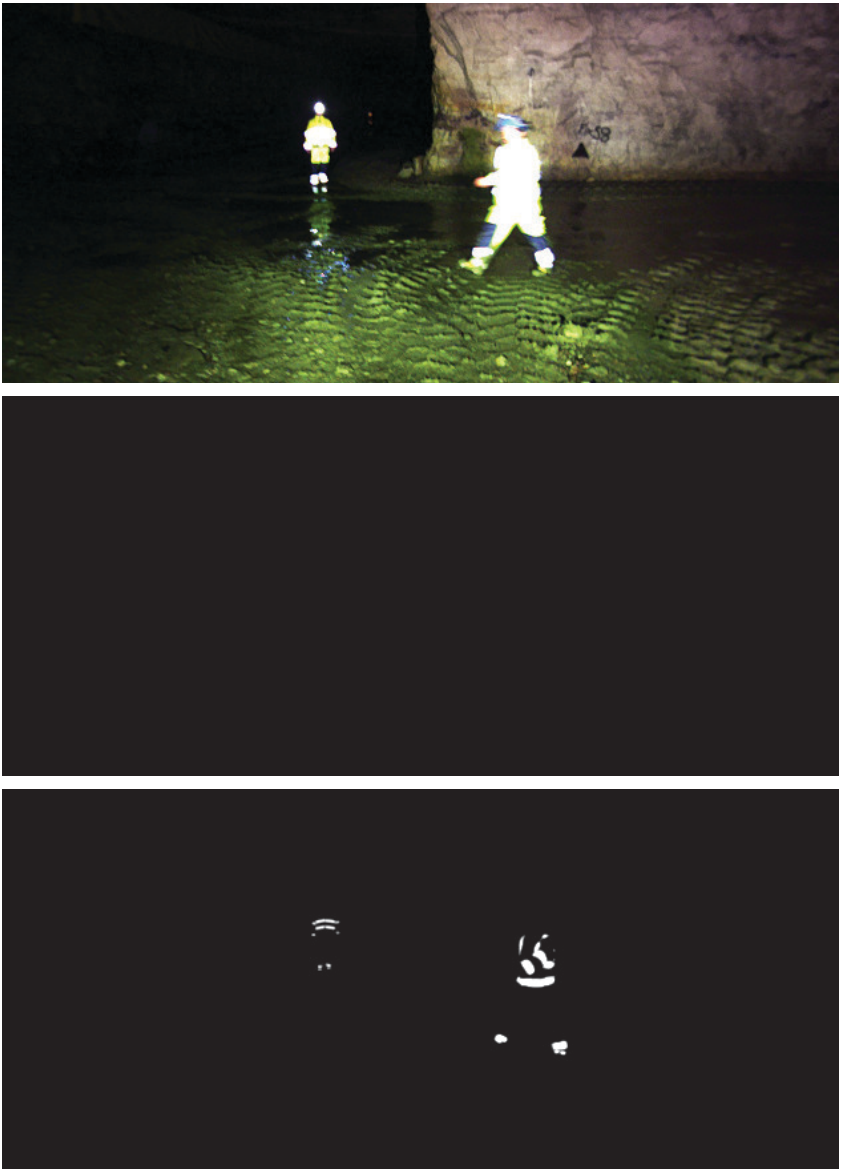

Figure 3.4: Image acquisition in absence of secondary NIR light sources: RGB image (top) and corresponding NIR images taken without (centre) and with (bottom) active illumination. The bottom picture shows a distinct separation between highly reflective markers and the image background, offering sufficient information for the extraction of reflective markers. The centre picture instead does not contain any relevant information.

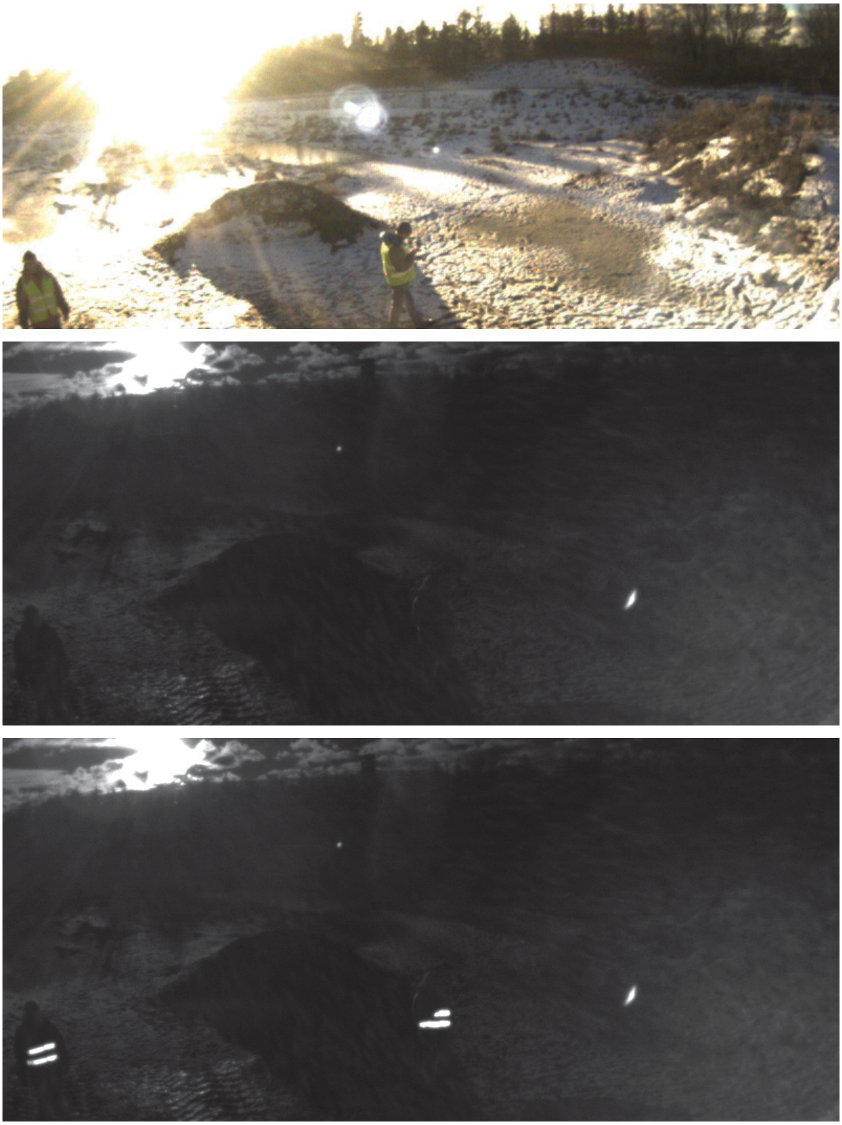

Figure 3.5: Image acquisition under the presence of a strong secondary NIR light source (the sun): RGB image (top) and corresponding NIR images taken without (centre) and with (bottom) active illumination. In contrast to Figure 3.4, the bottom picture alone does not provide sufficient information for differentiating between bright areas that reflected the emitted NIR flash pulse and areas that were illuminated by the secondary light source or show the light source itself. Therefore, the centre picture serves as a reference for the background illumination. Note: the bright white spot on the right hand side of the NIR images is a lens artefact caused by the strong backlight.

Chapter 4
Models and Methods

This chapter offers an overview and summary of the principal models and algorithms developed for the purpose of detecting industrial workers with the sensor configuration discussed in the previous chapter. The text further provides references to the different scientific articles where the respective parts of the models and methods are detailed and analysed in more depth. Models and algorithms are the result of a specific design process that takes into account the particular nature of the sensor data acquired with the infrared imaging device discussed in Section 3.1.

The chapter contains three main sections. Section 4.1 covers the extraction of retro-reflective markers from a stream of infrared images and discusses the challenges posed by additional infrared light sources in the environment. Section 4.2 then summarises the methods employed for analysing the reflective regions extracted from an image and for computing further entities such as class probabilities or depth estimates. Finally, Section 4.3 discusses a particular human appearance model for industrial workers equipped with protective garments. The discussion addresses the incorporation of several distinctive features of the highly characteristic work clothing in terms of reflectivity and colour into a multi-spectral appearance model that fuses NIR and RGB data.

4.1 Reflector Extraction

This section resumes the proposed approach for identifying regions that depict retro-reflective markers in images acquired with the NIR camera configuration described in Section 3.1. It is assumed that the observing sensor unit is mounted on an industrial vehicle and potentially in motion. It is further assumed that the environment in which the sensor system is placed might contain other secondary NIR light emitting sources than the camera's own infrared flash unit.

The extraction procedure takes a pair of monochrome NIR images as input, both taken by a single NIR camera unit in short succession and with a short exposure time (1–3 ms). One of the two images, I_{nf} (nf: no flash), is taken without active illumination and serves as a reference image for the momentary ambient illumination caused by other NIR light emitting sources in the environment. The second image, I_f (f: flash), is then taken with active illumination and registers the response from retro-reflective markers in the scenery. An example of such an image pair under the presence of a strong NIR emitting light source (e.g. the sun) is illustrated in Figure 3.5 (centre and bottom figure). Using the two input images I_{nf} and I_f, the extraction of reflectors is achieved in two subsequent steps, a *candidate generation* and a *verification* step. The algorithm is discussed with some variations in Paper I, Paper II and Paper III.

Due to the active NIR illumination during image acquisition, retro-reflective markers appear as high-intensity blob-like regions in image I_f. The candidate generation step therefore aims at locating these high-intensity regions in image I_f. To this end, two different methods have been used over the course of the work. In Paper I and Paper II, a circular blob detector, the *Center Surround Extrema (CenSurE)* feature detector [1] was employed, resulting in a set of circular blob features with respective centre coordinates and a scale measure. In Paper III, this procedure was replaced by local adaptive thresholding followed by the extraction of connected components from the resulting binary image. The latter approach comes with the advantage that reflectors are extracted as coherent units instead of a loose set of feature points, which makes it possible to compute additional variables regarding size and geometry of a reflector.

Under the presence of other NIR light sources, the candidates extracted from I_f are not directly guaranteed to represent reflective items. In fact, the image regions might depict a secondary NIR light source itself, or regions that are brightly illuminated by such a light source. In both cases however, the respective image regions will have similar appearance when compared in I_{nf} and I_f, assuming that the secondary light source is not drastically changing intensity in the short time window between the acquisition of the two images. In contrast, truly reflective items will appear significantly brighter in image I_f, as it is depicted in Figure 3.5. In a verification step, the detected candidate regions in I_f are therefore compared with the corresponding regions in I_{nf}. Due to camera motion or changes in the scene, the exact image coordinates of the interest regions can differ by small amounts between the two input images. The algorithm proposed in Paper I and Paper II therefore aims at relating the respective image regions using the Lucas-Kanade feature tracking method [12] before measuring the intensity difference within a close neighborhood of the candidate regions and rejecting candidates with low difference.

4.2 Reflector Classification

An analysis of various video sequences acquired in different industrial environments revealed that the retro-reflective markers on the protective garments are not the only items with highly reflective properties. Typical examples of other reflective objects include windows, mirrors, cat's eye reflectors on vehicles, and different types of reflective signage attached to the walls or floor. In consequence, it has to be assumed that the set of reflectors, as extracted by the procedure outlined in the previous section, contains items which do not belong to the class of interest. Specifically detecting humans with reflective garments therefore requires the approach to discriminate between different types of reflectors. A simple separation in terms of size or geometry of reflectors has thereby proven unsuccessful. The task is therefore formulated as a two-class classification problem and approached with supervised machine learning methods. To do so, a large amount of training samples are collected depicting either a retro-reflective marker on a safety garment, or another arbitrary reflective item observed in a range of different industrial environments. Each training sample is assigned a label designating to which of the two classes it belongs. Particular focus has to be given to covering a large variety of body positions in the acquisition of the training samples, as the observed reflective patterns of a garment as observed in the NIR images vary with the body position and the angle from which a person is observed.

Reflector classification is not performed on the raw NIR image data but on local image feature descriptors extracted from the neighborhood of the previously detected reflective regions. Depending on the choice, a feature descriptor can thereby represent a vector of either numeric (e.g. SIFT [54], SURF [6]) or binary (e.g. BRIEF [20], BRISK [53], FREAK [2]) variables. Local image feature descriptors lead to a considerable dimensionality reduction with respect to the raw image data and aim to encode the most distinctive information from an image region in less variables.

Two of the most popular classifiers described in literature were initially considered for solving the classification problem, namely *support vector machines* (SVMs) [24] and *random forests* [14]. Experiments conducted in the scope of Paper I and Paper II indicated that random forests outperform SVMs for the problem at hand, that is, the appearance based classification of reflectors on the basis of the NIR image data. Furthermore, random forests provide a convenient basis for building the particular human appearance model that will be discussed in Section 4.3.2, and were therefore the preferred choice of learning algorithm for the remainder of the work.

4.2.1 Randomised Classification and Regression Forests

Randomised decision forests [48, 3] are a machine learning method for building classification and regression models from an ensemble of decision trees that are trained using randomised feature selection. They have been shown to successfully counteract the problem of training data overfitting that is frequently observed with conventional decision trees. The principle was further enhanced by the technique of *bootstrap aggregating* [13], where each classifier of an ensemble is trained on a different subset of the training data, and was subsequently trademarked under the name of *random forests* [14].

During the supervised learning stage, a separate bootstrap sample for every tree is taken from the labelled training data, using sampling with replacement. Each tree is then trained independently on its own bootstrap sample in a recursive manner, starting from the root node. At each node, the set of training samples are split into two subsets by submitting each sample to a binary test. The binary tests evaluate a function of one or several elements of the feature vectors and compare it to a threshold. Depending on the test result, a sample is then propagated to either of the two subnodes. A large number of randomly generated tests is evaluated at each node, and the test which leads to the highest information gain in the two subnodes is selected and added to the tree model. The same procedure is then applied to the respective subnodes, and data splitting continues until either a specified depth is reached or until a node contains only one training sample. The leaf nodes finally store a distribution over the annotated categorical (classification) or numerical (regression) entity of interest.

Inference is performed by propagating a test sample down the tree of the learned random forest and applying the feature tests selected during the training phase. The procedure thereby aggregates the posterior distributions over the entity of interest stored in the final leaf nodes. This aggregation of votes from multiple independent decision trees happens through majority voting in the case of a classification problem, or by prediction of the mean value in a regression problem. As it will be discussed in Chapter 5, random forests have been used throughout Paper I to Paper IV for binary classification of reflector families as well as for the estimation of depth from monocular NIR input in Paper I and Paper II.

4.3 Human Appearance Model

The previous two sections presented methods for the extraction, analysis and classification of reflective markers observed in images acquired with the customised infrared camera described in Section 3.1. This section focuses on two additional topics, namely the relation between a set of observed reflectors and individual person occurrences and the fusion of NIR data with traditional RGB images in order to build a more distinctive human appearance model, incorporating both the reflectivity and the characteristic colour of protective garments.

As defined in the introduction, it is part of the problem statement to not only detect whether or not a person is present in the field of view of the camera, but also to infer the number of humans and their individual locations with respect to the sensor. Extracting these entities from pure NIR data becomes more complex as multiple persons are located close to each other, and as the number of retro-reflective markers on one person increases. The problem is well illustrated by the example in Figure 4.1, where workers are equipped with both protective jackets and trousers that in total feature 14 retro-reflective markers of different size and shape. In addition, several retro-reflective cones are placed in the same scene. The depicted scenario poses several challenging tasks. First,

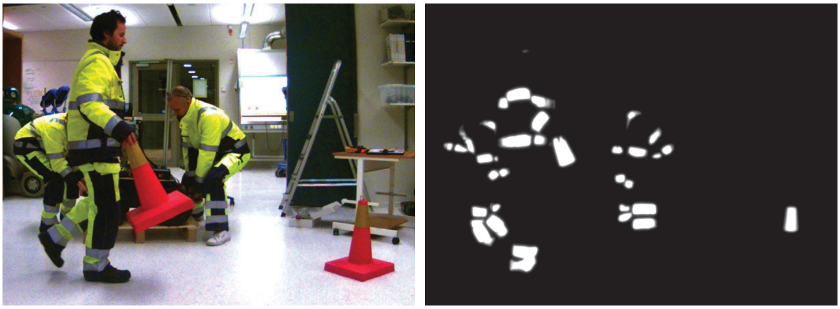

Figure 4.1: The figure illustrates the challenging problem of associating reflectors in the NIR data to individual person occurrences. While the three persons are easily distinguished in the RGB image (left), extracting the number of persons and their location from the corresponding NIR data (right) is significantly more challenging.

the question arises which of the reflectors in the NIR data actually originate from a protective garment. Then the portion of reflectors that is believed to do so, needs to be associated with specific persons in the image and the locations of these persons need to be derived by incorporating the evidence provided by the entire set of the reflectors.

To address these problems in a unified approach, a human appearance model building on the popular Hough forest method [36] is proposed. The model learns the characteristic spatial distribution of local image feature patches around a defined reference point on the human body. By doing so, the model establishes a relation between the specific appearance of local features and the spatial location where these features typically appear with respect to a defined reference point on a person. During the detection stage, a generalised Hough voting procedure then collects the evidence provided by the entirety of observed image features and locates the defined reference points of individual person occurrences in the image.

The approach was first evaluated on the basis of pure NIR data in combination with the standard implementation of Hough forests which processes single channel images. From the point of view of practical applicability this comes with the advantage that the entire processing chain remains independent from ambient lighting conditions. However, in applications where good illumination conditions can be assumed, the additional information provided by an RGB camera in terms of colour, texture and structure can contribute to building a stronger, more distinctive appearance model. An extended version of Hough forests, coined *multi-band Hough forests*, therefore introduces a convenient way of fusing information from NIR and RGB images in the same model. The method was proposed in Paper V and is summarised in Section 4.3.2 of this thesis.

4.3.1 Hough Forests

Hough forests [36] are random forests enhanced with the ability of performing a generalised Hough transform [5]. Building on the concept of the *implicit shape model* [52], they represent a codebook of local prototypic image patches with given relative location from a predefined reference point on the object, usually the object centre. Thereby, the approach stands in strong contrast to the family of holistic models (e.g. [26]) that are frequently used in pedestrian detection and which model an object as a single entity. In order to cover the large variety of body positions and articulations observed in certain industrial scenarios, a holistic approach would require analysing the images simultaneously with a multitude of different templates. A more appropriate approach represents the family of *deformable part models* (DPMs) [33], which detect individual connected parts of the human body and by design handle high degrees of articulation much better. However, they are computationally expensive and involve considerable body part annotation efforts as part of the model learning process. Hough forests, with their representation of local image feature patches, have been shown to successfully handle larger degrees of body articulations [78, 42] while they are at the same time more computationally efficient than DPMs. Similar to DPMs, they are further able to detect and locate object instances even under partial occlusion.

During the supervised learning stage, a Hough forest learns a mapping from local image feature patches to a probability distribution over a defined parameter space. Here, this parameter space consists of the likelihood of an image patch to depict a part of the object class, and for training samples of the foreground class, the two-dimensional location of the patch with respect to a defined reference point of the object class. To build a Hough forest, an ensemble of randomised decision trees are trained recursively on a large collection of image feature patches. The patches are sampled from both training images of the given object class and images depicting various background scenes. Each patch is further labelled accordingly. Patches representing the object class also store an offset vector indicating the location of the training patch with respect to an annotated object reference point. Given the labelled training data, building a Hough forest largely follows the randomised procedure known from conventional random forests. However, splitting a set of training samples at a given node now pursues two different objectives. A split attempts to create an information gain either by reducing the uncertainty in the class membership of patches, or by reducing the variance of their offset vectors.

During the detection stage, local feature patches are extracted from the input images and propagated down every tree of the forest. Compared to conventional random forests, the model now performs classification and regression at the same time. In fact, the stored leaf-node distributions not only allow the model to assign foreground probabilities to the analysed images regions but also to perform a generalised Hough transform and estimate the location of the reference point of hypothesised objects. By the dense sliding window based sampling of feature patches from previously defined regions of interest in the test image, the method thereby aggregates evidence from the observation of a large amount of local features in a unified voting scheme.

4.3.2 Multi-band Hough Forests

The concept of *multi-band Hough forests* has been introduced in Paper V. Multi-band Hough forests combine the advantages of local feature based models, such as Hough forests, with the ability to conveniently fuse information from multiple spectral bands in a unified model. They represent an extended concept of Hough forests, adapted to the case where an appearance model is to be learned from multiple images showing the same object at identical resolution, but sensed within different spectral bands. Similar to standard Hough forests, the patches in a multi-band Hough forest incorporate various stacked feature channels which are calculated from the raw input data. However, the feature stack now represents a collection of channels computed from either of the input images.

In the framework of this thesis, the concept of multi-band Hough forests has been evaluated by fusing NIR and RGB input images. The motivation comes from the fact that key features that provide evidence for the presence of a person wearing protective garments are to be found in both types of images. While NIR images depict distinctive high-intensity image regions wherever a retro-reflective marker appears on a garment, RGB images provide complementary distinctive cues including colour, structure and texture which are not captured in the infrared band. Depending on the location from where features are extracted, one or the other channel might contain more discriminative information, as illustrated in Figure 4.2. Imposing rules for when to best exploit features from either spectral band is therefore difficult, as the answer is location-dependent and differs for individual types of protective garments with varying reflective patterns. The core idea is therefore to fuse multiple spectral channels in the scheme of Hough forest feature patches and exploit the randomised feature selection mechanism of Hough forests to find the most discriminative features for splitting a given set of patches at each node in the decision trees.

Figure 4.2 shows two examples of feature patches sampled from images depicting a person equipped with protective jacket and trousers. One patch is extracted from an image region in the neighbourhood of a retro-reflective marker, while the other patch is sampled from a non-reflective region. The figure illustrates that while the NIR data contains highly discriminative appearance information if the patch depicts a reflector, the complementary RGB input is clearly the richer source of information in regions that depict non-reflective structures.

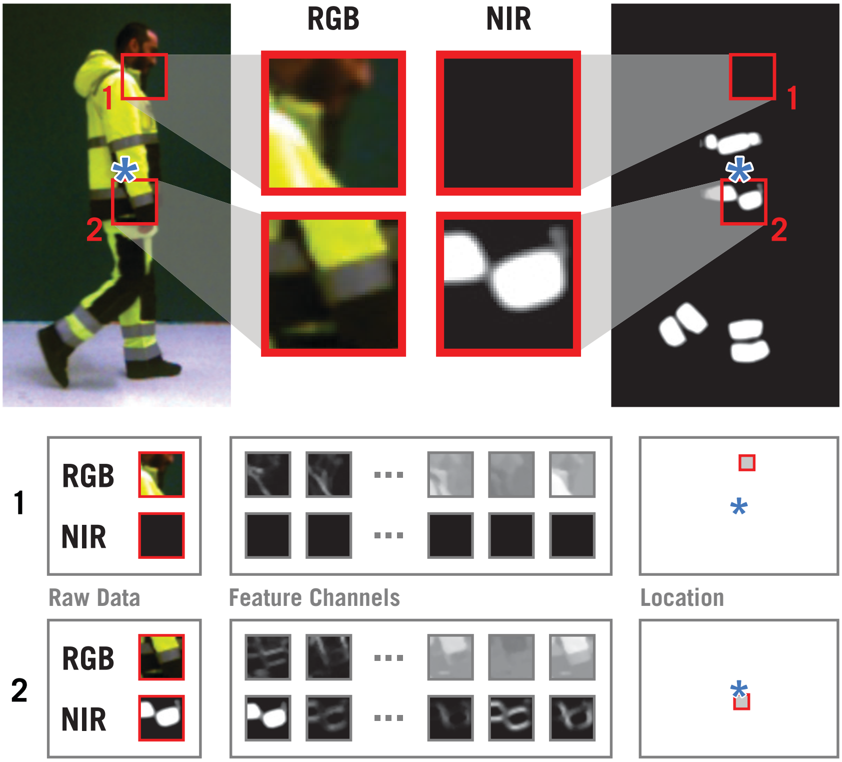

Figure 4.2: The figure illustrates the concept of multi-band feature patches which build the basis for learning multi-band Hough forests. Square patches (outlined red) are sampled from corresponding regions in the RGB and NIR training images. Each patch is resized to a reference scale and stores a set of stacked feature channels that are computed from either RGB or NIR image data. Each feature patch further holds an offset vector indicating the location where the patch has been extracted from with respect to an annotated reference point (blue star).

4.3.3 Model Learning

The learning procedure of a multi-band Hough forest model starts with the acquisition of a large set of training images, captured simultaneously in the RGB and NIR spectrum. As illustrated in Figure 4.3, the training material covers a large variety of different body positions and articulations in order to account for the broad range of work tasks potentially carried out by industrial workers. The figure shows several exemplary body positions with corresponding RGB and NIR snapshots. The training data is then annotated with a reference point, which was defined as the point passing through the centre of the torso and lying between the two horizontal reflectors of the safety jacket (see Figure 4.4a). Furthermore, a bounding box delimiting the extent of the person in the image is annotated. The image background is masked in the RGB data in order to avoid the particular background scene of the training environment to be incorporated into the appearance model. This results in the advantage that training images can be conveniently acquired in a controlled environment with a background scene different from the ones encountered in industrial scenarios where the model is applied later.

Reflectors are then extracted from the NIR training images according to the procedure discussed in Section 4.1. Subsequently, the reflectors observed in the NIR data have to be related to the corresponding regions in the RGB image. However, due to the fact that the cameras observe the scene from slightly different perspectives, there exists a difference in the location of corresponding regions in the RGB and NIR images. This difference, referred to as the *binocular disparity*, is inversely proportional to the depth of an object depicted in an image. Provided that the camera setup is in frontal parallel configuration, disparity exists only in one dimension. To account for it, either a range sensor can be employed and disparity is computed from range measurements, or disparity is directly computed from stereo NIR data acquired by a camera configuration such as the versions presented in Figure 3.1c or 3.1d.

The training images are subsequently rescaled to a reference size. Various feature channels are computed both from the RGB images and the NIR images. Features channels may include amongst others image derivatives of different order, channels from different colour spaces such as RGB or LUV, or histogram of oriented gradient (HoG) features. Multi-spectral feature patches are then extracted from various locations on the human silhouette. As depicted in Figure 4.2, the patches consists of a set of feature channels extracted from the region delimited by a rectangular sampling window. Furthermore, each patch stores a two-dimensional offset vector designating the location from where the training patch was extracted with respect to the annotated reference point. All patches sampled from the human silhouette are finally given a class label marking them as foreground patches and added to the set of training patches.

A complementary set of feature patches is then sampled from a collection of background scenes depicting numerous characteristic industrial indoor and outdoor environments. Here, it is of particular importance to also include frames that depict various instances of frequently encountered reflective items, so that the model can learn to discriminate the appearance of these types of reflectors from the reflective markers on the workers' garments. The patches sampled from these background scenes are marked with a respective background class label before being added to the collection of training patches.

Building the multi-band Hough forest from the set of extracted feature patches finally follows the approach known from conventional Hough forests. By means of the supervised data splitting procedure during which a large set

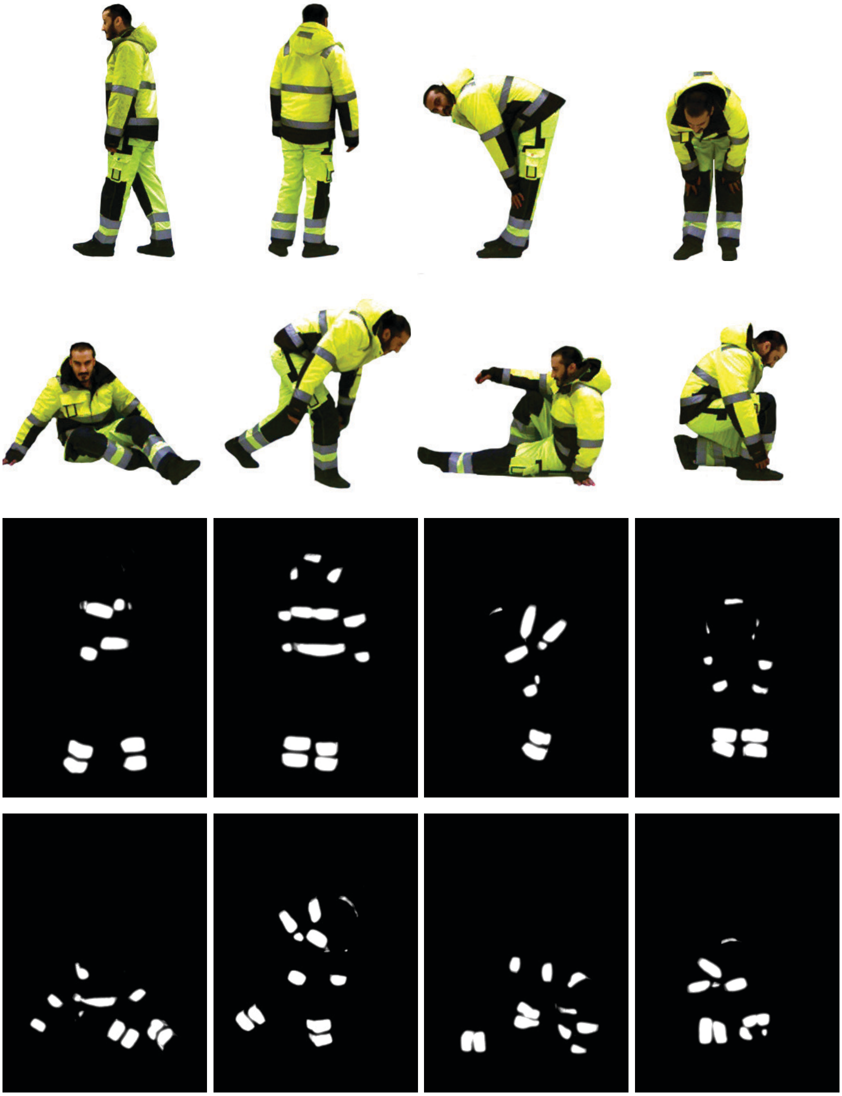

Figure 4.3: The figure shows examples of a rich set of training images acquired for learning a multi-band human appearance model from RGB and NIR data. The training samples are specifically selected to cover a wide variety of different body positions.

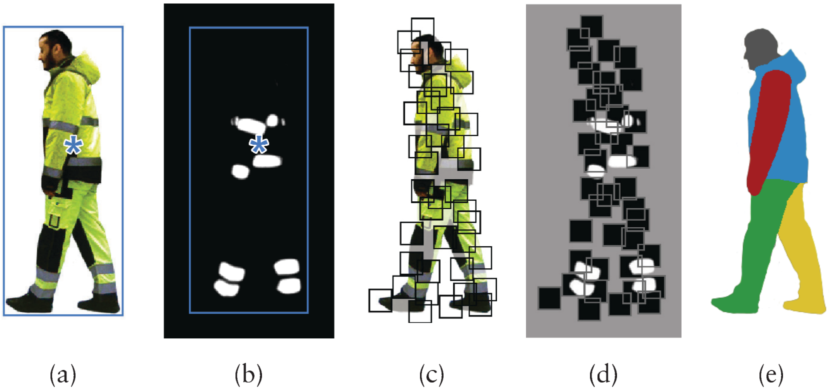

Figure 4.4: The figure illustrates the sampling procedure that extracts corresponding feature patches from RGB and NIR images during the multi-band Hough forest learning stage: (**a–b**) training image annotations in terms of bounding box and object reference point (blue star), (**c–d**) square training patches sampled from corresponding NIR and RGB image regions within the human silhouette, and (**e**) the additional pixel-wise body-part annotations shown to improve detection performance if used for supervision during the training procedure of the Hough forests.

of potential candidate splits are evaluated and compared, the Hough forest structures the patches step by step and reduces the uncertainty in terms of foreground/background class as well as the variance of offset vectors at each node, as proposed in [36]. However, it has been shown in the context of Paper V that the detector performance can be increased by introducing an additional third supervision criterion which is based on ground-truth pixel-wise body part annotations. Exploiting these additional annotations, the training procedure has been extended in order to not only reduce class and offset uncertainty, but alternately also attempt to regroup patches with similar body part labels while the data is structured in the tree. In the framework of Paper V, this approach was evaluated on persons wearing reflective vests only. A specific automated pipeline for the efficient extraction of such body part labels was proposed, which involved equipping the persons that appeared in the training frames not only with a reflective vest but additionally with trousers and sweaters that featured individually coloured parts. By doing so, an automated ground-truth image segmentation according to various body parts could be applied before the training stage in order to extract the necessary pixel-wise annotations. However, the approach is not generally applicable for other types of protective clothing, especially not to the case where workers are equipped with both upper- and lower-body garments as shown in Figure 4.3. In this case, pixel-wise body part labels need to be annotated manually.

4.3.4 Model Inference

After the learning stage is concluded, the learned appearance model is applied to perform human detection on unseen image material. Again, this requires that reflective markers have been previously extracted from the raw NIR input according to the method described in Section 4.1. Successful model inference further requires that per-reflector depth estimates have been obtained.

The detected reflective markers in the input images are then processed one by one. Figure 4.5 visualises the generalised Hough voting procedure for one particular reflector extracted from the input images. It starts with the generation of a square region of interest (ROI) in the NIR image, covering a wider but still locally confined area in the neighbourhood of the reflector under consideration. Region correspondence is established between the NIR and RGB images by computing the respective ROI disparity from the available depth of the reflector. The ROIs are then rescaled to the reference size adopted during training before the feature channels are computed from each type of image data. A sliding window with dimensions corresponding to the training session then densely scans the ROI to extract local feature patches.

The extracted patches are propagated down the learned Hough forest until ending up in a leaf node within each tree. The distributions stored in the respective leaf nodes are exploited to perform a generalised Hough transform. Probabilistic votes for the presence of an object at a specific location are accumulated in a three-dimensional Hough space, consisting of a stacked set of two-dimensional Hough images. Each layer of the stack thereby corresponds to a scale. Into which of the scale layers a vote is cast, is inferred from the per-reflector depth estimate. The location where a vote is cast within the given layer is inferred from the offset vector of a matched training sample. Finally, the probabilistic weight with which a vote is cast is inferred from the class distribution stored in the respective leaf node. Using a three-dimensional voting space allows the Hough transform to discriminate individual object occurrences not only in the image space but also the scale space. This is especially helpful in situations where humans appear close to each other in the image space but appear at different distances, as in the example shown in Figure 4.1.

Once all reflectors of a frame have been processed and all votes have been cast, each layer of the Hough space is smoothed with a two-dimensional Gaussian kernel adapted in size to match each scale layer. Object hypotheses, comprising the 2D location of the reference point in the image plane, the characteristic scale and a voting score, are subsequently extracted from the 3-dimensional Hough space using non-maxima suppression.

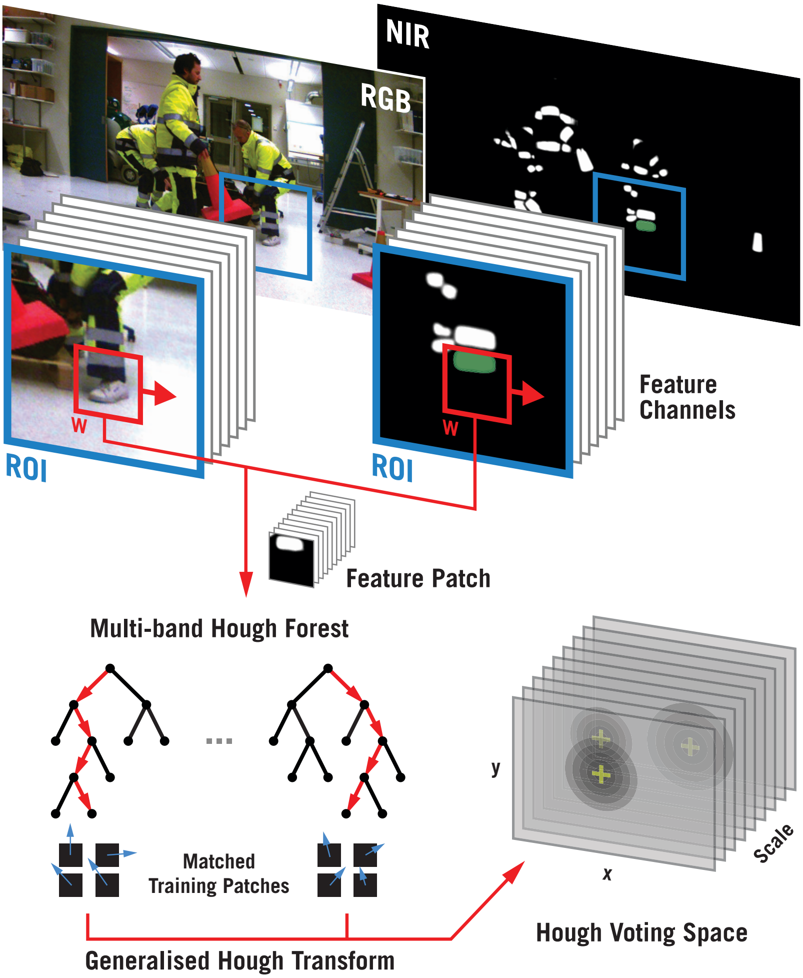

Figure 4.5: The figure illustrates the generalised Hough voting procedure triggered by the presence of one particular example reflector (marked green in the NIR image). The reflector generates a region of interest (ROI, blue), which is scaled to a reference size. Feature channels are computed from RGB and NIR data within the defined ROI before feature patches are densely sampled by a sliding window (red). The features patches are propagated down the learned multi-band Hough forest, and the class and offset vector distribution in the resulting leaf nodes are used to perform a probabilistic Hough transform in a three-dimensional Hough voting space. Subsequent Gaussian smoothing and non-maxima suppression serve to isolate the locations of the individual persons.

Backprojection and Bounding Boxes Estimation

Apart from detecting object occurrences and estimating the position of their reference points in the image, it is desirable to produce a bounding rectangle, usually referred to as the bounding box, that delimits the image region in which the object appears. Given the fact that humans can occur at various distances and that no prior assumption is made in terms of body position, defining an accurate bounding box involves estimating three entities:

- **Scale:** How large should the area of the bounding box be, e.g. what is the number of pixels covered by it?
- **Aspect Ratio:** What is the ratio of the width to the height of the rectangle? This entity is dependent on the body position. Many pedestrian detectors fix it to a value of around 0.4, corresponding to an upright person.
- **Centre Position:** Where should the bounding box be centred? Depending on the choice of reference point during data annotation (cf. Figure 4.4a), the box centre does not necessarily coincide with the reference point.

The scale of the bounding box can directly be estimated by using the average area of all annotations in the training data (in a normalised reference scale) and rescaling it to the scale corresponding to the layer in the Hough space where an object has been extracted from. To estimate the remaining entities, two additional voting steps are carried out after performing a feature patch selection process referred to as *backprojection*.

Backprojection involves densely scanning the ROIs around the defined reflectors a second time, extract feature patches in identical manner as in the first pass, and propagating them through the Hough forest a second time. This again leads to the same matches of training patches, however, during this second pass only the portion of patches is retained which cast a vote into the neighbourhood of a detected object. Here, the neighbourhood is defined by the size of the Gaussian kernel previously applied to smooth the individual layers in the Hough space. For each local maximum extracted from the Hough space, this results in a set of feature patches that *support* the detected object. This subset of feature patches is then used to cast additional votes for the aspect ratio of the associated bounding box as well as its centre location with respect to the reference point of the object.

Figure 4.6 and 4.7 illustrate the output from this final bounding box estimation stage and show how the algorithm manages to produce bounding boxes with an aspect ratio adapted to the observed body position. It is to be noted that the estimation of the bounding box centre was not part of the work presented in Paper V and has been investigated later. In Paper V bounding boxes were centred around the reference point. A visual inspection reveals that bounding boxes are more accurately estimated if their centre location is individually estimated, however the influence of the additional procedure on the detection rates has not been investigated.

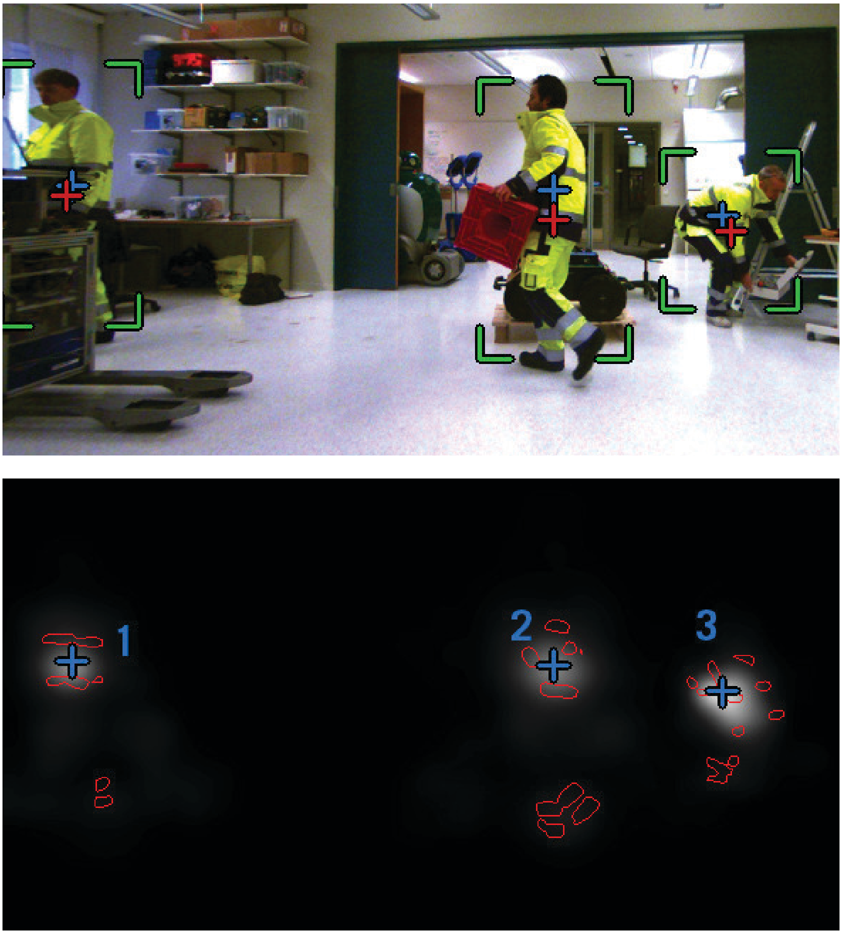

Figure 4.6: The figure illustrates the output obtained after applying the generalised Hough transform, backprojection and bounding box estimation. The top figure shows the input scene with three person occurrences, together with the detected object reference points (blue), corresponding to local maxima in the Hough space, estimated bounding box centres (red), and the final bounding boxes (green). The bottom figure visualises the Hough voting space, integrated over all the scale layers. The extracted local maxima (blue) and the reflector contours (red) are overlayed. The aspect ratios and centre positions of the respective bounding boxes are estimated according to the voting procedure illustrated in Figure 4.7.

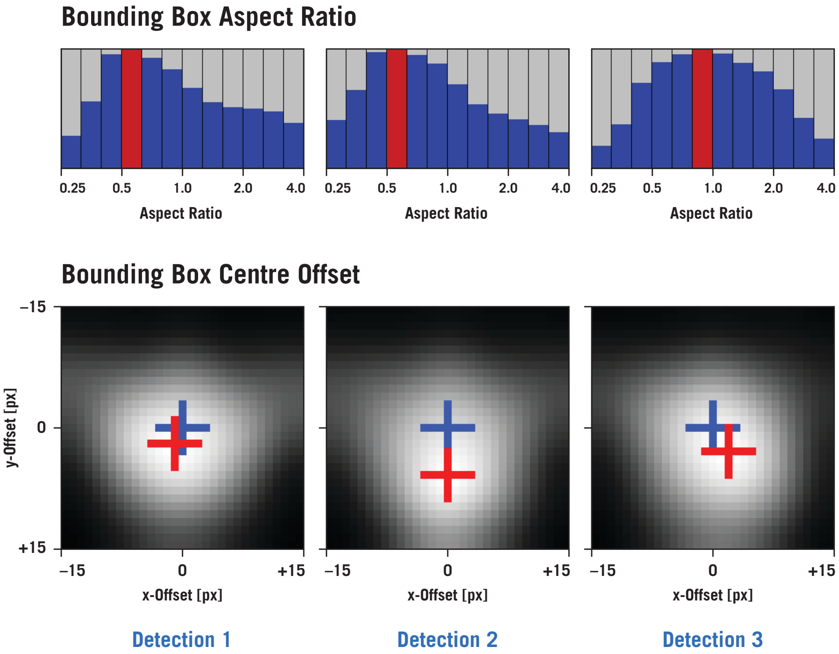

Figure 4.7: The figure illustrates the bounding box estimation process for the three detected persons shown in Figure 4.6. The subset of feature patches that, after backprojection, were found to have contributed to the detection of a person, are selected to vote for the aspect ratio of the respective bounding box (top) and its centre position with respect to the object reference point (bottom). A one-dimensional histogram with logarithmic binning is used for the aspect ratio, while voting for the bounding box centre position is done inside a two-dimensional histogram in scale-normalised pixel coordinates.

Chapter 5
Systems and Applications

This chapter presents an overview and discussion of the different variations and configurations of vision systems implemented in the course of Paper IV–Paper V. All systems are built on the basis of the sensors, models and algorithms described in Chapter 3 and 4. Even though the underlying sensing approach offers generic perception of arbitrary retro-reflective markers, the systems discussed in Section 5.1 focus on the task of detecting and localising industrial workers with high-visibility garments. All the system configurations have been evaluated on image sequences recorded in real-world industrial settings with the aim of demonstrating the practical relevance of the approach. Among the studied configurations feature monocular and stereoscopic setups relying on pure NIR vision (Section 5.1.1 and 5.1.2) as well as a setup with one RGB and two NIR cameras which fuses multiple spectral bands in a unified detection scheme (Section 5.1.3).

An alternative interesting application of the underlying research is illustrated in Section 5.2.1, where NIR stereo vision input was used for tracking the 3D pose of small-scale vehicles fitted with multiple retro-reflective markers. The example shows that the combination of proposed sensors and algorithms offer a more generic flexible toolbox for building computer vision systems with a range of interesting applications. Potential use cases include scenarios where a sensor system is required to robustly perceive retro-reflective markers under possibly difficult lighting conditions. Positioning systems on autonomous vehicles that are based on reflective beacons placed in the environment are a good example. Here, an active NIR vision sensor as presented in Chapter 3 can be a cost-effective alternative to a lidar based ranging device. Furthermore, the high resolution of the acquired sensor data allows for a more detailed analysis of the shape and geometry of reflectors than it is possible with common lidar systems. The proposed NIR sensing scheme is therefore particularly suitable if an application requires a specific analysis of reflector geometry.

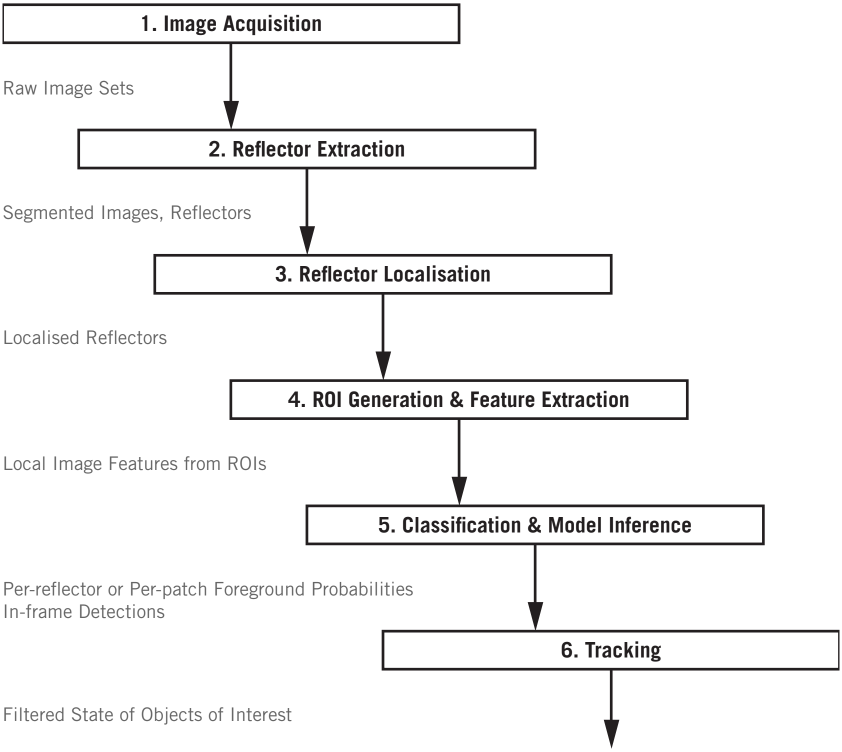
1. Image Acquisition
Raw Image Sets
2. Reflector Extraction
Segmented Images, Reflectors
3. Reflector Localisation
Localised Reflectors
4. ROI Generation & Feature Extraction
Local Image Features from ROIs
5. Classification & Model Inference
Per-reflector or Per-patch Foreground Probabilities
In-frame Detections
6. Tracking
Filtered State of Objects of Interest

Figure 5.1: The figure illustrates a general detection pipeline that all three proposed variations of vision systems are based upon.

5.1 Human Detection and Tracking

Within the scope of this thesis, the principal application of the sensors, models and algorithms proposed in Chapter 3 and 4 has been the detection and tracking of humans with retro-reflective workwear. In the course of the research, this problem was studied on the basis of three different hardware configurations with varying number and type of cameras. The study included monocular and stereoscopic NIR-only systems, as well as a setup with two NIR and one RGB camera that exploits several spectral bands in a more comprehensive sensing approach. Each configuration entails specific design choices and adaptions on the algorithmic level to make optimal use of the sensor data at hand. This section summarises and compares the multiple system implementations and discusses their respective advantages and limitations.

Figure 5.1 presents a generic system pipeline, which all three proposed system variations, by and large, are built upon. For each time frame, the processing chain departs from the synchronised acquisition of a set of multiple images, where the exact number of images depends on the hardware setup under consideration. Raw input images are undistorted to account for geometric lens distortions. Furthermore, in the case of multiple cameras, the images are rectified by projecting them on a common image plane, which reduces the correspondence problem arising when relating image regions from different cameras to a search in only one dimension.

Subsequent processing steps involve the extraction of image regions that depict reflective markers by relating images taken with and without active illumination, as described in Section 4.1. No assumption is made at this stage regarding the object family a given reflector originates from. Reflector extraction is followed by a depth estimation procedure which localises reflectors in a 3D space relative to the camera. This step involves either machine learning based depth regression in the case of monocular NIR, or more traditional computation of per-reflector disparity maps in the case of stereoscopic NIR vision.

Regions of interest (ROIs) covering a certain neighbourhood around the extracted reflective markers are then generated and features are computed individually within each ROI. The area covered by an ROI thereby depends on whether features are extracted from NIR data only, or, whether complementary information is sampled from the additional RGB input. On the basis of the computed features, a learned appearance model of industrial workers is applied. This step involves classification of local features in order to identify image regions that effectively depict a human worker, and discard ROIs generated through the presence of other reflective items in the environment. Furthermore, if a model of the spatial distribution of image features within the class of interest has been learned, such as a Hough forest, the model is applied for regressing the position of the reference point of individual person occurrences in the image space by means of a generalised Hough transform.

A final tracking layer then establishes and maintains temporal correspondences over a series of frames and recursively estimates the state of detected workers in terms of their 3D location and velocity relative to the camera unit.

5.1.1 Monocular NIR Vision

Monocular vision systems have gained high popularity in the field of pedestrian detection for road traffic scenarios [31]. Single-camera modules are compact and can therefore be seamlessly integrated in automotive applications. They further come at a lower cost than their multi-camera counterparts, consume less power, and involve less calibration effort.

Detection of industrial workers wearing high-visibility vests using an active monocular NIR vision device has been the major topic of the work presented in Paper I and Paper II. While the focus of Paper I concentrates on the estimation of depth from monocular NIR image features, Paper II describes a complete and low-cost single-camera system performing detection, classification, localisation and tracking of workers with retro-reflective vests. Figure 3.1a depicts the compact single-camera module deployed for the experimental evaluation. Equipped with an optical bandbass filter and 8 high-power infrared LEDs, the imaging device costs a mere 500€ and was shown to successfully perceive reflective markers up to 10 metres distance.

Figure 5.2 illustrates the temporal cycles of acquiring and processing the monocular image stream. Pairs of images are registered at a frequency of T_1^{-1}, where a first image I_{nf}, taken without NIR flash, serves as a reference image for the background illumination by ambient NIR light sources. A second image I_f is registered under active NIR illumination. The exposure time T_{NIR} is kept short (1–3 ms) to minimise the perceived amount of light emitted by potential other NIR light sources. On the other hand, the exposure window has to be long enough to saturate the reflectors in image I_f and allow for their clear discrimination from the background. The time delay T_2 between the acquisition of I_{nf} and I_f is kept as short as possible in order to minimise viewpoint changes under camera motion. The image pair is then processed together and enters the detection pipeline.

With respect to Figure 5.1, the system presented in Paper II represents a simplified version of the outlined scheme. In fact, step 5 (classification and model inference) only involves an appearance model that acts on a per-reflector level. A binary random forest classifier models the appearance of reflector families to be detected or discarded, while a random forest regressor learns the relation between specific reflector appearance and depth. No spatial distribution of the reflectors with respect to the human silhouette is learned. This simplified approach is only feasible if the retro-reflective markers on the garments under consideration are concentrated in one spot. The condition holds true for the typical sleeveless vests that are most frequently used in industry and feature two horizontal reflective stripes around the torso. The mapping from a set of observed reflectors to a set of individual persons is then handled by assigning reflectors to bounding boxes of individual tracked persons as described below.

Learning Depth from Appearance

One of the major challenges in the monocular sensing approach lies in the localisation of detected persons in 3D space and in thereby obtaining a notion of depth. In road traffic scenarios and other applications where a vehicle is consistently moving on even ground, monocular depth estimation is often performed by relying on spatial information about the ground plane and geometrical constraints that force detections to be located directly on the ground, such as in [75]. Another method observes several consecutive frames and obtains depth via structure from motion cues, such as in [30]. Due to the nature of the available sensor data, both approaches are not suitable for addressing

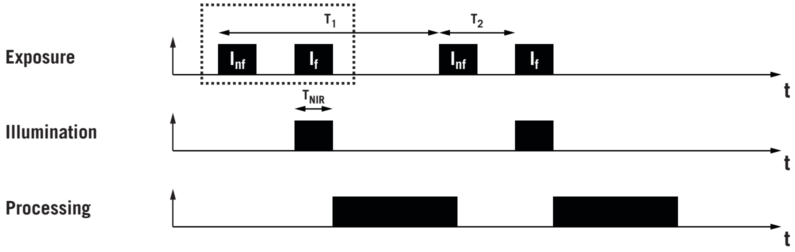

Figure 5.2: Temporal image acquisition and processing cycle for the monocular NIR vision system. Images are alternately taken with and without active illumination. Once an image pair is acquired, an image pair is processed together.

the problem at hand. As stated in Section 1.2, it is a fundamental requirement that the proposed solution is applicable to outdoor environments with potentially rough and uneven terrain. Trying to estimate a ground-plane and spatially constraining detections to according image regions is therefore not advisable. Persons located on terrain elevations or in cavities would easily be missed. In addition, the input images from the customised infrared camera lack any scene structure apart from the sparse set of reflective markers and therefore do not permit extraction of information about a ground-plan or scene structure.

Instead, it was shown in Paper I and Paper II that approximate depth estimates for individual reflective markers can be obtained by supervised learning from monocular NIR image data. To do so, a random forest regressor is trained on the same local appearance features that also serve for the discrimination of different families of reflective markers or objects. However, in contrast to the random forest classifier, the leaf-nodes of the decision trees no longer store a class distribution but the average depth estimate computed from the annotated training samples in a leaf-node. These entities are later used during detection to vote for the most probable object distance when a reflective pattern is observed. The machine learning based approach has been shown to estimate depth with an accuracy ranging between one to two decimetres at 2.5 metres and half a metre at 10 metres distance.

While monocular camera systems enjoy many advantages over multi-sensor configurations, the method comes with the distinct drawback of requiring a considerable amount of training samples labelled with ground-truth depth measurements. In Paper I and Paper II, this procedure was substantially simplified by acquiring the training material from a static sensor unit in an uncluttered environment and on even ground, so that the ground-truth locations of persons were conveniently extracted using a two-dimensional lidar in combination with simple statistical background subtraction methods.

Tracking

A tracking layer that takes individual reflectors as input, initialises tracking targets, and maintains their state over multiple frames has been discussed in Paper II for the case of a single person and in Paper III and Paper IV for multiple persons. The tracker by design incorporates all the detected reflectors together with their classification scores and distance estimates. By doing so, all the reflective items are being kept track of, regardless of the type of object they represent. This offers the advantage that a decision on whether an object is a person or not can be taken after integrating several classification scores over a series of frame and thus base the decision on more evidence.

The tracking layer is implemented using a particle filter based on the standard sequential importance resampling algorithm [43]. The particle filter performs recursive Bayesian estimation of a 6-dimensional state variable consisting of the 3-dimensional location and velocity of a tracked target. Two important entities that need to be defined in the particle filter are the state transition (or motion) model and the measurement model. The measurement model defines the probability of making a certain observation given a state. In the particular case of the proposed monocular system, where depth is inferred from learning, the measurement model needs to take into account that the accuracy of position estimates for individual reflectors is significantly lower in the depth dimension than in the lateral (horizontal and vertical) dimension.

Defining an appropriate model for the motion of the tracked targets is difficult in the underlying case. Both the camera and the observed targets may or may not be in motion. No information is assumed available with regard to the vehicle motion and inferring such information from visual odometry is no option here given the type of NIR images produced by the camera. A common solution in such cases it to use a constant velocity model in combination with a relatively large process noise that accounts for unmodelled velocity changes.

5.1.2 Stereoscopic NIR Vision

As described in the previous section, monocular depth estimation by means of supervised learning requires a large amount of annotated training samples. In particular, if the algorithm must simultaneously handle several types of protective garments with different reflective patterns, the acquisition and annotation of these training samples can prove cumbersome. Therefore, if the compactness and the manufacturing cost of the camera module are not the primary factors, it is advisable to opt for stereoscopic NIR input. In this configuration, depth estimates are obtained by computing a disparity map from the input images of two NIR cameras. As detailed in Paper III and Paper IV, disparity is not densely computed over the entire image but efficiently restricted to local areas where a reflective marker indicates the potential presence of a person. The 3D locations of reflectors are then inferred by triangulation from the disparities.

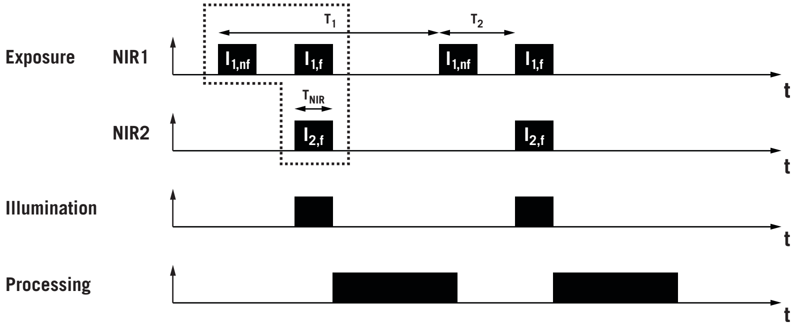

Figure 5.3: Temporal image acquisition and processing cycle for the stereoscopic NIR vision system. Camera NIR1 alternately takes images with and without active illumination which permits identifying reflective interest regions. The images captured by camera NIR2 then serve as complementary input for depth estimation through triangulation.

For the work in Paper III and Paper IV, the camera module was therefore enhanced by adding a second NIR camera of identical type. The respective unit is depicted in Figure 3.1c. Figure 5.3 illustrates the temporal acquisition and processing cycle adapted to the stereoscopic configuration. As in the monocular case, one camera (NIR1) continuously acquires pairs of images without and with active illumination ($I_{1,nf}$ and $I_{1,f}$). In addition, a second camera (NIR2) acquires images under active illumination ($I_{2,f}$), synchronised with image $I_{1,f}$. The image triplets then build a set of input images for the detection pipeline described in detail in Paper III and Paper IV. Images $I_{1,nf}$ and $I_{1,f}$ are thereby related to identify reflective markers in the image, while images $I_{1,f}$ and $I_{2,f}$ are related for computing a local disparity map in the neighbourhood of reflective interest regions and estimating the depth of each reflector.

For a comparison between monocular and stereoscopic depth estimation, Figure 5.4 quantifies the estimation error for both methods on a per-reflector basis. It is evident from the plot that the monocular, learning based approach neither yields the same accuracy nor the same precision as conventional triangulation from stereo images. While the error for monocular depth estimation is in the range of several decimetres, stereo-based estimation provides depth at an accuracy which is nearly an order of magnitude higher. In both cases, the measurements become less accurate and less precise for increasing distances from the sensor module.

At the tracking stage, the same approach can be adopted that was described in the context of the monocular system. To do so, the measurement model of the particle filter is adapted to take into account that the uncertainty in the depth dimension is lower in the underlying stereoscopic case with respect to the previously discussed case of monocular depth estimation.

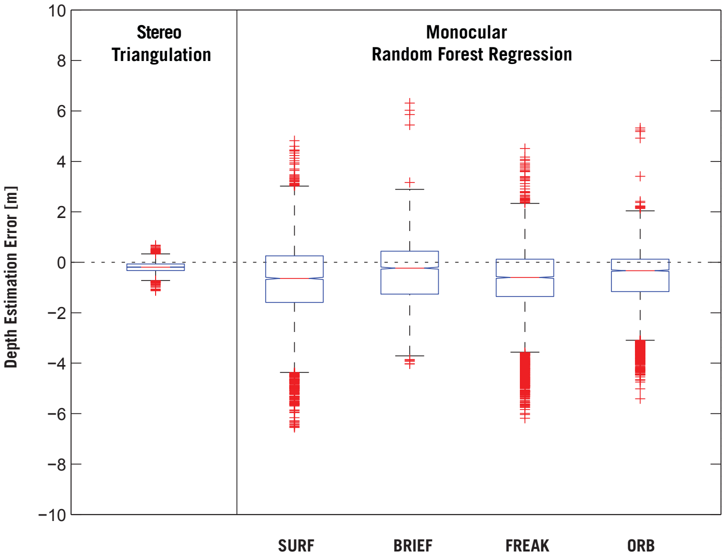

Figure 5.4: The figure quantifies the depth estimation error on a per-reflector level in the form of box plots. The first column depicts the error in estimating depth via triangulation from stereo NIR images, while the remaining columns indicate the error for depth estimation from random forest regression in combination with different popular image feature descriptors. The plots indicate that stereo triangulation yields accuracy and precision which are around an order of magnitude higher than the ones obtained with the monocular learning based approach.

5.1.3 Combined NIR and RGB Vision

A key advantage of the two previous NIR-only systems is their robustness towards various lighting scenarios. In fact, low-light conditions have no negative effect whatsoever on detector performance due to the active sensing principle, while the effect of glare is highly reduced by the optical narrow-band filter of the NIR sensor. Furthermore, focusing the computational efforts on spatially limited reflective interest regions proves much more efficient than a full image analysis as it is done in many vision applications.

However, by solely analysing the appearance of retro-reflective markers in the NIR data, the system is not exploiting the potential lying in other feature cues that can be helpful for discriminating reflectors of different families and detecting and locating individual persons in an image. If lighting conditions can be assumed within reasonable bounds, an additional visible-light camera

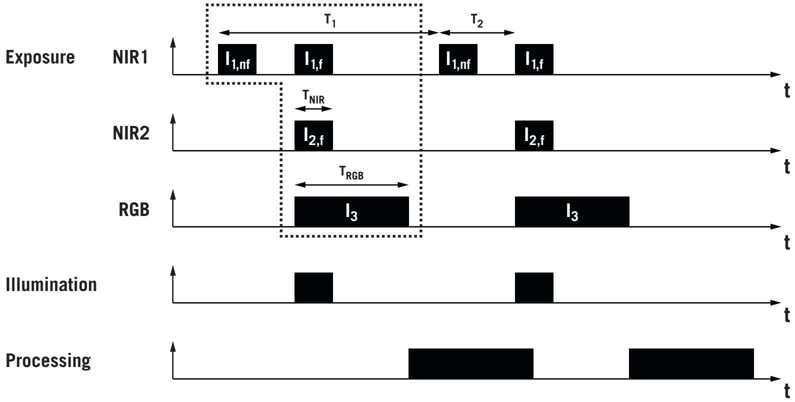

Figure 5.5: Temporal image acquisition and processing cycle for the combined stereo NIR and RGB vision system. Camera NIR1 serves the purpose of extracting reflective interest regions, while the additional input from camera NIR2 permits to localise reflectors in 3D space using triangulation. Image features are finally sampled from corresponding regions in the NIR and RGB cameras in order to apply the learned multi-band Hough forest appearance model and identify individual person occurrences in the images.

can provide a valuable complementary source of information for the task at hand. Protective high-visibility clothing is most often fabricated fully or partially from material with bright fluorescent colours. Sensing the characteristic colour properties of the garments with help of an RGB camera can therefore further help to discriminate the object class form the background. Furthermore, gradient patterns related to the human silhouette can be observed in RGB or grayscale images. Such features are entirely filtered out from the data provided by the customised NIR setup which depicts reflective markers exclusively.

Paper V therefore investigated the fusion of NIR and RGB data in a multi-spectral appearance model according to the multi-band Hough forest discussed in Section 4.3.2. The respective camera module (see Figure 3.1c) features a stereo NIR camera with a baseline of 20 cm and an RGB camera located in the centre. Figure 5.5 shows the temporal acquisition and processing cycles adapted for the NIR+RGB sensing scheme. A significantly longer exposure time is needed for acquiring the additional RGB images. It has thereby proven convenient to use a camera with automatic exposure control to ensure that the acquired images are not perturbed by small changes in lighting.

The algorithm proposed in Paper V is able to detect industrial workers and localise their respective centre points in the image space by performing a generalised Hough transform. In addition, no fixed bounding box aspect ratio is assumed but the entity is estimated during an additional voting step. The approach thereby accounts for the fact that persons in industrial working sce-

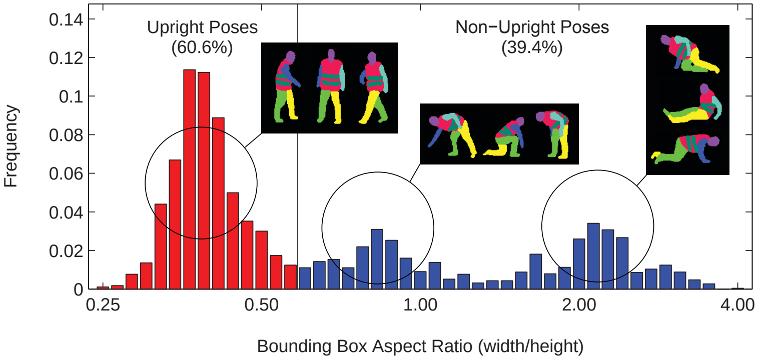

Figure 5.6: Distribution of person annotations in the test sequences of PAPER V with respect to the bounding box aspect ratio.

narios cannot simply be assumed upright, which adds another dimension to the search space if meaningful bounding box detections are to be produced. Obtaining an estimate of the bounding box aspect ratio can also help inferring information regarding the body position of a detected person. Figure 5.6 therefore illustrates the relationship between bounding box aspect ratio and body position and shows the distribution of person annotations in the real-world test sequences evaluated in of PAPER V.

Figure 5.7 shows two of the main results obtained from the scientific evaluation of the approach. The curves illustrate first and foremost that performance on upright pedestrians is significantly higher than on non-upright person occurrences. This observation is in full accordance with conclusions drawn from many other experiments conducted in the field of vision-based human detection, where body articulation and a high degree of body pose variability are often named as particularly challenging problems. Figure 5.7 (top) further shows that the proposed fusion of RGB and NIR images improves detector performance compared to the sole use of NIR data. However, the difference is shown to be modest. In fact, using precision and recall at the equal error rate (EER) as a comparison measure, NIR+RGB vision outperforms pure NIR vision by only 3%. Figure 5.7 (bottom) finally indicates that the proposed additional supervision criterion based on pixel-wise body part labels, introduced in the training procedure of the Hough forest, leads to a further consistent improvement in precision and recall at the equal error rate in the range of approximately 4%.

Nevertheless, it cannot be generally concluded that the benefit from the multi-band sensing approach is minor. The extent to which the additional RGB vision benefits the detection task depends on the type of garments in use. It is illustrated in Figure 5.8 that in the the specific work environment in which the experimental evaluation took place, the working personnel is equipped with reflective vests as the sole item of high-visibility clothing. In contrast, arms and legs are most often covered by dark clothing which in front of an also frequently dark background fails to provide significant contrast and gradient information. The additional evidence for the presence of a human, gained from sampling features patches from regions in the RGB image that depict these body parts, is therefore rather limited. However, if the personnel is equipped with garments such as in Figure 4.1, the amount of additional information contained in the RGB images is more significant and a higher benefit can be expected.

Tracking

At the tracking stage, a particular challenge that is generally encountered is the association of in-frame detections with individual targets that are currently being tracked. The human appearance model discussed in Section 4.3 which spatially connects local image feature patches to a defined reference point of the object class, significantly facilitates this data association issue. In fact, while an association between observed reflectors and tracked targets was previously handled on the tracking level, the procedure has been moved back to the detection stage by adopting the Hough forest detector. In higher-level terms, the Hough forest maps the evidence observed inside the scanned regions of interest to the two-dimensional locations where the reference point of the detected object is believed to be located. While both the observed reflectors in the NIR image as well as the various parts of the articulated human body observed in the RGB space may be moving significantly over a series of frames, the reference point locations produced by the Hough forest offer a much more stable indication of where the detected object is located in the image. As a result, detections are less ambiguously associated with a tracked target. While Paper V purely focused on the detection stage, it was later shown that a tracking layer can be implemented using a simple Kalman filter that continuously incorporates the in-frame detections provided by the Hough forest detector.

5.1.4 The Role of the Protective Garments

The type of protective garments worn by workers plays a significant role for the performance at certain stages in the detection pipeline. The largest part of experimental evaluation has been carried out in work environments where the personnel is equipped with simple sleeveless yellow vests that feature two horizontal retro-reflective stripes around the torso. This type of garment has shown to be the most frequent choice in many of the addressed industrial environments. However, the experiments in Paper V have shown that the correct estimation of the bounding box aspect ratio is difficult, especially if arms and legs are covered by black clothes and the background is dark. In fact, most missed detections were correctly localised by the multi-band Hough forest, but an erroneous estimation of the bounding box aspect ratio prevented a successful matching of detected and annotated bounding boxes.

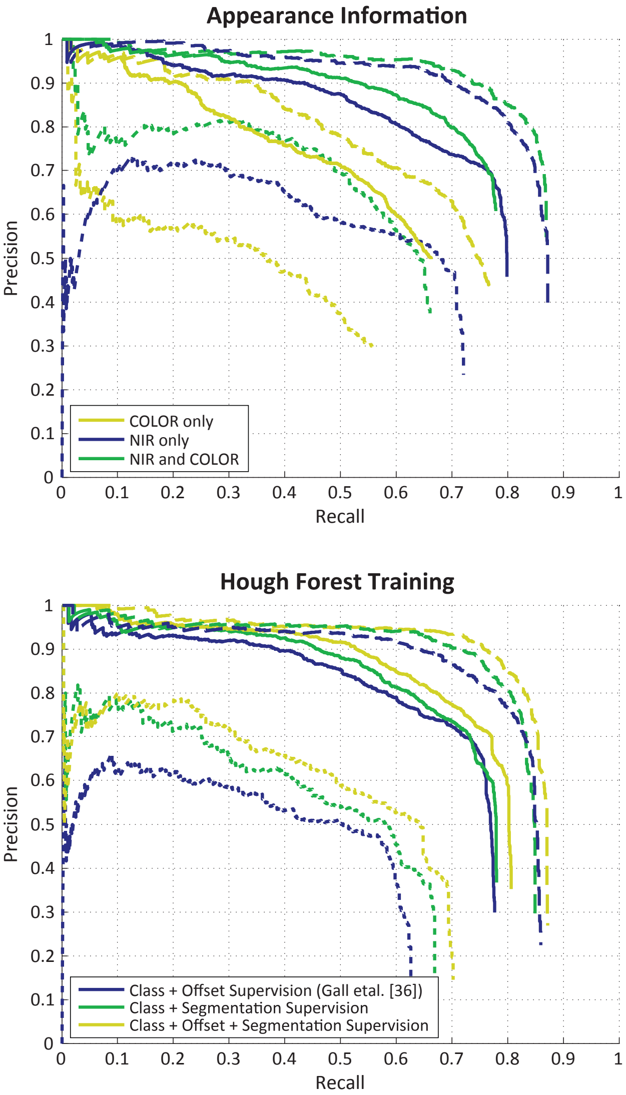

Figure 5.7: The figures show detector performance in the form of precision-recall curves obtained during the evaluation of the combined NIR and RGB vision system presented in Paper V. Solid lines show overall performance, dashed lines performance on upright, and dotted lines performance on non-upright person occurrences.

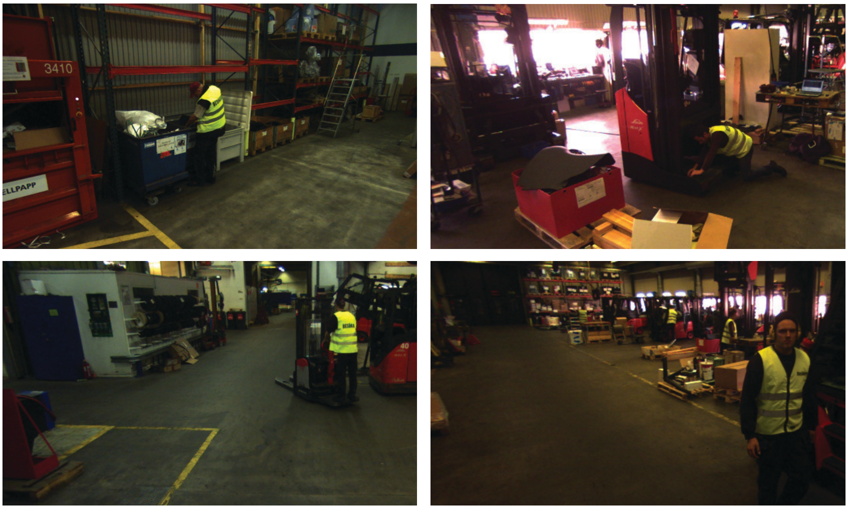

Figure 5.8: The figure shows several examples from the image sequences evaluated in the course of Paper V. The fact that most workers are wearing reflective vests and black trousers, leads to only a slight improvement in detector performance if additional RGB features are used during detection.

Further experiments not included in Paper I–V have shown that the estimation of the bounding box aspect ratio as the final step of a generalised Hough voting procedure provides significantly better results if a person is equipped with both upper- and lower-body garments, such as in Figure 4.3. In this case, it is easily observed that the individual feature channels computed from both NIR and RGB images contain a much higher amount of distinctive information than if a simple vest is worn on top of black clothing.

5.2 Alternative Applications

The presented vision system lends itself exceptionally well for the the task addressed in this thesis, namely the detection of industrial workers equipped with retro-reflective garments. For this purpose, an NIR sensitive camera was specifically enhanced with the aim of facilitating the perception of reflective markers and reducing the complexity of the required image processing methods. The resulting camera configuration is therefore suitable for a broader range of applications where a system needs to be able to sense and localise reflective markers that are deliberately placed in a working environment. Such scenarios may include, among others, on-board mapping and localisation tasks from autonomous vehicles, but also on- or off-board localisation and pose estimation of vehicles for fleet management and vehicle coordination in logistics scenarios.

5.2.1 Vehicle Pose Tracking

The NIR stereo vision input delivered by the camera module in Figure 3.1d has been successfully exploited for estimating the 3D pose of several smaller-sized indoor vehicles. The term pose hereby refers to an object's position and orientation in a given coordinate system. For estimating pose from NIR images, a vehicle is fitted with multiple retro-reflective markers on the outer sides. The pattern of the reflectors is chosen such that the vehicle pose can be unambiguously inferred from reflector geometry only. The pose tracking algorithm is further provided with a simple vehicle model defining the three-dimensional position and orientation of each retro-reflective marker with respect to a defined reference point on the vehicle.

Pose estimation then starts with the extraction and analysis of retro-reflective markers from stereoscopic NIR images, by the same means as described in Chapter 4. Binary reflector classification based on BRIEF [20] image feature descriptors has shown to successfully discriminate reflectors on vehicles and reflectors on the personnel's safety garments. To infer the vehicle pose, a particle filter is employed that maintains a large set of hypothesised 4-dimensional vehicle poses that are continuously evaluated and updated with incoming measurements. The state space is restricted to four degrees of freedom by assuming that the observing camera remains static and that the observed vehicle is moving on a planar surface. A measurement model then computes the expected visible reflector pattern for each hypothesised pose in the particle filter and compares the result with the set of actually observed reflectors extracted from the NIR images. Simple nearest-neighbour matching of expected and observed sets of reflectors finally allows the system to efficiently compute particle weights as the sum of absolute differences between matched reflector pairs. Two examples of estimated 3D vehicle poses are shown in Figure 5.9. The approach has been tested on relatively simple vehicle layouts, and more research is necessary to study its applicability to larger and potentially articulated vehicles.

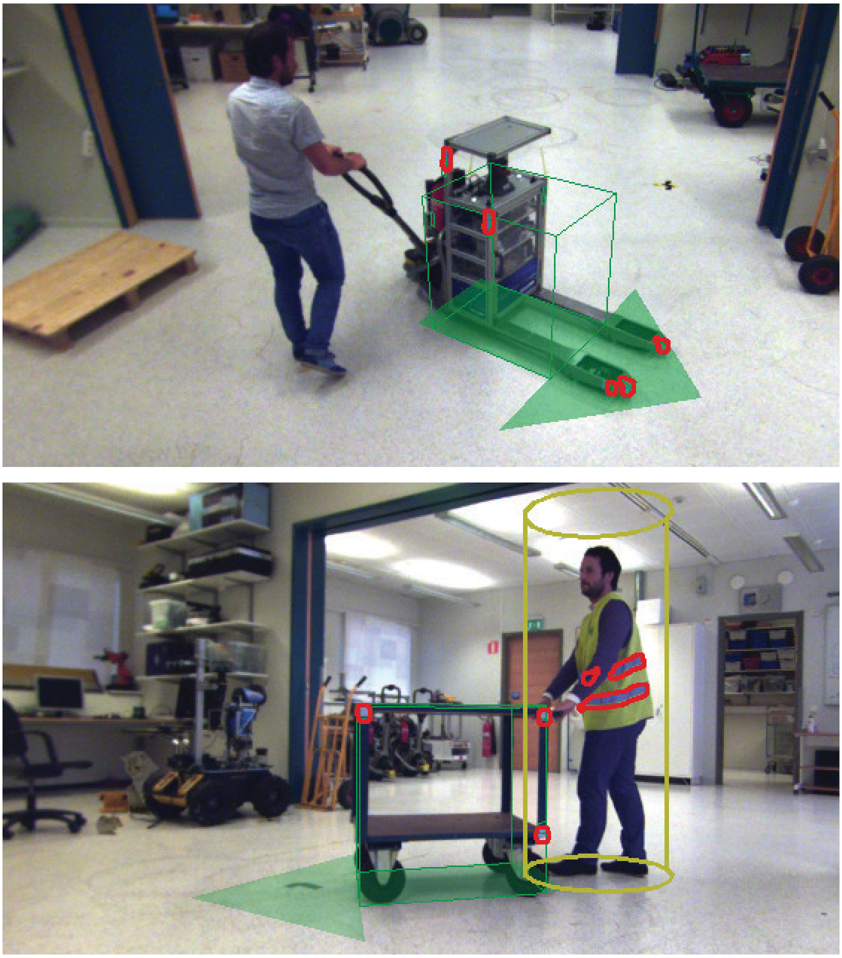

Figure 5.9: The figure shows examples of the output of a vehicle pose tracker (top) and a combined vehicle and person tracking system (bottom). In both cases, detection is entirely based on processing stereo NIR images and vehicle pose is inferred by matching observed reflective patterns with a simple vehicle model.

Chapter 6
Conclusion and Future Work

This thesis investigated a novel method for detecting human workforce from mobile industrial machinery with an active NIR vision system and optional complementary RGB input. A collection of algorithms that handle the extraction of reflective markers from infrared images, their description and classification as well as their association with individual person occurrences were proposed throughout multiple scientific articles. In this thesis, the underlying sensors, methods and algorithms were summarised and compared. Furthermore, the advantages and limitations of three concrete system implementations have been discussed, covering monocular and stereoscopic infrared-only configurations as well as a multi-spectrum approach fusing NIR and RGB information in a combined appearance model.

Targeting industrial environments where protective garments are a de facto standard, it was shown that building a detection pipeline around a reflector sensing scheme offers an efficient approach as computational resources are concentrated on processing spatially limited interest regions in the image. Due to the narrow NIR filter band which excludes the vast amount of ambient lighting, the method was shown to be robust to a wide range of illumination and lighting conditions. The approach was first evaluated on the basis of pure NIR vision input. There, a particular challenge was the fact that the protective clothing of workers is not the only family of reflective items commonly observed at industrial work sites. A closer analysis and discrimination of different groups of reflectors observed in the NIR images was therefore necessary to delimit the number of false positives under the presence of reflective items. To exploit additional feature cues from RGB images in the detection process, the method was extended and the learning of a unified human appearance model based on multi-band Hough forests was suggested. Based on local NIR-RGB feature patches connected in a star-shaped model, it represents a convenient way of fusing multiple spectral bands without the need for explicitly defining feature selection criteria. The task of evaluating the most discriminative feature channels for structuring a given set of feature patches is rather left to the randomised feature selection mechanism of the Hough forest framework. Furthermore, Hough forests handle considerable amounts of body articulation without the need for computationally expensive models.

Keeping a strong focus on industrial applicability, the approach has been consistently evaluated on video sequences recorded from several different types of industrial vehicles and in multiple realistic work environments. The image material features a broad range of indoor and outdoor scenarios with strongly varying and partly very challenging lighting conditions. Person occurrences include both upright standing and walking persons as well as various workers carrying out tasks in challenging non-upright positions that typically pose a significant challenge in vision-based human detection. Due to the particular nature of the acquired sensor data as well as the authenticity of the depicted scenes, the recorded and evaluated data sets are unique and no similar data sets have been made publicly available by other authors in the research field. For reasons of confidentiality, the acquired data sets cannot be made fully public but are available on individual request.

Like every sensor technology, the proposed method also has a number of limitations. Most notably, it is emphasised that the suggested vision system is not suitable for environments where it cannot be assumed with high certainty that the persons are equipped with reflective high-visibility clothing. The obvious reason is that the entire sensing approach by design requires people to wear garments with retro-reflective markers. In the vast majority of harsh work environments, employers impose strict safety regulations regarding the protective equipment employees are required to wear in order to access a work site. However, there still exist considerable geographical differences in the consistency of and adherence to work place safety policies, which could in certain industrial sectors hinder a full reliance on the proposed sensor technology. Even though an extension of the method for fusing information from RGB and NIR data was presented, the detection cue is still triggered only after successful perception of at least one reflective marker. As a result, the vision system will entirely fail to detect the presence of a person without protective clothing, if not combined with complementary detection methods.

Future research will involve identifying and investigating further application scenarios where the robust perception of retro-reflective markers under potentially varying lighting conditions is a crucial requirement or facilitates the achievement of a task. An exemplary candidate domain is the increasing automation in construction and logistics, where a new market for autonomous vehicles and machinery is gradually developing. Intelligent vehicle coordination and fleet management therefore become important tasks which require that different vehicles are aware of each other's presence. The vehicle pose estimation scheme outlined in Section 5.2.1, that makes use of stereoscopic NIR vision and specific patterns of retro-reflective markers fitted to a target vehicle, can represent a promising step to in that direction.

References

[1] M. Agrawal, K. Konolige, and M. R. Blas. CenSurE: Center surround extremas for realtime feature detection and matching. In *European Conference on Computer Vision (ECCV)*, volume 5305 of *Lecture Notes in Computer Science*, pages 102–115. Springer, 2008.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst. FREAK: Fast retina keypoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 510–517, 2012.

[3] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[4] L. Andreone, F. Bellotti, A. De Gloria, and R. Lauletta. SVM-based pedestrian recognition on near-infrared images. In *Proceedings of the International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 274–278, 2005.

[5] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, 110:346–359, 2008.

[7] M. Bertozzi, A. Broggi, C. Caraffi, M. Del Rose, M. Felisa, and G. Vezzoni. Pedestrian detection by means of far-infrared stereo vision. *Computer Vision and Image Understanding*, 106(2–3):194–204, 2007. Special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[8] B. Besbes, A. Rogozan, A.-M. Rus, A. Bensrhair, and A. Broggi. Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF. *Sensors*, 15(4):8570–8594, 2015.

[9] G. Blumrosen, B. Fishman, and Y. Yovel. Noncontact wideband sonar for human activity detection and classification. *IEEE Sensors Journal*, 14(11):4043–4054, 2014.

[10] C. Böddeker, M. A. Mohamed, and B. Mertsching. Detection and tracking of construction workers and equipment. *Bildverarbeitung in der Automation*, 2014.

[11] P. V. K. Borges, R. Zlot, and A. Tews. Integrating off-board cameras and vehicle on-board localization for pedestrian safety. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):720–730, 2013.

[12] J.-Y. Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm. *Intel Corporation Microprocessor Research Labs*, 2000.

[13] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[14] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[15] A. Broggi, R. L. Fedriga, and A. Tagliati. Pedestrian detection on a moving vehicle: An investigation about near infra-red images. In *IEEE Intelligent Vehicles Symposium*, pages 431–436, 2006.

[16] M. Bui, V. Fremont, D. Boukerroui, and P. Letort. People detection in heavy machines applications. In *IEEE Conference on Cybernetics and Intelligent Systems (CIS)*, pages 18–23, 2013.

[17] M. Bui, V. Fremont, D. Boukerroui, and P. Letort. Deformable parts model for people detection in heavy machines applications. In *13th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 389–394, 2014.

[18] M.-T. Bui, V. Fremont, D. Boukerroui, and P. Letort. Multi-sensors people detection system for heavy machines. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 867–872, Qingdao, China, 2014.

[19] L. Cai and J. Qian. A method for detecting miners based on helmets detection in underground coal mine videos. *Mining Science and Technology (China)*, 21(4):553–556, 2011.

[20] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *Proceedings of the European conference on Computer Vision (ECCV)*, volume 6314, pages 778–792, 2010.

[21] J. L. Carr, C.C. Jobes, and J. Li. Development of a method to determine operator location using electromagnetic proximity detection. In *IEEE International Workshop on Robotic and Sensors Environments (ROSE)*, pages 1–6, 2010.

[22] S. Chang, N. Mitsumoto, and J. W. Burdick. An algorithm for UWB radar-based human detection. In *Radar Conference, 2009 IEEE*, pages 1–6, 2009.

[23] P. Corke, J. Roberts, J. Cunningham, and D. Hainsworth. Mining robotics. In *Springer Handbook of Robotics*, pages 1127–1150. Springer Berlin Heidelberg, 2008.

[24] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[25] A. D. Costea, A. Vatavu, and S. Nedevschi. Obstacle localization and recognition for autonomous forklifts using omnidirectional stereovision. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 531–536, 2015.

[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[27] J. S. Dickens, M. A. van Wyk, and J. J. Green. Pedestrian detection for underground mine vehicles using thermal images. In *Proceedings of IEEE Africon Conference*, 2011.

[28] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–311, 2009.

[29] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[30] M. Enzweiler, P. Kanter, and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. In *IEEE Intelligent Vehicles Symposium*, pages 792–797, 2008.

[31] Markus Enzweiler and Dariu M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:2179–2195, 2009.

[32] A. Fackelmeier, C. Morhart, and E. Biebl. Dual frequency methods for identifying hidden targets in road traffic. In *Advanced Microsystems for Automotive Applications 2008*, VDI-Buch, pages 11–20. Springer Berlin Heidelberg, 2008.

[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[34] A. Fernández-Caballero, M. C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás. Optical flow or image subtraction in human detection from infrared camera on mobile robot. *Robotics and Autonomous Systems*, 58(12):1273–1281, 2010. Intelligent Robotics and Neuroscience.

[35] American Society for Testing and Materials. *Standard Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37°Tilted Surface*. ASTM International, 2003.

[36] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(11):2188–2202, 2011.

[37] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[38] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007.

[39] J. Ge, Y. Luo, and G. Tei. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298, 2009.

[40] D. Geronimo, A. M Lopez, A. D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1239–58, 2010.

[41] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine. Pedestrian detection and tracking in an urban environment using a multilayer laser scanner. *Intelligent Transportation Systems, IEEE Transactions on*, 11(3):579–588, 2010.

[42] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. In *IEEE International Conference on Computer Vision (ICCV)*, pages 81–88, 2011.

[43] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, 1993.

[44] M. Haselich, B. Jobgen, N. Wojke, J. Hedrich, and D. Paulus. Confidence-based pedestrian tracking in unstructured environments using 3D laser distance measurements. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4118–4123, 2014.

[45] T. Heimonen and J. Heikkilä. A human detection framework for heavy machinery. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 416–419, 2010.

[46] S. Heuel and H. Rohling. Two-stage pedestrian classification in automotive radar systems. In *Proceedings of the International Radar Symposium (IRS)*, pages 477–484, 2011.

[47] M. Heuer, A. Al-Hamadi, A. Rain, M.-M. Meinecke, and H. Rohling. Pedestrian tracking with occlusion using a 24 GHz automotive radar. In *Proceedings of the International Radar Symposium (IRS)*, pages 1–4. IEEE, 2014.

[48] T. K. Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, 1995.

[49] C. Jobes, J. Carr, J. DuCarme, and J. Patts. Determining proximity warning and action zones for a magnetic proximity detection system. In *IEEE Industry Applications Society Annual Meeting (IAS)*, pages 1–7, 2011.

[50] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura. Pedestrian recognition using high-definition lidar. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 405–410, 2011.

[51] J. Koch, J. Wettach, E. Bloch, and K. Berns. Indoor localisation of humans, objects, and mobile robots with rfid infrastructure. In *Hybrid Intelligent Systems, 2007. HIS 2007. 7th International Conference on*, pages 271–276, 2007.

[52] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008.

[53] S. Leutenegger, M. Chli, and R.Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2548–2555, 2011.

[54] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[55] Y. Luo, J. Remillard, and D. Hoetzer. Pedestrian detection in near-infrared night vision system. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 51–58, 2010.

[56] V. Milanés, D. F. Llorca, J. Villagrá, J. Pérez, I. Parra, C. González, and M. A. Sotelo. Vision-based active safety system for automatic stopping. *Expert Systems with Applications*, 39(12):11234–11242, 2012.

[57] J. Miseikis and P. V. K. Borges. Joint human detection from static and mobile cameras. *Intelligent Transportation Systems, IEEE Transactions on*, 16(2):1018–1029, 2015.

[58] M. Moebus, A.M. Zoubir, and M. Viberg. Parametrization of acoustic images for the detection of human presence by mobile platforms. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 3538–3541, 2010.

[59] S. Montabone and A. Soto. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*, 28(3):391–402, 2010.

[60] F. Oleari, M. Magnani, D. Ronzoni, and L. Sabattini. Industrial AGVs: Toward a pervasive diffusion in modern factory warehouses. In *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 233–238, 2014.

[61] N. Onkarappa and A. D. Sappa. On-board monocular vision system pose estimation through a dense optical flow. In *Image Analysis and Recognition*, volume 6111 of *Lecture Notes in Computer Science*, pages 230–239. Springer Berlin Heidelberg, 2010.

[62] M.-W. Park and I. Brilakis. Construction worker detection in video frames for initializing vision trackers. *Automation in Construction*, 28:15–25, 2012.

[63] R. H. Rasshofer, D. Schwarz, E. Biebl, C. Morhart, O. Scherf, S. Zecha, R. Grünert, and H. Frühauf. Pedestrian protection systems using cooperative sensor technology. In *Advanced Microsystems for Automotive Applications*, VDI-Buch, pages 135–145. Springer Berlin Heidelberg, 2007.

[64] H. Ritter and H. Rohling. Pedestrian detection based on automotive radar. In *Proceedings of the IET International Conference on Radar Systems*, pages 1–4, 2007.

[65] T. Ruff and D. Hession-Kunz. Application of radio-frequency identification systems to collision avoidance in metal/nonmetal mines. *Industry Applications, IEEE Transactions on*, 37(1):112–116, 2001.

[66] L. Sabattini, V. Digani, C. Secchi, G. Cotena, D. Ronzoni, M. Foppoli, and F. Oleari. Technological roadmap to boost the introduction of agvs in industrial applications. In *2013 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 203–208, 2013.

[67] S. Sato, M. Hashimoto, M. Takita, K. Takagi, and T. Ogawa. Multilayer lidar-based pedestrian tracking in urban environments. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 849–854, 2010.

[68] W. H. Schiffbauer. Active proximity warning system for surface and underground mining applications. *Mining Engineering*, 54(12):40–48, 2012.

[69] H. Son, C. Kim, H. Kim, S. Han, and M. Kim. Trend analysis of research and development on automation and robotics technology in the construction industry. *KSCE Journal of Civil Engineering*, 14(2):131–139, 2010.

[70] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In *IEEE Intelligent Vehicles Symposium*, pages 206–212, 2006.

[71] T. Teixeira, G. Dublon, and A. Savvides. A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. *ACM Computing Surveys*, 5:427–450, 2010.

[72] J. Teizer. Magnetic field proximity detection and alert technology for safe heavy construction equipment operation. In *Proceedings of the 32nd International Symposium on Automation and Robotics in Construction and Mining (ISARC)*, 2015.

[73] J. Teizer, B. S. Allread, C. E. Fullerton, and J. Hinze. Autonomous proactive real-time construction worker and equipment operator proximity safety alert system. *Automation in Construction*, 19(5):630–640, 2010.

[74] P. Vähä, T. Heikkilä, P. Kilpeläinen, M. Järviluoma, and E. Gambao. Extending automation of building construction: Survey on potential sensor technologies and robotic applications. *36*, pages 168–178, 2013.

[75] A. Wedel, U. Franke, J. Klappstein, T. Brox, and D. Cremers. Realtime depth estimation and obstacle detection from monocular video. In *Pattern Recognition*, volume 4174 of *Lecture Notes in Computer Science*, pages 475–484. Springer Berlin Heidelberg, 2006.

[76] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li. Robust multi-resolution pedestrian detection in traffic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3033–3040, 2013.

[77] L. Yang and N Noguchi. Human detection for a robot tractor using omni-directional stereo vision. *Computers and Electronics in Agriculture*, 89:116–125, 2012.

[78] A. Yao, D. Uebersax, J. Gall, and L. Van Gool. Tracking people in broadcast sports. In *Pattern Recognition*, volume 6376 of *Lecture Notes in Computer Science*, pages 151–161. Springer Berlin Heidelberg, 2010.

Publications *in the series* Örebro Studies in Technology

1. Bergsten, Pontus (2001) *Observers and Controllers for Takagi – Sugeno Fuzzy Systems*. Doctoral Dissertation.
2. Iliev, Boyko (2002) *Minimum-time Sliding Mode Control of Robot Manipulators*. Licentiate Thesis.
3. Spännar, Jan (2002) *Grey box modelling for temperature estimation*. Licentiate Thesis.
4. Persson, Martin (2002) *A simulation environment for visual servoing*. Licentiate Thesis.
5. Boustedt, Katarina (2002) *Flip Chip for High Volume and Low Cost – Materials and Production Technology*. Licentiate Thesis.
6. Biel, Lena (2002) *Modeling of Perceptual Systems – A Sensor Fusion Model with Active Perception*. Licentiate Thesis.
7. Otterskog, Magnus (2002) *Produktionstest av mobiltelefonantenner i mod-växlande kammare*. Licentiate Thesis.
8. Tolt, Gustav (2003) *Fuzzy-Similarity-Based Low-level Image Processing*. Licentiate Thesis.
9. Loutfi, Amy (2003) *Communicating Perceptions: Grounding Symbols to Artificial Olfactory Signals*. Licentiate Thesis.
10. Iliev, Boyko (2004) *Minimum-time Sliding Mode Control of Robot Manipulators*. Doctoral Dissertation.
11. Pettersson, Ola (2004) *Model-Free Execution Monitoring in Behavior-Based Mobile Robotics*. Doctoral Dissertation.
12. Överstam, Henrik (2004) *The Interdependence of Plastic Behaviour and Final Properties of Steel Wire, Analysed by the Finite Element Metod*. Doctoral Dissertation.
13. Jennergren, Lars (2004) *Flexible Assembly of Ready-to-eat Meals*. Licentiate Thesis.
14. Jun, Li (2004) *Towards Online Learning of Reactive Behaviors in Mobile Robotics*. Licentiate Thesis.
15. Lindquist, Malin (2004) *Electronic Tongue for Water Quality Assessment*. Licentiate Thesis.
16. Wasik, Zbigniew (2005) *A Behavior-Based Control System for Mobile Manipulation*. Doctoral Dissertation.

17. Berntsson, Tomas (2005) *Replacement of Lead Baths with Environment Friendly Alternative Heat Treatment Processes in Steel Wire Production.* Licentiate Thesis.

18. Tolt, Gustav (2005) *Fuzzy Similarity-based Image Processing.* Doctoral Dissertation.

19. Munkevik, Per (2005) *"Artificial sensory evaluation – appearance-based analysis of ready meals"*. Licentiate Thesis.

20. Buschka, Pär (2005) *An Investigation of Hybrid Maps for Mobile Robots.* Doctoral Dissertation.

21. Loutfi, Amy (2006) *Odour Recognition using Electronic Noses in Robotic and Intelligent Systems.* Doctoral Dissertation.

22. Gillström, Peter (2006) *Alternatives to Pickling; Preparation of Carbon and Low Alloyed Steel Wire Rod.* Doctoral Dissertation.

23. Li, Jun (2006) *Learning Reactive Behaviors with Constructive Neural Networks in Mobile Robotics.* Doctoral Dissertation.

24. Otterskog, Magnus (2006) *Propagation Environment Modeling Using Scattered Field Chamber.* Doctoral Dissertation.

25. Lindquist, Malin (2007) *Electronic Tongue for Water Quality Assessment.* Doctoral Dissertation.

26. Cielniak, Grzegorz (2007) *People Tracking by Mobile Robots using Thermal and Colour Vision.* Doctoral Dissertation.

27. Boustedt, Katarina (2007) *Flip Chip for High Frequency Applications – Materials Aspects.* Doctoral Dissertation.

28. Soron, Mikael (2007) *Robot System for Flexible 3D Friction Stir Welding.* Doctoral Dissertation.

29. Larsson, Sören (2008) *An industrial robot as carrier of a laser profile scanner. – Motion control, data capturing and path planning.* Doctoral Dissertation.

30. Persson, Martin (2008) *Semantic Mapping Using Virtual Sensors and Fusion of Aerial Images with Sensor Data from a Ground Vehicle.* Doctoral Dissertation.

31. Andreasson, Henrik (2008) *Local Visual Feature based Localisation and Mapping by Mobile Robots.* Doctoral Dissertation.

32. Bouguerra, Abdelbaki (2008) *Robust Execution of Robot Task-Plans: A Knowledge-based Approach.* Doctoral Dissertation.

33. Lundh, Robert (2009) *Robots that Help Each Other: Self-Configuration of Distributed Robot Systems.* Doctoral Dissertation.

34. Skoglund, Alexander (2009) *Programming by Demonstration of Robot Manipulators*. Doctoral Dissertation.

35. Ranjbar, Parivash (2009) *Sensing the Environment: Development of Monitoring Aids for Persons with Profound Deafness or Deafblindness*. Doctoral Dissertation.

36. Magnusson, Martin (2009) *The Three-Dimensional Normal-Distributions Transform – an Efficient Representation for Registration, Surface Analysis, and Loop Detection.* Doctoral Dissertation.

37. Rahayem, Mohamed (2010) *Segmentation and fitting for Geometric Reverse Engineering. Processing data captured by a laser profile scanner mounted on an industrial robot.* Doctoral Dissertation.

38. Karlsson, Alexander (2010) *Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making*. Doctoral Dissertation.

39. LeBlanc, Kevin (2010) *Cooperative Anchoring – Sharing Information About Objects in Multi-Robot Systems*. Doctoral Dissertation.

40. Johansson, Fredrik (2010) *Evaluating the Performance of TEWA Systems*. Doctoral Dissertation.

41. Trincavelli, Marco (2010) *Gas Discrimination for Mobile Robots.* Doctoral Dissertation.

42. Cirillo, Marcello (2010) *Planning in Inhabited Environments: Human-Aware Task Planning and Activity Recognition.* Doctoral Dissertation.

43. Nilsson, Maria (2010) *Capturing Semi-Automated Decision Making: The Methodology of CASADEMA*. Doctoral Dissertation.

44. Dahlbom, Anders (2011) *Petri nets for Situation Recognition*. Doctoral Dissertation.

45. Ahmed, Muhammad Rehan (2011) *Compliance Control of Robot Manipulator for Safe Physical Human Robot Interaction.* Doctoral Dissertation.

46. Riveiro, Maria (2011) *Visual Analytics for Maritime Anomaly Detection*. Doctoral Dissertation.

47. Rashid, Md. Jayedur (2011) *Extending a Networked Robot System to Include Humans, Tiny Devices, and Everyday Objects*. Doctoral Dissertation.

48. Zain-ul-Abdin (2011) *Programming of Coarse-Grained Reconfigurable Architectures*. Doctoral Dissertation.

49. Wang, Yan (2011) *A Domain-Specific Language for Protocol Stack Implementation in Embedded Systems*. Doctoral Dissertation.

50. Brax, Christoffer (2011) *Anomaly Detection in the Surveillance Domain*. Doctoral Dissertation.

51. Larsson, Johan (2011) *Unmanned Operation of Load-Haul-Dump Vehicles in Mining Environments*. Doctoral Dissertation.

52. Lidström, Kristoffer (2012) *Situation-Aware Vehicles: Supporting the Next Generation of Cooperative Traffic Systems*. Doctoral Dissertation.

53. Johansson, Daniel (2012) *Convergence in Mixed Reality-Virtuality Environments. Facilitating Natural User Behavior*. Doctoral Dissertation.

54. Stoyanov, Todor Dimitrov (2012) *Reliable Autonomous Navigation in Semi-Structured Environments using the Three-Dimensional Normal Distributions Transform (3D-NDT)*. Doctoral Dissertation.

55. Daoutis, Marios (2013) *Knowledge Based Perceptual Anchoring: Grounding percepts to concepts in cognitive robots*. Doctoral Dissertation.

56. Kristoffersson, Annica (2013) *Measuring the Quality of Interaction in Mobile Robotic Telepresence Systems using Presence, Spatial Formations and Sociometry*. Doctoral Dissertation.

57. Memedi, Mevludin (2014) *Mobile systems for monitoring Parkinson's disease*. Doctoral Dissertation.

58. König, Rikard (2014) *Enhancing Genetic Programming for Predictive Modeling*. Doctoral Dissertation.

59. Erlandsson, Tina (2014) *A Combat Survivability Model for Evaluating Air Mission Routes in Future Decision Support Systems*. Doctoral Dissertation.

60. Helldin, Tove (2014) *Transparency for Future Semi-Automated Systems. Effects of transparency on operator performance, workload and trust*. Doctoral Dissertation.

61. Krug, Robert (2014) *Optimization-based Robot Grasp Synthesis and Motion Control*. Doctoral Dissertation.

62. Reggente, Matteo (2014) *Statistical Gas Distribution Modelling for Mobile Robot Applications*. Doctoral Dissertation.

63. Längkvist, Martin (2014) *Modeling Time-Series with Deep Networks*. Doctoral Dissertation.

64. Hernández Bennetts, Víctor Manuel (2015) *Mobile Robots with In-Situ and Remote Sensors for Real World Gas Distribution Modelling*. Doctoral Dissertation.

65. Alirezaie, Marjan (2015) *Bridging the Semantic Gap between Sensor Data and Ontological Knowledge*. Doctoral Dissertation.

66. Pashami, Sepideh (2015) *Change Detection in Metal Oxide Gas Sensor Signals for Open Sampling Systems*. Doctoral Dissertation.

67. Lagriffoul, Fabien (2016) *Combining Task and Motion Planning*. Doctoral Dissertation.

68. Mosberger, Rafael (2016) *Vision-based Human Detection from Mobile Machinery in Industrial Environments*.