UMEÅ UNIVERSITY

Object Detection and Recognition in Unstructured Outdoor Environments

Ahmad Ostovar

PhD Thesis, November 2019
Department of Computing Science
Umeå University
Sweden

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

ahmado@cs.umu.se

ISBN 978-91-7855-147-7
ISSN 0348-0542
UMINF 19.08

Printed by Cityprint i Norr AB, Umeå 2019

Abstract

Computer vision and machine learning based systems are often developed to replace humans in harsh, dangerous, or tedious situations, as well as to reduce the required time to accomplish a task. Another goal is to increase performance by introducing automation to tasks such as inspections in manufacturing applications, sorting timber during harvesting, surveillance, fruit grading, yield prediction, and harvesting operations. Depending on the task, a variety of object detection and recognition algorithms can be applied, including both conventional and deep learning based approaches. Moreover, within the process of developing image analysis algorithms, it is essential to consider environmental challenges, e.g. illumination changes, occlusion, shadows, and divergence in colour, shape, texture, and size of objects.

The goal of this thesis is to address these challenges to support development of autonomous agricultural and forestry systems with enhanced performance and reduced need for human involvement. This thesis provides algorithms and techniques based on adaptive image segmentation for tree detection in forest environment and also yellow pepper recognition in greenhouses. For segmentation, seed point generation and a region growing method was used to detect trees. An algorithm based on reinforcement learning was developed to detect yellow peppers. RGB and depth data was integrated and used in classifiers to detect trees, bushes, stones, and humans in forest environments. Another part of the thesis describe deep learning based approaches to detect stumps and classify the level of rot based on images.

Another major contribution of this thesis is a method using infrared images to detect humans in forest environments. To detect humans, one shape-dependent and one shape-independent method were proposed.

Algorithms to recognize the intention of humans based on hand gestures were also developed. 3D hand gestures were recognized by first detecting and tracking hands in a sequence of depth images, and then utilizing optical flow constraint equations.

The thesis also presents methods to answer human queries about objects and their spatial relation in images. The solution was developed by merging a deep learning based method for object detection and recognition with natural language processing techniques.

Sammanfattning

System för datorseende och maskininlärning utvecklas ofta för att ersätta människor i svåra, farliga eller tråkiga situationer, samt för att minska tidsåtgången att utföra en uppgift. Ett annat mål är att öka prestandan genom att automatisera arbetsuppgifter såsom inspektioner i tillverkningsprocesser, sortering av träd vid avverkning, övervakning, bedömning av frukt, prediktion av skörd, och avverkning. Beroende på uppgiften kan olika objektdetekterings- och igenkänningsalgoritmer tillämpas, både konventionella och metoder baserade på djupinlärning. När man utvecklar algoritmer för bildanalys är det dessutom viktigt att ta hänsyn till utmaningar i omgivningen, t.ex. förändringar i ljusförhållanden, delvis dolda objekt, skuggor och skillnader i färg, form, struktur och storleken på objekten.

Målet med denna avhandling är att angripa dessa utmaningar för att stödja utvecklingen av autonoma jordbruks- och skogsbrukssystem genom förbättrad prestanda och minskat behov av mänsklig inblandning. Denna avhandling beskriver algoritmer och tekniker baserade på adaptiv bildsegmentering för träddetektering i skogsmiljö, och även detektering av paprikor i växthus. För segmentering användes *seed point generation* och en metod för *region growing* för att upptäcka träd. En algoritm baserad på *reinforcement learning* utvecklades för att upptäcka paprikor i bilder. RGB- och djupdata kombinerades och användes i klassificerare för att detektera träd, buskar, stenar och människor i skogsmiljöer. En annan del av avhandlingen beskriver metoder baserade på djupinlärning för att upptäcka stubbar och klassificera rotningsnivån baserat på bilddata.

Ett annat bidrag i denna avhandling är en metod som använder infraröda bilder för att upptäcka människor i skogsmiljöer. För att detektera människor användes en formberoende och en formoberoende metod. Algoritmer för att detektera människors avsikt baserat på handgester presenteras även. 3D-handgester detekterades genom att först detektera händer i en sekvens av 3D-bilder, och sedan applicera ekvationer för optiska flöden. Avhandlingen presenterar också metoder för att besvara frågor från människor om objekt och deras spatiella relation i bilder. Lösningen utvecklades genom att kombinera en djupinlärningsmetod för objektdetektering och igenkänning med tekniker för analys av naturligt språk.

Preface

This thesis contains a brief description of methods for object detection and recognition in unstructured outdoor environments such as forests and greenhouses, a discussion on improving detection and recognition performance, and the following papers.

Paper I Thomas Hellström, Ahmad Ostovar. Detection of Trees Based on Quality-Guided Image Segmentation. *Proceedings of the Second International Conference on Robotics and associated Hightechnologies and Equipment for Agriculture and forestry (RHEA-2014), pp. 531-540.*

Paper II Ahmad Ostovar, Ola Ringdahl, Thomas Hellström. Adaptive Image Thresholding of Yellow Peppers Based on Reinforcement Learning. *Robotics, MDPI 2018, Vol. 7, (1).*

Paper III Mostafa Pordel, Thomas Hellström, Ahmad Ostovar. Integrating kinect depth data with a stochastic object classification framework for forestry robots. *Proceedings of the 9th International Conference on Informatics in Control, Automation and Robotics: Volume 2, SciTePress, 2012, pp. 314-320.*

Paper IV Ahmad Ostovar, Bruce Talbot, Stefno Puliti, Rasmus Astrup, Ola Ringdahl. Detection and Classification of Root and Butt-Rot (RBR) in Stumps of Norway Spruce Using RGB Images and Machine Learning. *Sensors, MDPI 2019, Vol. 19, (7).*

Paper V Farid Abedan Kondori, Shahrouz Yousefi, Ahmad Ostovar, Li Liu, Haibo Li. A Direct Method for 3D Hand Pose Recovery. *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR -2014), pp. 345-350.*

Paper VI Ahmad Ostovar, Thomas Hellström, Ola Ringdahl. Human Detection Based on Infrared Images in Forestry Environments. *Proceedings of the Image Analysis and Recognition: 13th International Conference, ICIAR 2016, pp. 175-182.*

Paper VII Ahmad Ostovar, Suna Bensch, Thomas Hellström. Natural Language Guided Object Retrieval in Images. *Submitted to Sensors, MDPI 2019.*

Financial support for this work is provided in part by the European Commission (CROPS GA no.246252 and SWEEPER GA no.66313).

Acknowledgements

It would not have been possible to accomplish this period of my academic life and write this thesis without support and help of many kind people. First of all, I would like to thank my supervisor, **Thomas Hellström** for unlimited support and help throughout my studies over the years. His knowledge and experience made our discussions and meetings effective and interesting. I very much appreciate his patient and the time he spends working with me. His support in my studies and life means a lot to me. Besides, I would also like to thank my co-supervisor **Ola Ringdahl** for his continued support, constructive discussions and comments on my works. I am grateful for all his help. Many thanks to **Peter Hohnloser** as a colleague who has always been there for me and making life enjoyable. I am really honored to be part of the Robotic group and I would like to show my greatest appreciations to other members of the group, **Suna**, **Maitreyee**, **Michele**, **Neha** and **Avinash**.

I would like to express my gratitude to **Carl Christian**, **Birgitte**, **Lili**, **Muyi**, **Juan Carlos**, **Angelika**, **Mikael**, **Mirko**, **Xuan-son**, **Chanh**, **Mahmoud** and **Esteban** for making life easier and more colorful. I would like to thank other colleagues at the department, especially **Yvonne**, **Anne-lie**, **Lennart**, **Pedher** and **Carina** for handling administrative aspects and accompany. I would also like to thank our system support administrators, **Tomas**, **Bertil** and **Mattias** for helping me with technical issues and also the beautiful Christmas decorations.

My deepest and sincere gratitude goes to my parents and brother, **Mojedeh**, **Reza** and **Nima** for the permanent and infinite love and support and also encouragement towards excellence. I am deeply in debt to them.

Lastly and most importantly, I am grateful to my lovely wife **Masoumeh**, for her precious understanding, endless encouragement and support with her love and patient. I really appreciate all her endeavors to provide me a peaceful environment, making me feel happy and also her heartful accompany.

Umeå, October 2019
Ahmad Ostovar

Contents

Chapter 1

Introduction

1.1 Research Motivation

Manual agricultural and forestry operations in outdoor environment is always challenging due to environmental conditions such as extreme temperatures and difficult working situations. Manual harvesting of crops are highly labor intensive and are becoming increasingly costly at the same time as the availability of skilled labor force is decreasing. Furthermore, working in such environments poses a high risk to the health of the humans working there. The labor issue is predicted to become even more critical with both cost and shortage of labour increasing in the future [77]. Therefore, utilizing robots for automatic fruit harvesting, farming, and forestry operations are essential as a response to this. Autonomous systems decrease the risk of human injuries, lower the harvesting cost by speeding up the operation, save money and energy, reduce the dependency on labor, and generally increase the performance of operations. This is beneficial for both producers and consumers and has resulted in a growing interest for developing robots for harvesting fruits, vegetables, and trees over the past three decades [10]. In the 1980s Sistler [239] and Pejsa et al. [195] discussed the advantages of using machine vision and robotics to improve sustainability of crop production and also indicated that citrus and apples have a high potential for robotic harvesting. The first computer vision based approaches were reported for peach and apple detection [240], shape and orientation detection of sweet peppers [278], and computer vision based self-driving tractors [214].

1.2 Aspects of Designing Computer Vision Based systems

The most crucial step in developing autonomous systems is perceiving the environment by using data received from different sensors such as RGB cameras, 3D, lasers, and infrared sensors. Localization/detection and recogni-

tion/classification of the target objects (e.g. fruits and trees) are also essential for such systems, which are performed by using machine vision techniques for analysing the data. Object detection provides a possible location of the target object and can also serve as a collision avoidance system in the navigation process. Moreover, object classification (recognition) brings the capability for the system of discriminating between detected objects (i.e. background vs. foreground, fruit ripeness, or tree species). Detected objects and the class that they belong to are further used for physical processes such as automatic gripping of objects (harvesting) by a robotic end-effector. As shown in Figure 1.1, designing a computer vision based system for object detection and recognition include several major steps: 1) image acquisition; 2) image pre-processing; 3) object detection/segmentation; 4) morphological operations; 5) feature extraction and 6) object recognition/classification. There are two bottlenecks for proper analysis of visual data:

1. **Environmental Conditions:** Outdoor environments such as orchards, greenhouses and forests are unstructured and dynamic. Shadows, light intensity, occlusions, and variable shape and size of target objects are some examples of challenges which results in degraded performance.

2. **Selection of Methods:** Since there are a large variety of sensors and methods to select from for data acquisition and developing each step of the algorithm, it is critical to select methods which work well together and result in a system with expected performance. The selected methods should be able to overcome environmental challenges. Moreover, selecting proper values for parameters have an important role in developing an autonomous system.

Despite improvements and extensive research on harvesting robots, the performance of these robots are not yet competitive in comparison to manual harvesting [91] which keeps agricultural operations as a important and challenging topic in computer vision and machine learning.

1.3 Evolution of Computer vision Based Systems

During the past decades an enormous number of algorithms have been developed for object detection and recognition in agricultural and forestry fields. Development approaches that have been used can be divided to three periods: using basic methods for segmentation, neural network based learning methods, and the deep learning era.

1. **Basic Methods:** up to a few years ago, the majority of the methods were threshold-based segmentation or color index-based approaches. These methods suffer from a significant problem, they cannot adopt to changes in the environment. The initial settings in these methods could work sufficiently for static images with high color and intensity differences

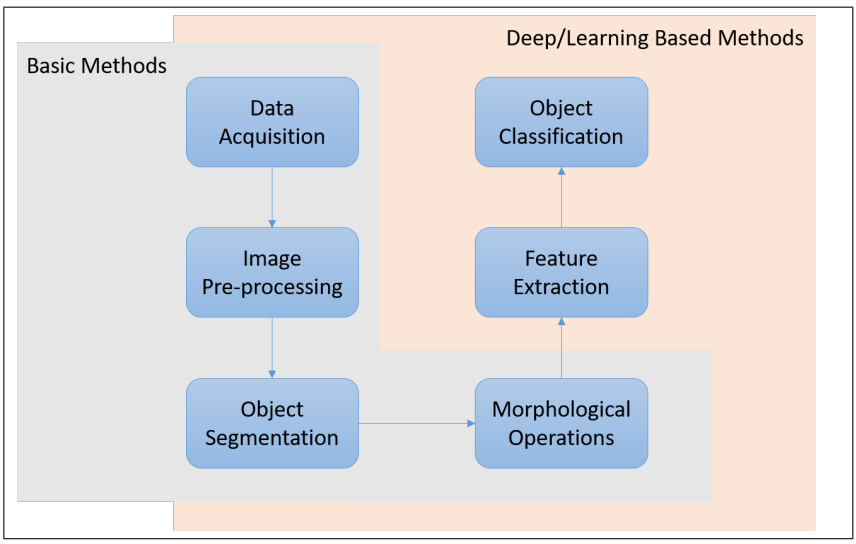

Figure 1.1: Block diagram of the steps that are generally involved in an object detection and recognition algorithm.

between objects in the scene, however when it is not static, for example if illumination conditions, fruit color (ripeness), or background changes, these methods are subject to failure.

2. **Learning Based Methods:** One approach to overcome problems with the basic methods is to use learning based methods, which can adopt to changing environmental conditions. They can be divided into unsupervised and supervised machine learning methods. In the beginning, unsupervised algorithms were popular, e.g. K-means clustering and fuzzy clustering. Later, learning methods based on artificial neural networks were applied, starting with random forest methods and later convolutional neural networks (CNN).

3. **Deep Learning Based Methods:** Development of more advanced machine learning methods for object detection and recognition in computer vision field resulted in generation of deep neural networks (DNNs). Deep learning methods have recently started to be adopted to the agricultural and forestry fields. These networks drop the need of hand-crafted feature extraction and instead automatically learn feature representations. Algorithms based on these networks proved to have higher performance and more capability in generalization. DNNs are generally divided into one-stage and two-stage methods for detection and recognition.

It worth mentioning that following the evolution trend within the field, this thesis presents methods for object detection and classification using approaches

from all three aforementioned periods of computer vision based methods.

1.4 Research Contribution

To provide the society with the benefits of autonomous harvesting operations, several projects have been financed by EU with focus on detection of objects in forestry and agricultural environments in which Umeå University had participated, such as CROPS[1] and SWEEPER[2]. The FP7 CROPS project focused on detection of trees, bushes, humans and rocks in forestry environments besides selective harvesting of fruits in orchards and greenhouses. The H2020 project SWEEPER, aimed at developing a sweet pepper harvesting robot for use in greenhouses. Moreover, the PRECISION[3] project, financed by Research Council of Norway, has the aim of improving resource utilization and reducing wood decay in Norwegian forests. Umeå University also participated in this project. During my PhD, mainly based on these three projects, I have evaluated a variation of imaging sensors, features of target objects, and image analysis algorithms. Using a combination of these, I developed algorithms for object detection and recognition. The main research problem investigated is object detection and recognition in unstructured environments, including forests and greenhouses.

The contribution of this thesis is as follows: in Paper I, an algorithm for detecting trees in a forest environment using colour images is presented. The algorithm couples a novel seed point generator with a segmentation method similar to region growing for tree detection. A novel method for adaptive image thresholding of yellow peppers in greenhouses based on reinforcement learning is demonstrated in Paper II. In Paper III a forestry robotic system is developed for detecting and classifying objects as bushes, trees, stones, and humans. In this research work, images from a depth sensor and an RGB camera were used and a set of stochastic classifiers were applied for object recognition.

Root and butt-rot (RBR) has a substantial impact on the quality of harvested timbers in forestry and wood processing industry. Detecting the presence of RBR and measuring its covering area during timber harvesting operation would benefit the industry in many ways, such as accurate sorting of harvested logs (affected trees cannot be used for saw timber for example) and recording the location of affected trees for later analysis of areas with high presence of RBR. Therefore, as presented in Paper IV, a system for automatic detection of tree stumps and classification of RBR using both conventional and deep learning based approaches was developed.

Paper V presents a 3D hand-gesture pose estimation, using 3D data from a Kinect camera, for gestural interaction systems. Paper VI proposes a human detection algorithm in forestry environments using a thermal camera. In this

[1] http://crops.sweeper-robot.eu/
[2] http://www.sweeper-robot.eu/
[3] https://www.researchgate.net/project/PRECISION-2

paper two human detection methods are introduced; one shape-dependent and one shape-independent approach. Paper VII presents an approach for generating responses to natural language based queries regarding objects and their spatial relations in images. In this approach, natural language processing techniques and image analysis algorithms are merged together. The main idea in this work is to develop an approach for human robot interaction.

1.5 Research Methodology

When developing algorithms we employed the Constructive Research (CR) method [50], a common research methodology in the field of computer science. Using the CR method results in a system that resolves a domain specific problem and also generates a theoretical contribution with academic value. Following the steps in the CR method, first we identified the research problem, which could be acquired from a project (See section 1.4). Then, we performed a systematic literature review to gain an overview of the specific research problem and understand the approaches that other researchers have proposed. In the next step, we designed and developed algorithms to detect and classify the target object. To demonstrate the feasibility of the designed algorithm, and also validating it, we tested the algorithm/framework with different set of images, containing different environmental conditions such as shadows and occlusions. Moreover, considering the important role of setting parameters for achieving a good performance, different sets of parameter values were tested. Furthermore, the proposed approach was evaluated by comparing it with similar approaches. However, the comparison is not always possible, due to lack of benchmarks within the field.

1.6 Thesis Outline

This thesis focuses on three main criteria: imaging sensors, visual characteristics of the target objects, and the image analysis algorithms, to provide an extensive overview of available sensors and methods within the domain of autonomous agricultural and forestry applications. The rest of this thesis is organized as follows. In Chapter 2 and 3, different imaging sensors and visual cues respectively are described. Image analysis algorithms are divided into conventional and deep learning based methods. Chapter 4 discusses conventional object detection and classification approaches and Chapter 5 provides a brief introduction of deep learning methods, detection and classification approaches, and some applications in agricultural and forestry environments. A short summary of the contributions within the included papers are discussed in Chapter 6.

Chapter 2

Sensor systems

In this section the focus is on data acquired by sensors/cameras. Since data acquisition is the first step in localizing and recognizing objects in the scene, it worth to start with analysing possible sensors and their configurations. The variability of types and sensors configurations for the purpose of object detection and recognition is large, ranging from a single monochrome camera to combinations of hyperspectral cameras without any visual sensors.

2.1 Single Camera

Using a single camera for object detection is the simplest approach. Within object detection and recognition applications in agricultural and forestry domain, the sensors are typically located on the robot (machinery) platform to present a single view of the scene [24, 181, 289]. In other systems the camera is mounted on the robot's arm (end-effector) [150, 217, 137, 185] which provides closer view of the object. It also can be used for path-planning of the robot movement toward the object of interest. Combining these two views in robotic systems [74, 284], would increase accuracy of the robot or its end-effector movements.

2.1.1 Black and White (B/W) Cameras

B/W cameras were used in some of the earliest researches in the 1980's for detecting objects such as fruits, based on its shape features [277]. D'Esnon et al. [58] developed a self-propelled robot for fruit harvesting in the MAGALI project, using a B/W camera. Object detection algorithms for detecting melons based on its shape features using B/W camera were developed by Cardenas-Weber et al. [34] and Dobrusin et al. [59] in the early 1990's. Edan et al. ([67]) detected melons based on reflectance, shape and texture features using B/W camera in the early 2000's. In this work the authors concluded that using several sensors and combination of their data could improve the object

detection performance. During this period other researchers used B/W camera with color filters to detect objects of interest within the agricultural field [240, 120]. The major disadvantage of B/W cameras is lack of color data, which is one of the most important features of fruits [89]. This limitation makes it difficult to achieve desired performances on fruit detection using only B/W cameras. Therefore, B/W cameras were later replaced with color cameras.

2.1.2 RGB Cameras

Color cameras with Charged Coupled Device (CCD) or Complementary Metal-Oxide-Semiconductor (CMOS) sensors are widely used within machine vision systems in robotics and autonomous agricultural and forestry operations [48, 151, 237] for localization, detection, and tracking objects of interest. Baeten et al. [11] used a color camera with a CMOS sensor to localize apples in trees, and also guiding the harvesting robot. They used the relationship between the intrinsic parameter of the camera (the focal length), the pixel size, and the center of the apples in images to compute the distance between the robot (camera) and the apples. At each step, several images were taken to calculate the remaining distance using triangulation. Kurtulmus et al. [130] developed an algorithm based on color cameras using a statistical classifier and a neural network to detect peaches.
Zhao et al. [6] used a CCD camera mounted on the end-effector of a harvesting robot to localize fruit in the scene. In the developed algorithm first the centroid of the fruit in the image was determined, then the robot arm was moved to the corresponding center point of the image and fruit together. Then the arm was extended until the fruit entered the gripper.
Mehta and Burks [162] used a single color camera, mounted on the robot base to estimate the location of citrus fruits based on depth information. They used pixel coordinates of the fruit, the fruit size, and the intrinsic parameters of the camera to obtain the euclidean position of fruit. Then by transforming the coordinates from camera to robot base reference frame, based on the extrinsic parameters of the camera, they calculated the position of the fruits with respect to the robot base. The authors of this work discussed that using estimated depth based on a single color camera could be a better approach than using computationally complex stereo vision or triangulation methods.//

A noticeable disadvantage of color cameras is that the acquired images are sensitive to varying light conditions. Objects in these images could suffer from over exposure or shadows, which directly affect the detection and classification accuracy. Additionally, detecting objects with less color distinction to the background is challenging. For example detecting green fruits such as cucumber or green bell peppers which have similar color as leaves and branches bring extra difficulties in comparison with detecting red apples.

As also noticed by Mehta and Burks [162] using a single camera for controlling movements of the robot or actuators is not efficient since it needs continu-

ous computation of distance and direction toward the object of interest, which slowing down the overall speed of the robot or its manipulator.

2.2 Stereo Vision

Stereo vision consist of two or more cameras which are located within a certain distance and are capable of providing depth information. Images captured simultaneously from all cameras are matched together to compute displacement or disparity of objects in the images. Using the cameras' relative location and orientation to each other and their focal lengths, the image disparity is converted to distances to the objects [235], thereby providing depth information. Researchers have widely used calibrated stereo cameras in agricultural harvesting robot applications to localize and identify fruits [79, 256, 126, 284]. Wang et al. [276] used a calibrated pair of cameras and a triangulation method to localize apples in global coordinates. The system could also identify repeated counting of apples, caused by having multiple images. Fruits that were located at a distance less than twice the diameter of the fruit were regarded as repeated occurrences of the same apple within multiple images.
Plebe and Grasso [198] placed stereo cameras in two arms of an orange harvesting robot. They performed stereo matching of orange centres by using an ANN to locate oranges in a 3D coordinate system. Multiple pairs of calibrated cameras can be used to improve visibility of the object, specifically occluded ones, and also to handle illumination variations.

Stereo vision based systems suffer from their complexity, time consuming computations [97] and low accuracy. Furthermore, stereo matching is particularly problematic in outdoor environment where objects might be relocated by wind and also changes in illumination conditions (sunlight, shadows) can affect the matching processes [198].

2.3 Time of Flight (TOF) 3D cameras

TOF 3D cameras operate based on the principle of Time of Flight of light, i.e the time it takes for light to travel from the camera to the object and back, which is used to compute the distance.They provide intensity and 3D coordinates of objects in the scene. 3D images are widely used for reconstructing models of buildings, navigation of self driving cars, and detecting objects. The Photonic Mixing Device (PMD) CamCube 3.0[1] and Microsoft Kinect[2] are examples of such cameras. Kinect is usually used as a low cost sensor for acquisition of short-range 3D data with high capability of fusion with other sensors such as RGB cameras [89].

[1]PMD Technologies GmbH, Siegen, Germany, https://www.pmdtec.com
[2]Microsoft Corporation, Redmond, WA

TOF cameras in agricultural automation systems have been used for autonomous 3D reconstruction of apple trees [123], localization of fruit in trees [88], determination of inter-plant spacing [173] and detection of objects in forestry environments [201]. Gongal et al. [88] used a TOF camera to detect 3D coordinates of apples in trees. Coordinate information was further used to identify the repeated apples in images acquired from two opposite sides of the tree canopy. Authors determined that some of the system errors were caused by error in registering 3D location data with detected apples in the color camera images, which resulted in lower accuracy of the system.

Using 3D cameras reduce the data acquisition and processing time in comparison to stereo vision cameras. TOF cameras also provide higher accuracy in 3D reconstruction than stereo vision systems [18]. These characteristics make these cameras a good option for fusion with other sensors such as color cameras. However, TOF cameras are expensive and provide low resolutions.

2.4 Thermal Cameras

Thermal cameras capture thermal response of objects, which is beneficial in discrimination of objects of interest and background. In thermal cameras the emitted radiation of heat is captured within the infrared range (9-14μm) [19]. Researchers have used thermal imaging for pedestrian detection [118, 110], in forestry for automatic detection of human in close range of harvesting machines for safety [184], and in agricultural environment to detect fruits [26, 25, 244]. In all developed systems using thermal imaging, detection is based on the temperature difference between the object of interest and the background.
As detection of green fruits is a challenging task using only color information, researchers have used thermal cameras for detecting these types of fruits. Stajnko et al. [244] used a thermal camera to detect apples in an orchard. In this work, pixel values of thermal image was first transformed to RGB color space and then to chromaticity coordinates. Afterwards, a global threshold based on normalized difference index (NDI) was used to perform segmentation. Authors discussed that the accuracy was limited by exposure to direct sunlight as well as the fruit size and recommended adding shape features in the detection process to improve the accuracy of the system.

Since the thermal response of objects are sensitive to the exposure to sunlight and heat accumulation, the detection can be less accurate if the object is shadowed or located deep in the background (tree canopy), as there would not be a significant temperature difference between the object of interest and the background in those sections.

2.5 Laser Range Finder

Laser range finders work either based on the principle of Time of Flight (TOF) of light or phase shift method and they can provide 2D or 3D information of the scene. A laser sensor consists of two units, a source that emits pulse laser beams and a sensor that receives the reflected beam from nearby objects. LiDAR (Light Detection and Ranging) can achieve 2D coverage of hundreds of meters by scanning the laser beam using a fast rotating mirror. By moving the sensor horizontally or vertically a 3D map of the environment can be created. Researchers widely used laser sensors in robotics, agricultural, and forestry autonomous systems to localize fruits in trees [76, 262, 273], estimate tree diameter [218], and estimating the position of the harvester head of forest machines [149]. Bulanon et al. [28] measured the distance to fruits in apple orchards using a laser sensor. For this purpose a laser sensor and a RGB camera were mounted on a manipulator. The target fruit center, which was detected using the color camera, was aligned in the center of the image using visual servoing and the distance to the fruit was measured using the laser sensor.

Laser sensors achieved better location accuracy compared to other sensors [24]. On the other hand, systems using 3D laser sensors are slow, costly and bulky in comparison to stereo vision and TOF 3D sensors.

2.6 Spectral Cameras

The development of sensors and spectroscopy technology increased the popularity of spectral imaging as a method for detecting objects based on their reflectance at different wavelengths. Spectral images provide both spectral and spatial information of target objects. Spectral imaging provide advantages to detect fruits even if they have similar color as the background [284, 27, 276, 121, 122]. Kane and Lee [122] utilized multispectral imaging by capturing images at three optical bands using near infrared (NIR) wavelengths (1064, 1150 and 1572 nm) and performed index computation[3] on the images. Afterwards, fruits were segmented using Otsu's threshold [49] followed by morphological processes. In this research work, the main challenges were fruit movements between image acquisition with different filters, due to wind and changes in lighting conditions. The problem of target movements could be solved by capturing multiple waveband images at the same time.
Hyperspectral imaging provides a complete spectral signature for each pixel in the scene, which can result in higher accuracy of object detection and classification [180, 227, 2]. Safren et al. [227] used hyperspectral data, a combination of visible and near infrared (NIR) wavelengths, for apple segmentation. Authors used a multistage algorithm to analyze the data. First the dimensionality of

[3]Image comparison between different wavelength images are commonly referred to as indices.

the data was reduced using principle component analysis (PCA). Then analogous objects were extracted and classified following by morphological processes, watershed, and blob analysis.

The accuracy of hyperspectral imaging is higher than multispectral imaging [124], as it provides more information about reflectance characteristics of the objects. However, it suffers from long data acquisition (in the order of minutes per image) and image processing time. Therefore, hyperspectral imaging are usually used for offline processing and also for preprocessing of data to recognize useful spectral channels which would provide beneficial information for real time processing.

2.7 Fusion of Imaging Sensors

Another approach for object localization and classification is fusion of multiple sensors, as one type could compensate the limitations of the other type. Object localization and classification in agricultural and forestry fields using fusion of an RGB camera and a 3D sensor (Microsoft Kinect) have been investigated with high accuracy and minimal computation demands [201, 82, 148]. Despite the lower resolution of the 3D camera, it provides accurate distance measurements to the objects of interest. Therefore, integrating low resolution images and high location accuracy from 3D images with high resolution color images can improve resolution of 3D images and result in high accuracy object detection and classification.

Fusion of a laser range finder and stereo camera was also studied for the purpose of object tracking and obstacle detection with promising results [125, 13, 133]. Since 3D laser sensors are expensive, it is possible to lower the cost in laser-stereo fusion using a grid of a small number of visible lasers [89]. The grid of laser beams provide both various features for stereo matching and accurate references for calibration of stereo based 3D data.

Jnaneshwar et al. [55] developed a system for automated monitoring enabling precision agriculture by fusing a LiDAR scanner, a thermal imaging camera, multispectral cameras, and GPS sensors. Using these sensors they extracted plant morphology, canopy volume, leaf area index, and fruit counts. In this work the fusion of thermal images with color images results in detection of green oranges, which were not visible enough in color images. Authors discussed the need of lowering the cost of the designed sensor fusion.

Sensor fusion increase the number of sensors in the system, which results in additional system complexity and cost. Therefore in designing such systems it is essential to investigate the variation of sensors to optimize the computation time, accuracy, and cost. Furthermore, it is necessary to identify ways to avoid or minimize the sensors used in a robotic system. For example, instead of mounting several sensors on an end-effector of a robotic system for accurate localization and classification, they can be replaced with one sensor with higher

resolution and accuracy and also improving the image processing capabilities of the system to control a large number of robotic arms in parallel. Scarfe et al. [229] developed an approach to provide location information of fruits to four robotic arms using a pair of sensors capturing a wider area of canopy.

The major challenges of the sensing systems are limited robustness, high cost, excessive complexity, insignificant computation speed, and sensitivity to environmental conditions (See Table 2.1). Therefore, to improve accuracy of autonomous object detection and recognition in agricultural, horticultural and forestry fields it is essential to consider these limitations and design approaches that minimize the effect of these complications.

Table 2.1: Summary of sensors used for autonomous object detection in agricultural, forestry and horticultural environments.

Sensor	Advantage	Limitation
B/W camera	Less sensitivity to lighting conditions	Lack of color information
Color camera	Provides color, shape and texture features	Sensitivity to lighting conditions
Stereo Vision cameras	Providing 3D information	Time consuming computation
TOF 3D camera (Kinect)	High accuracy in distance measurement	Low resolution, Sensitive to lighting conditions
Thermal camera	Independent of object color	Sensitive to the temperature, low resolution
Laser camera	High accuracy in distance measurement, less sensitive to lighting conditions	Low resolution, expensive sensors
Spectral camera	Providing spectral and color data	High computation cost and acquisition time

Chapter 3

Visual Cues for Object Detection

Visual features (cues) of objects refer to their visual characteristics which provide information for determining them from the background in images. Feature extraction (detection) is a low-level processing step in which the input is pixel intensities and the output is visual features. A wide variety of visual features are studied and applied to computer vision based applications such as object localization, object recognition, object retrieval and visual tracking. The fundamental goal is to extract highly stable features to guarantee robustness of the system.
The main challenge in computer vision tasks is the gap between high-level concepts (computer vision tasks) and image pixels. Visual cues are utilized to bridge the gap. Since visual features directly impact the system performance, it is crucial that extracted features being discriminant and descriptive. Selecting a proper set of features is a challenging task due to variety of imaging conditions, changes in scale, illumination conditions, viewpoints and image quality [143].

3.1 Color Features

Color is an important feature which is widely applied in object detection systems. It has been used in computer vision based systems, for example to distinguish fruits from background areas such as leaves, branches and stem. Researchers segmented fruits for detection based on their distinct colours, including citrus, oranges, red apples, tomatoes, pineapples and peaches [204, 99, 141, 181]. The accuracy of object detection in these works are limited mainly due to variable illumination conditions. As an example, apples in a tree could have different colours depending on its maturity level, exposure to sun light and location (shadowed), which makes detecting them challenging.

Color features can be extracted from different color spaces such as RGB, HSV, HSI, etc, depending on the task and illuumination conditions of the environment. Zhou et al. [295] used color features in both RGB and HSI spaces for apple detection. Rather than the differences between the RGB color channels, which was used to differentiate between red and green apples, red apples were further segmented by applying a threshold on the saturation channel in the HSI space.

Ostovar et al. [185] converted images from RGB color space to HSV and extracted features from hue and saturation channels followed by an adaptive thresholding method to segment yellow peppers in a greenhouse environment. Hellström et al. [107], converted RGB images to HSV color space to detect trees in a forest environment. The authors discussed that conversion to other color spaces were also evaluated, but trees were more distinguishable in the HSV space.

In all the proposed approaches for object detection based on color features, described above, the accuracy of detection was affected by variation of object color, object variety, dynamic and varying background features, similarity of color between object and its background, variable illumination conditions, and occlusion. Therefore, to achieve high precision in object detection it is necessary to use other features such as shape, texture, and reflectance in addition to color features [89].

3.2 Texture Features

Texture is a repeated visual pattern (either stochastically or regularly) which covers regions and surfaces of objects. It goes beyond purely local features (e.g. color and spectral reflectance) to illustrate characteristics of image patches [174, 80]. Texture can be used as an effective feature for object detection in unstructured and dynamic environments, when color and reflectance cues are not discriminative and stable enough, because of the its stability under varying illumination conditions.

Texture analysis has has been used extensively for fruit detection [37, 129, 208, 181]. In all these works they used the property of smooth surface of fruits, such that fruits surface generate lower edge density than the background. This make them distinguishable from the background, e.g. leaves and stems. Additionally, as the surface color does not affect the texture analysis, it can be used to detect fruits with similar color as other parts of the plants. An example of this is a work by Kurtulmus et al. [129] who developed a system to detect and count green citrus using circular Gabor texture analysis, proposed by Zhang et al. [287]. The results of texture analysis was then integrated with blob analysis and a 'Eigenfruit' approach used to identify fruits.

It is recognized that accuracy of texture-based segmentation in uncontrolled

outdoor environments is mainly limited by extreme variation of lighting conditions, varying object size and occlusions [289]. Overexposure and lack of light are cases of extreme lighting conditions where the texture details of the objects are not recognizable. In case of varying sizes and occlusions combination of textural features with other visual cues could be a solution [208].

3.3 Geometric Features

Geometric features, e.g. shape and size, provide a set of distinguishable features for object detection. They are less susceptible to variable illumination conditions, which generally cause variation in appearance of objects. This characteristic make them a good candidate for object detection in unstructured environments. Geometric features implies an exceptional spatial relation between points, contours and surfaces of an object which specify the physical properties of the object.
In forestry and agricultural fields, objects such as tree trunks, canopy, and fruits can be detected by their shape properties. For example, a trunk is a long vertical object, canopies are wider and have larger areas while fruits are smaller, round objects with generic size constraints. Additionally, detection of fruits which have similar color signature as the background (e.g. leaves and branches) is easier using geometric features.
Geometric feature are popular in harvesting robots [126, 154, 208, 184, 107, 284, 201]. Ostovar et al. [184], used the ratio of width to height in infrared images as a shape-dependent feature to detect humans in forestry. Also, Hellström et al. [107], considered the predominantly vertical orientation of trees as a feature to refine generated segments in images for tree detection.

Despite the strengths of geometric features such as persisting to extreme variation of illumination conditions, they are extremely sensitive to occlusions, as it would change the projection of objects shape. Also, large variability of objects shape within a class of target object affects the accuracy of detection using only geometric features. Therefore, in most of works where shape features are used explicitly, only simple shape models such as spheres are employed [126, 154, 284]. Furthermore, extracting and analysing geometric features are computationally intensive, which may limit their application in real-time environment.

3.4 Thermal Responses

Thermal responses (features) of objects are highly related to spectral reflectance, based on temperature difference of objects and their surroundings within infrared range. They are used in agricultural and forestry environments to detect fruits and humans [25, 81, 244, 26, 276, 184]. Since fruits and other parts of a plant have different characteristics in absorbing and radiating heat (fruit absorb

and radiate more heat than the rest of the plant), thermal features provides an option for automatic object detection in harvesting machines [25, 272], specifically fruits with the same color as the plants (e.g. cucumber, green apples and citrus) [244]. Ostovar et al. [184] used thermal features to automatically detect the presence of humans in close range to forestry machines.

As discussed in Section 2.4, due to sensitivity of thermal features to temperature differences and emissivity characteristics of the object, location of fruits (e.g. if located in shadowed areas) might result to lowering the system performance based on this cue [25, 244]. It was also discussed in [184], that different clothing changes the heat radiation map of the human body, which bring difficulty in human detection. Similar to other features, it is not possible to overcome all challenges within the agricultural and forestry fields such as occlusion by using only thermal features.

3.5 Spectral Reflectance

Spectral reflectance can be used as an effective discriminatory feature to detect objects in the outdoor environment. They have been used widely in fruit harvesting robots, specifically when the chromatic differences between the fruit and the plants foliage is insignificant (e.g. green peppers, cucumbers, green citrus). Object with similar color might not have the same spectral signature and can have different reflectance properties in selected channels (either inside or outside of visible light range), called metamerism [279]. This makes spectral reflectance based features a good candidate for object detection. For this purpose either the entire spectral signature [180, 227], a set of selected spectral channels [284, 265], or specific wavebands [256] are used.

Regardless of number of channels used to capture spectral reflectance, object detection only based on this visual cue is limited. Due to its sensitivity to varying illumination conditions [284, 265] and also occlusions. For example, it is possible that fruit and plants parts have the same spectral signature (e.g. young leaves and green fruit). Therefore, it would be beneficial to fuse this cue with other features to overcome challenges in outdoor environments.

3.6 Integration of Visual Cues

As discussed above, a single visual feature is often not capable of representing the object of the interest in unstructured, dynamic environments (see Table 3.1 for more details). In autonomous object detection systems, efforts to detect target objects using one visual cue usually failed due to variable illumination conditions, occlusions, variability in shapes and size [33, 194, 89]. For example, lighting conditions in orchard affects the color of the object and occlusion by other parts of the plant results in changes of the geometric characteristics of

the object in images. In such common scenarios, detection based on only one feature would generate poor accuracy.
Since each feature illustrate different characteristic of the object, it is reasonable to compensate the limitation of one visual cue by including other features into the system. As a result, fusion of features might improve performance of the object detection. This approach is generally noted in computer vision [291] and particularly in the agricultural field [192]. It also worth mentioning that integration of visual cues can be done either at the image level [25] or at the algorithm level [271].
Researchers [151, 245, 193, 194, 33] commonly use integrated feature from different groups to enhance the performance and robustness of object detection algorithms in outdoor environment. Pordel et al. [201] combined features from RGB and HSV color spaces with depth based features to enhance performance of object classification in forestry environment. Hannan et al. [98] integrated color and geometric features to detect oranges in tree images. Chromaticity in red channel was used for segmentation and then a threshold based on the perimeter feature was applied to detect oranges. Most of the studies concluded that the main causes of limited accuracy of object detection in unstructured outdoor environment are occlusions and variable lighting conditions. These issues could partially be resolved on two levels. On the field level, agricultural and horticultural operations such as tree training, pollination, pruning, and thinning can potentially increase the visibility of fruits and reduce occlusion and clustering to improve performance of object detection. On the sensor level, it is common to address varying illumination conditions by having a controlled lighting environment. Wang et al. [276] and Payne et al. [194] simplified the problem by suggesting nighttime imaging under artificial lighting. However, this approach limits the operational time of harvesting machines. Furthermore, Arad et al. [7] suggested using a flash-no-flash method [197] to stabilize the impact of the environmental illumination conditions in images. Utilizing these techniques can further settle using machine vision approaches.

Table 3.1: Summary of features used in object detection algorithms and their limitations.

Feature	Variability	Limitation
Color	Features from different color spaces	Varying lighting conditions
Geometric	Shape and size related features	Occlusions, Varying lighting conditions
Texture	Spatial variation in pixel intensity (edges)	Occlusions, Varying lighting conditions
Thermal	Temperature variation in pixels	Sensitive to temperature, Occlusion
Spectral	Pixel intensity in different wavebands	Sensitive to spectral signature, Occlusion
Integration of features	Combination of features from any group	Occlusion, Varying lighting conditions, Clustering

Chapter 4

Conventional Object Localization and Classification Methods

Object localization[1] is one of the fundamental and challenging areas in computer vision, in which the aim is to determine the existence of a predefined target object in the image. If the object is present, its spatial location and extent should be determined. Object detection (localization) is the basis for resolving high level vision tasks. Generally a greater attention is placed on localization of structured objects (e.g. cars, bicycles, ships) and articulated objects (e.g. cats, dogs, humans) than unstructured ones (e.g. grass, sky, sea) [152]. The spatial location and extent of an object can be determined in different ways: bounding box [70, 223], pixel based segmentation mask [288], or closed boundary [147, 224]. Representation of objects using bounding boxes are the most common method for the evaluation of the object localization approach [70, 224]. The location of the object in the image determines the region of interest (ROI), which means segmenting the object of interest (foreground) from the surroundings (background). Segmentation methods can be divided into traditional approaches, which are described in this Chapter, and deep learning based methods illustrated in Chapter 5, Section 5.2.

Object recognition[2] is another essential component in computer vision task with the goal of determining the presence of target objects from a set of given object classes in an image or assigning object class labels to a given (detected) object. An ideal object recognition approach should be able to distinguish different instances of the same object class, subject to intra-class appearance variations [152]. Intra-class variations can be divided into two categories: intrinsic factors and imaging conditions. The former term include different instances

[1] We use the terminology object *detection* and *localization* interchangeably.

[2] We use the terminology *recognition* and *classification* interchangeably.

of an object category, varying in visual characteristics such as shape, size, and color. Imaging condition variations are changes in object appearances caused by environments such as illuminations, occlusions, viewing distance, backgrounds, and cameras. For example, apples in a tree could have different colors and shapes depending on their location (e.g. shadowed), maturity, ripeness, and level of occlusion. For other types of objects, such as humans, the object instances could be different poses, clothing, and non grid deformations. Object classification approaches are also divided into conventional approaches, presented in this chapter and deep learning based, described in Chapter 5, Section 5.2.

4.1 Localization Methods

Objects of interest can be located at any position with variable scales and aspect ratios in an image. Localizing an object of interest (object detection) can be modelled as a segmentation of the image into foreground and background. Foreground serves as regions of interest (ROI) which are areas that are likely to contain the object of interest. Segmented areas are determined using bounding boxes, pixel based masks, or boundaries and represent distinctive characteristics, i.e. color, intensity, texture, or edges. These regions are later used to extract features and recognizing (classifying) objects in the image. Under the assumption that all objects of interest in the image have common visual features making them distinguishable from the background, it is possible to design or train an algorithm that outputs a set of proposed regions which contain objects. Different image segmentation methods have been used by researchers, but since the method to use directly depends on the task domain, the selection of the segmentation technique needs to be based on knowledge about the specific problem to be solved. Furthermore, as in each task objects of interest (e.g. shape, color, size) and environmental conditions (e.g. lighting and occlusion) are variable, a successful approach in one domain is not necessarily applicable in another domain. Methods for extracting ROIs can be divided into two categories, sliding window techniques and object proposal approaches.

4.1.1 Sliding Windows

In the era of handcrafted feature descriptors such as SIFT [156], HOG [53] and LBP [179], the most successful object detectors (e.g. DPM [73]) used sliding window techniques [53, 73, 101, 266, 268]. However, approaches based on sliding windows techniques suffer from an extensive number of windows, which increase with the number of pixels in the images. Furthermore, its demand to search at various scales and aspect ratios additionally expands the search space. Detection based on sliding windows generates about 10^4-10^5 windows per image, and it increase to 10^6-10^7 by considering various scales and aspect ratios [152]. As a result, it is computationally expensive method which makes

it a challenging approach for real time detection and classification. In the following some of the segmentation methods within this category are briefly described.

Template Matching

Template Matching (pattern matching) is a technique for matching parts of the image with a specific template pattern. These algorithms use the pixel intensity information from the template image as the source of features for matching. The process usually works by moving the template over the image, and in each position it computes how well the template matches with that portion of the image. The computation is based on measuring the similarity using approaches such as sum of squared differences (SSD) or cross-correlation. The similarity measure can be also based on visual cues such as edges and corners.

Template matching is useful when diversity of the objects is small, to increase the chance of the matching process. Therefore, when the environment is filled with different types of objects, which is the case in most of the forestry and agricultural environments (fruits, leaves, stems etc.), it is less beneficial. To overcome this problem, it is possible to apply template matching on binary images [286] when target objects are represented with some labels such as blobs to reduce the variability of objects for comparing with a shape like circle as a template.

Template matching accuracy is limited due to highly varying shapes of objects in typical agro-forestry environments. In addition, illumination conditions may change the shape of the target object (due to shadows and overexposure). Occlusions add further complications. Furthermore, computing the similarity in each step of the process is time consuming, which makes this approach inappropriate for real-time systems. Moreover, applying any pre-processing methods, such as to generate binary images with labels, adds to computation time. Generally, template matching is inappropriate to combine with sliding window techniques.

Shape Inference

Shape is a major visual cue that autonomous harvesting robots should be able to determine. It is beneficial specifically when harvesting operation must be selective and individuals. Shape inference (geometric matching) involve in finding a shape that best matches with the geometric features from portions of the image [80]. In these algorithms, geometric information from the template image is used as the primary features for matching. These features can range from low level features i.e. edge, line or curves, to high level feature such as geometric shapes made by low level features. Therefore, shapes can be divided to local and global inferences, obtained from low level and high level features. Determining global shape inferences requires predefined shape model of the target object, and also a fitting technique to compare the model to the image

regions. Local shape inference uses local shape descriptors to indicate presence of a particular shape instances. Methods such as voting, deterministic or stochastic optimisation and statistical inferences can be used to compute the inference.

The main challenge in both cases is construction of reliable models for the target object. In addition, global shape inference is computationally expensive due to matching process and detecting instances of the model in the image. Therefore, most of researchers limited the domain of its application to spherical fruits and also using local shape inferences to simplify the model and its detection in images and also lowering the computation process.

Applying local shape inferences, researches [198, 116] used edge fitting process to detect spherical shape characteristics by estimating the center and radius of edges in each part of the image. Spherical shape approach (global shape inferences) also used to segment target objects [181] by first generating the edge map of the image and then analysing it through a geometric template consisting a circular region and an outer ring with predefined dimensions. Fruits were located and separated from the background using the ratio between the number of edge points included in the circular region and the outer ring.

Voting

Voting is a technique in which local visual evidence within the image votes for all possible global interpretations that could be derived. In computer vision this method is used to detect shapes and patterns. A well-knows voting technique is the Hough transform [113], which is designed for line detection. A variation of this technique is the circular Hough transform (CHT), developed to detect circular patterns. Researchers have used it to detect spherical fruits such as oranges [116], coconuts [219] and apples [271]. Since the Hough transform and CHT are computationally expensive, researchers have proposed several techniques to reduce the complexity. One approach is to reduce the parameter space by setting the expected radius of fruits either globally [116] or dynamically (for each object) [271]. Another approach is to reduce the number of votes per image point using information of edge directions [116, 219]. Due to the expensive computations of CHT, faster approaches have been proposed and applied for fruit detection [150, 42].

Generally, approaches based on voting method are computationally expensive, due to the need of visiting all parts of the image, and also they usually require a pre-processing step. Furthermore, usage is limited to detection of spherical shaped objects.

4.1.2 Object Proposals[3]

Object proposal (or detection proposals) based approaches are developed to balance the strain between expensive computations and high detection accu-

[3]We use the terminology *object* proposals and *detection* proposals interchangeably.

racy. This technique is based on the idea of abjectness proposed by Alexe et al. [4], which are regions of an image that are highly probable to contain an object. Object proposals limits the number of regions that need to be evaluated by the detector, therefore they can speed up the process compared to the sliding windows approach. An object proposal based detection method should meet three requirements: 1) high localization accuracy such that they match to the object bounding boxes accurately, 2) high recall rate by using a small number of proposals, and 3) reduced computation cost to be appropriate for real-time detection.

Object proposal methods are generally similar to the interest point detection approaches [260, 163], in which feature descriptors are computed around interest points and then used for object detection, classification, and retrieval. The difference is that detection proposals are based on low level visual cues to generate candidate boxes. Two notable segmentation approaches within this category are defined in the following.

Thresholding

Numerous studies [141, 25, 154, 181, 208, 201] approach segmentation with thresholding methods based on low level visual cues, specifically color. Predefined thresholds often fails in unstructured, dynamic outdoor environments, mainly due to illumination changes. Therefore, many researchers have proposed adaptive thresholding methods [269, 274, 23, 200, 98, 185], where the threshold automatically adapts to the environmental illumination conditions. Adaptive thresholding provides higher accuracy in object detection, however its performance is limited due to high variability of typical unstructured agricultural environments. This lead researchers to develop other segmentation methods such seed point generation and region growing [126, 54], edge detection [54, 205, 289], and region merging [227].

In most of these approaches, only one visual cue was used, and other characteristics of the target object such as shape information are ignored. Therefore, when the target object is located within the cluster of objects (such as a cluster of peppers), the segmented area is the cluster of objects rather than individual fruits. This may not be a problem in application such as automated spraying of weeds, but is a definite challenge in harvesting applications, when individual fruits and their stem should be detected accurately. This problem has been approached by several researchers, without shape information [213, 227] using watershed method. In this method gray scale image is seen as topographical surfaces, and boundaries are considered as the curves separating basins of flooding.

Clustering

Clustering is an unsupervised learning approach that has been used in machine vision algorithms for autonomous harvesting robots to segment an image

into object and background. It is beneficial when fusion of several visual cues are employed. Despite development of advanced clustering methods, the classical K-means algorithm applied on various color spaces is the most popular approach for fruit detection [271, 24, 139]. Selecting a measure of distance between data points in feature space is a major aspect of clustering. While Euclidean distance is sensitive to scaling of variables, and cannot handle correlation of variables, Mahalanobis distance shows better performance, and have, for example, been used in harvesting robots [290]. Since the Mahalanobis distance takes into account the covariance among variables to calculate distance, the problems of scale and correlation are resolved.

Similar to the thresholding approach, clustering is also sensitive to lighting conditions. Furthermore, one has to deal with data points that do not cluster, or get assigned to incorrect clusters. Moreover, since shape characteristics of objects are not considered, clustering based approaches often have poor accuracy in outdoor environments, unless combined with other methods in which shape is considered.

Segmented regions could further processed to generate more accurate object proposals using the following approaches.

1. **Segmentation Grouping Methods**. Attempts to generate multiple segments with high probability of containing an object. The simplest approach in this method is to directly use image segmentation algorithms such as thresholding. To increase number of candidate regions, it is also possible to either perform multiple low level segmentation [36, 261] or start with an over segmentation and then randomly merging them [160]. The merging decision is mostly based on cues of segmented areas such as superpixel shape, object visual cues, and boundary estimation [8, 60]. Grouping methods are categorized into three types based on how they generate object detection proposals [108]: grouping superpixels, using multiple graph cut with diverse seeds, or based on edge contours.

   Methods in this category include approaches such as selective search [264, 261], constrained parametric min-cuts (CPMC) [35, 36], and multiscale combinatorial grouping (MCG) [9]. The output of these methods is a segmentation mask of the objects.

2. **Window Scoring Methods**. A substitute method for generating object proposals is to score each candidate region based on the probability of containing an object. The window scoring approach include methods based on objectness [4, 3], edgeboxes [297], binarized normed gradients (BING) [41], and also methods proposed by Rahtu et al. [207] and Feng et al. [75]. In comparison to grouping methods, window scoring usually return bounding boxes and they are also faster. On the other hand, object proposals generated by these methods have a lower localization accuracy [108]. The accuracy can further be improved by refining the location of detection proposals.

3. **Neural Network Based Methods**. In this method, a neural network is used to score object proposals, which are produced by a segmentation algorithm, for determining existence of an object (or a portion of it). The trained network is basically used to regress the number of proposals in the image. Neural networks in this approach are designed based on multi layered models, segmentation as the first layer and then a segment classifier as the next layer. Most of the work based on this approach [4, 93, 264, 35, 68] first trained a classifier to distinguish candidate boxes containing instances of an object. The classifier is then used to score bounding boxes in the test images. Visual cues from candidate boxes are used to train the classifier. Instances of the object are localized using local maxima of the score. A distinguished method that fits in this category is the multibox approach [253, 69] which is further improved by using deep neural networks.

Within the object proposal approaches based on low level visual cues (color, edge, texture and gradients), Selective Search [264, 261], MCG [9] and Edge-Boxes [297] are methods which received the most attention within the field. Object proposals based detection results in avoiding applying sliding window search on the image. It benefits the detection by reducing the search area and consecutively decreasing the computation time.

4.2 Reducing Localization Error

To improve the efficiency of the segmentation process, such that the accurately cover the surface of the target object, further developments are required. These processes are described in followings.

4.2.1 Segmentation Modification

After segmentation, the detected regions typically need further refinement to make sure that they cover the target object completely. A common approach is to apply morphological operations such as dilation, erosion, closing and opening to the detected regions. In this process, assumed geometric characteristics of the target object are taken into account to achieve higher accuracy. Also, methods such as region growing based on visual cues of the object (such as color) could be beneficial. Moreover, merging of segmented regions based on similarity of features (analogous to grouping method in the object proposal approach, see Section 1) can improve the accuracy.

4.2.2 Segmentation Refinement

As the number of segmented regions could be very large, and not all segments contain the target object, it is beneficial to reduce the number of segments for further processes. Discarding segments is mostly performed based on some

predefined measurements based on visual cues of the target object. For example Hellström et al. [107] developed a quality assessment function based on vertical properties of trees to refine segmented areas with the aim of improving tree detection performance. We can interpret this process as a variant of the window scoring method (see Section 2).

4.3 Segmentation Evaluation

Evaluating accuracy of segmentation is typically done by comparing the predicted (detected) regions with manually labeled areas (ground-truth) in the images. Generally the output of the segmentation process is an irregular shape that covers the target object surface. However, it is common practise to define a minimal rectangular bounding box that covers the detected shape.

There are several methods to evaluate segmentation accuracy, either pixel-wise, based on the segmented mask, or using bounding boxes. For comparison, both methods should use ground-truth of its own kind. The most common approach to evaluate accuracy of detected areas is the Intersection Over Union (IoU) method, as shown in Equation 4.1.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{4.1}$$

The numerator is the area of overlap between the predicted region and the ground-truth. The denominator is the area encompassed by both the predicted region and the ground-truth. The IoU score is between 0 to 1, where 1 means complete match between the segmented region and the ground-truth. IoU score computation is often combined with a threshold value, such that regions with an IoU greater than the threshold are considered as containing target objects.

Based on IoU and the performance index in [71], Ostovar et al. proposed an evaluation method for segmentation quality (sq) based on segmentation-overlap and segmentation-efficiency measures [185, 186] as shown in Figure 4.1 such that $sq = \frac{\frac{O}{L}+\frac{O}{S}}{2}$.

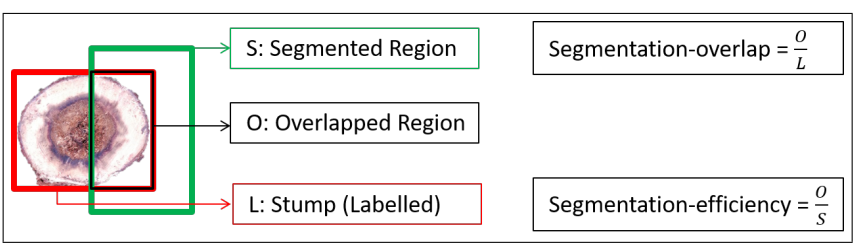

Figure 4.1: Segmentation quality using Segmentation-overlap and segmentation-efficiency. Figure from [186].

4.3.1 Bounding Box Refinement

The aim of the object detection as part of a detection and classification algorithm is to accurately localize objects by maximizing the IOU or some other metric. Since a bounding box (in the following also denoted BB) is a coarse estimate of the location of a target objects, a number of improvements are possible. It is crucial to refine the BBs, discard BBs which represent the same object, and also take into account the fact that background pixels are included in BBs. Methods to resolve these issues are discussed in the followings.

1. **Bounding Box Regression** This is an important technique with the aim of refining the location of a BB based on an initial proposal region. An efficient method is to consider the target object's characteristics, such that if the BB contains object features. Then further processes based on these features can be selected and performed. In the history of computer vision, the usage of BB regression can be divided into three periods. In early stages of detection, VJ and HOG detectors did not use BB regression, instead they directly utilized the sliding window as the result of the detection process. To attain accurate location of objects, researchers generated feature pyramids and then densely slide the detector on each location.

   BB box regression was first used by Felzenszwalb et al. [72] when developing a deformable part-based model (DPM) for detection. The method, uses "root" filter score to generate the one final detection window. The root filter location, based on Dalal and Triggs model [53], defines detection windows by capturing coarse resolution edges of objects. Later, more advanced methods were introduced to predict bounding boxes based on the complete configuration of objects and formulated the the process as a linear least-square regression problem.

   After introduction of convolutional neural networks (CNN), the third generation of BB regression were developed, as described in Chapter 5, Section 5.4.

2. **Non-Maximum Suppression (NMS)** Duplicate detection, i.e. multiple overlapping BBs containing the same object instance, is a common problem. As neighbouring windows usually have the same detection score, NMS methods, as a post-processing step, is used to eliminate the duplication and obtain the final BB. NMS methods are divided into three categories, greedy selection, BB aggregation and learning to NMS.

   - Greedy selection is the simplest method with the idea that from a set of overlapped BBs, the one with maximum detection score (could be based on some predefined measurements based on visual cues) is selected, and the neighbouring boxes are eliminated according to a predefined overlap threshold. This process is applied iteratively to the BBs in a greedy manner.

- Bonding box aggregation is based on the idea of combining several overlapped BBs to generate a final BB. The advantage of this method is that it considers object relationships and also their spatial layout. The Viola Jones detector [268] and Overfeat [233] use this approach.
- The main idea of learning NMS is to use NMS as a filter to re-score all generated BBs, and to train the NMS as a network in an end-to-end fashion. Learning NMS have shown improvements in object detection accuracy over traditional NMS methods, when objects are occluded or located densely.

3. **Hard Negative Mining (HNM)** Training an object detector or a classifier is often an imbalanced learning problem. For successful training, both positive (foreground) and negative (background) examples are essential. However, sliding window methods often generate a very large number of background windows. Using all BBs that contain background as negative examples for training will disturb the learning process. The HNM method is developed to deal with this problem. Bootstrap and HNM in deep learning are two main categories of developed HNM methods. In bootstrap, training starts with a small number of background samples and new samples of miss-classified background are then added. It was initially introduced to reduce the extreme number of background samples in [268, 191], and was later used as a training technique in DPM and HOG detectors to solve imbalance of data. HNM in deep learning based detectors is discussed in Chapter 5, Section 5.4.

4.4 Recognition Methods

In an autonomous object detection system, after localizing objects of interest, the system should be able to resolve two sub-tasks. 1) Determine if a segmented ROI contains the object of interest or part of the background, 2) Determine the type of the object. Determining if an extracted ROIs contains a fruit or background, and recognizing the type of fruit (orange, apple, peach) are examples of these two sub-tasks, respectively.

A recognition task consist of two steps, feature extraction and classification. These steps are described in the followings.

4.4.1 Feature Extraction

Extracting features from regions of interest and training classifiers is a general approach for object recognition in computer vision. Selecting a proper set of features is essential for high accuracy classification.

In early stages of object recognition, researchers used template matching techniques and simple part-based models [78]. In these methods the focus was

on recognizing objects with rigid spatial layout. Later on, more attention was put into object recognition using geometric representation of objects [170, 199]. Later the focus moved to utilizing statistical classifiers such as neural networks [222], Support Vector Machines (SVM) [187] and Adaboost [280] based on object features [171, 230] instead of geometry and prior models.

The general direction of developing object features moved from global to local representation in which features are invariant to changes in translation, rotation, and scale. Local invariant features started from Scale Invariant Feature Transform (SIFT) [156], HOG [53] and Local Binary Pattern (LBP) [179]. These local features are then usually combined using simple concatenation or feature pooling encoders such as Bag of Visual Words [51], BoW models [135] and fisher vector [196].

The common approach in the agricultural and forestry domains is extracting features (see Chapter 3), that describe characteristics of objects, from regions of interest, and then feed them to a set of classifiers for the recognition task. Expertise in the field, and prior knowledge about the target object, is essential. For example an apple can be recognized using its red color feature, but since illumination conditions are dynamic in outdoor environments, it is not possible to rely only on the color feature. On the other hand, smoothness of its surface (geometric feature) can be seen as a distinguishing feature of apples, while it may happen that fresh leaves also show the same properties. Moreover, changes in illumination conditions may result in representing other parts of the tree with the same smoothness, usually due to overexposure of the area.

Since a single feature cannot appropriately represent the target objects, features are usually aggregated to increase recognition performance. Pordel et al. [201] extracted 12 color based features including mean and variance of RGB and HSV values of the object in addition to six depth based features for recognizing human, tree, bush and stone. Ostovar et al. extracted two types of features from ROIs geometric based features and pure pixel values from infrared images to detect human in forestry environments [184]. In another project, the author used a grouping approach for feature extraction by utilizing BoF, based on SURF features, to classify rot on stumps of trees [186]. In all these researches, recognizing the best set of features for recognizing objects of interest were the most complicated and critical part of the project. It is complicated as features should be able to characterise objects in a way to maximize its persistence to environmental conditions such as lighting and occlusion. Moreover, they should be discriminative enough to represent objects within the same category with different sizes, shapes and colors. Therefore, extracting informative features needs expertise and additional considerations.

It also worth to note that increasing number of features and their diversity make the model more complex and affect recognition performance, thus it is beneficial to carefully design the feature extraction processes and in case of combining features performing dimensionality reduction. A machine learning model with a large number of features is highly dependent on the data it is trained on. The may result in an overfitted model with poor performance on

real (test) data. Dimensionality reduction can be made in two ways, feature selection (keeping the most relevant features from the original set of features) and dimensionality reduction (finding a smaller set of new features, each being combination of the original features, containing the same information as the original set). Principle component analysis (PCA) and related methods [119], independent component analysis (ICA) [111] and linear discriminant analysis (LDA) [21] are some common approaches for dimensionality reduction.

In the next step, a set of extracted features are fed into classifiers for the recognition task.

4.4.2 Object Classification

In machine learning, classification is the task of, given a set of features of an object, decide which object category it belongs to. Object classification methods can be divided to two groups: unsupervised and supervised. These two groups and some of their underlying methods within agricultural and forestry machine vision tasks are described in the following.

Unsupervised Classification

Unsupervised learning is the training of the system using data which is neither classified nor labeled. Here the task of the system is to group unsorted data according to similarities, patterns and differences without any prior training. One of the widely used unsupervised classification method for developing autonomous systems in agricultural environments is k-means clustering [234]. Also other unsupervised classifiers such as fuzzy clustering and Gaussian mixture models are used. These methods and their applications are briefly defined in the following.

1. **K-means Clustering** with the K-means algorithm divides input data points into a number of clusters. It aims at minimizing the distance between each point and its associated cluster center. It iteratively moves objects between clusters until the sum of distances is minimized. Most developed algorithms for fruit detection employes the K-means clustering algorithm [271, 29, 24].

   Wachs et al. [271] used thermal and color images to detect green apples. Bulanon et al. [29] used K-means algorithm to detect red apples in chromacity 'rgb' color space, which was achieved by transforming the RGB color space.

2. **Fuzzy Clustering** or soft clustering assigns each point in the dataset the probability of belonging to each cluster, while in the k-means algorithm each point just belongs to one cluster. The fuzzy approach is used when data points in some respect may belong to several clusters. The

distance function in fuzzy clustering is the measure of having the probability relative to the inverse of distance [84]. Instead of fixed assignments of data points to clusters they are assigned to a cluster that has maximum posterior probability. Fuzzy clustering is commonly used in agricultural environment for different tasks such as soil management based on soil or vegetation characteristics [259, 270], crop segmentation [95, 221] and crop disease detection [159, 167].

Guijarro et al. [95] classified crops from the soil and sky as background, using a threshold over different color indices. The fuzzy clustering was used to cluster textures within the same class to differentiate background from crops pixels. Tellaeche et al. [258] and Majumdar et al. [159] developed a classification system using fuzzy clustering to detect weed in barely fields, and to detect four different disease of wheat plants based on their leaves, respectively.

3. **Gaussian Mixture Models (GMMs)** GMMs are similar to fuzzy clustering, suitable for cases where the data points overlap between different classes. GMMs can be used specifically when the prior information about data distribution is known and also classes are normally distributed.

   GMMs applications in agricultural field includes yield estimation [57], vegetation segmentation and mapping [12, 56] and fruits and crop classification [56, 254, 16]. Tabb et al. [254] developed an autonomous harvesting system for apple detection from video in which used Global Mixture of Gaussians to model the background. De Rainville et al. [56] performed gaussian mixture clustering to classify weed and crop using their leaves features. Bauer et al. [16] classified two types of disease on leaves of sugar beet from multispectral stereo images using an automatic detection methods developed based on GMMs.

Supervised Classification

Supervised classification is a learning process that maps labeled input data into classes, with the goal of using the learned model to predict classes of new data (test data). Several approaches of supervised classification have been used to develop computer vision based autonomous system in forestry and agricultural environment. In the following, the most common methods are briefly described.

1. **Bayesian Classifier** is a probabilistic classifier based on Bayes' theorem. It makes decision based on prior knowledge and probability distribution with the assumption of conditional independence among predictor features such that the presence of a specific feature in a class is not related to any other feature. Bayesian classifier works by maximizing the posterior probability depending on the priori probability [64].

   Kurtulmus et al. [130] extracted color and textural features from hue and saturation channels, from HSV color space, and Gabor texture and used

them to train a Bayesian classifier to detect immature peaches. Bandi et al [14] used textural features which were extracted from HSI color co-occurance matrices (CCMs) with Naive Bayesian classifier to detect diseases on citrus leaves. Stegmayer et al. [247] fused color, shape and texture features to train a Bayesian classifier to automatically classify infected citrus fruits. Mursalin and Mesbah-Ul-Awal [172] used nine shape features with Bayesian classifier to classify four types of weeds. Caglayan et al. [31] developed a method based on Bayesian classifier using fusion of shape and color features from leaf images for automatic plant recognition.

The main drawback of Bayesian classifier is that in the process it cannot learn the relation between predictor features due to its conditional independence assumption. As a result, adding more features might not increase the classification performance but decreasing the accuracy because of their correlation [216].

2. **K-Nearest Neighbour (KNN)** is widely used for classification and regression tasks. KNN is instance based and does not learn a model during a training process. Instead it classifies unknown feature vectors to the class of its K nearest neighbour in training data [234].

   KNN has been used by many researchers to classify fruits [151, 231, 130]. Linker et al. [151] developed a methods based on KNN to classify apples using color and textural features from ROIs. Seng et al. [231] used three types of features including color-based, shape-based and size-based features, and KNN to classify apples, bananas, lemons and strawberries. Ahmad et al. [1] used Haar wavelet transform to extract features with a KNN classifier to classify weed. Li et al. [142] extracted three features based on the R and B channels of the RGB color space and Hue values with a KNN classifier to classify blueberry fruit into different growth stages.

   One major drawback with the KNN algorithm is the time consuming process of computing the distance to other observations. Another drawback is that classification accuracy typically decreases with increased dimensionality of the data [236].

3. **Artificial Neural Networks (ANN)** learn from the environment by an iterative training process, and improves its performance after each iteration. It consist of multiple layers of neurons in input, hidden and output layers. Neurons in layers are connected where each connection has an associated weight. Number of layers varies from task to task, the more complex a task, the more layers of neurons (as hidden layers) are used.

   ANN based classifiers received enormous attention in agricultural fields for developing computer vision based systems such as fruit detection [198, 271, 213, 20], crop recognition [190] and disease detection [168, 169].

Plebe and Grasso [198] used color features and a neural network to identify oranges for a robotic harvester. Wachs et al. [271] developed a system by fusing three ANN classifiers trained with back propagation to detect apples. Each ANN was trained and tested for the color spaces L*a*a, HSV and RGB. Regunathan and Lee [213] used a multi-layer ANN with back propagation to detect citrus based on hue, luminance and saturation values in HLS color space. Hue and saturation values of each pixel were used to classify the pixel in the fruit. These pixels were then segmented and morphological processes and watershed transformation were applied to these segments. Bhatt et al. [20], developed a system based on ANN to classify apples based on physical characteristic of apples such as color, size and also their external defects.

Although ANNs has demonstrated high performance for classification tasks, there are several consideration to design a successful network. Issues such as network size, learning rate, number of training cycles and thresholds for acceptable errors can affect the design and performance of the network. Therefore, it is essential to consider criteria such as determining input and output variables, selecting proper number of training sets, initializing network weights, choosing training parameters such as learning rate and also selecting the training stop criteria, in the designing process of an ANN network.

4. **Support Vector Machine (SVM)** is a binary classifier used for linear and non-linear regression and pattern classification. It is generally used to classify data into two disassociated classes. For linear classification, SVM separates two classes by maximizing the margin between them using a linear hyper plane. For non-linear separable classification, the feature vector is transformed into a higher-dimensional space that is linearly separable [30], and the hyperplane that separates data with maximum margin is then computed.

   The SVM classifier has successfully been applied in different areas of forestry and agricultural machine vision systems such as fruit detection [115, 275, 204, 232], tree detection and rot classification [201, 186], vegetation classification [44] and plant disease detection [46, 182]. Qiang et al. [204] developed a system for automatic detection of citrus fruit using multi-class SVM classifier. In this work Radial basis function (RBF) was used to classify citrus from background (leaves and branches) using features extracted from RGB color space. Wang et al. [275] and Ji et al. [115] used SVM classifier to identify apples. For classification three different SVM kernels functions including Poly, RBF and Sigmoid was used based on color, shape and merging of color and shape features. Sengupta and Lee [232] used CHT approach to localize spherical objects as ROIs, then extracted local texture and Tamura features [255] to identify green citrus in images.

SVM algorithms are not appropriate for large datasets as it needs long training time. It is sensitive to high noise levels in data i.e. when target classes overlap. Moreover, it is difficult to understand the structure of the final model [267]. Additionally, it is not easy to fine-tune the hyper parameters (C and γ) of SVM.

5. **Fusion of classifiers** as it is usually difficult to predict which classifier will perform best for a given classification task, a common approach is to utilize several classifiers for the same task [128, 225]. It provides various benefits, the researcher can determine which one fits the best to the problem and focus on that specific one to improve performance by changing its hyper-parameters. Moreover, it gives the possibility of combining predictions by several classifiers to further improve performance.

   Fusion of classifiers has been broadly used by researchers within the filed of agricultural and forest domain [258, 257, 220, 176, 184, 201]. Pordel et al. [201] fused results from five classifiers to make the final decision as the class of detected objects in forest environments. Ostovar et al. [184] improved classification performance of a human detection system in forest environments by fusing outcomes of three classifiers.

   There are generally two methods to aggregate predictions of classifiers [296], 1) Hard voting, the class which gets the most votes is selected (majority-vote) 2) Soft voting, if classifiers can provide a probability value of an object belonging to a class, then soft voting predict the class with the highest class probability, averaged over all the individual classifiers. In both methods, classifiers can either contribute all equally in the ensemble prediction, or the contribution of each classifier to prediction is weighted proportionally based on its performance.

Other Classification Methods

In addition to previously introduced supervised and unsupervised classification algorithms, there are other classifiers in agricultural and forestry machine vision systems that can be regarded as both supervised and unsupervised. It includes approaches based on Hidden Markov Model (HMM) and Reinforcement Learning (RL). HMM is an extension of Markov models, with added hidden conditions and observable observations [22]. Leite et al. [140] developed a system based on HMM to classify different agricultural crops using features from satellite spectral images.

RL focuses on interaction between the agent and the environment to learn the best action [248]. It can be used in autonomous agricultural and forestry machines to teach them to improve their movements or selection according to their relation based on the changes of the surrounding environment. The RL have been used to develop path planning, navigation and object detection approaches for autonomous systems within the filed [17, 117, 185]. Ostovar et

al. used reinforcement learning to detect yellow peppers for a harvesting robot in greenhouse environment [185].

4.5 Classification Evaluation

Developing a classifier consist of two steps, training and testing. Therefore it is needed to split the dataset into two sets, one for each step. The hold-out method is the simplest approach, in which the original dataset is partitioned into two sets, randomly selecting instances as training and test sets. In this method typically 2/3 of the dataset is selected as the training set and the rest as the test set. The classifier is trained using the training set and then evaluated on the test set. In this method, the classifier is trained and tested using only a portion of all data, which means that samples are presented in either the training or test data. it might result to bias in the classification process. Therefore, it would be beneficial to use repeated holdout method which gives more reliable estimation of the classifier by repeating the process with different subsamples. But since different test sets may overlap or some data points may never appear in the training sets, thus this approach is also not optimum. Using these approaches classifier might suffer from either underfitting or overfitting.

A common approach to overcome these problems is using K-fold cross-validation. It splits the data into K-folds, then trains the data on K-1 folds and tests on the remaining fold. This process repeated for all combinations and averages the result. The advantage with this method is that all data points are used in training and test sets which means that each is used once in the test test. K value is usually set to 5 or 10 as they result in an acceptable balance between computational complexity and validation accuracy.

The performance of the classifier is usually evaluated by computing the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These four numbers constitute a confusion matrix. Using the confusion matrix, classification performance can be presented by: Accuracy: overall effectiveness of a classifier, Precision: proportion of correctly positive identification, Recall: effectiveness of a classifier to identify positive labels, Fscore: measure of test accuracy (considering both precision and recall) and Receiver Operating Characteristics (ROC curve): the trade-off between the true positive and false positive rates.

Chapter 5

Deep Learning Based Object Localization and Classification Methods

With the limited performance of methods based on multistage hand tuned pipelines of hand engineered features and discriminative classifiers, deep learning methods based on convolutional neural networks (CNNs) emerged as an influential techniques for learning feature representations automatically from data. These methods are able to learn robust and high level feature representation of data which make them a powerful approach to perform complex tasks in broad range of problems such as object detection, object recognition, natural language processing, speech recognition and genomics.

The very first usage of CNNs for object detection and recognition can be traced back to the 1990s [263, 222], within limited domains such as face detection. The rebirth of CNNs happened when the successful application of deeper CNNs (DCNNs) in object classification [127] was transferred to object detection resulted to development of Region-based CNN (RCNN) by Girshick et al. [87]. Since then, object detection methods evolved remarkably with unprecedented speed.

5.1 A Brief Introduction to Deep Learning

Deep learning revolutionized machine learning in a wide variety of applications, from image classification to natural language understanding. Convolutional Neural Networks (CNNs) are the most representative model of deep learning and has a hierarchical structure with a number of layers that learn representation of features (data) with several levels of abstraction [136]. CNNs usually consist of three main operations, convolution, nonlinearity and pooling, which

creates layers in the network. In the convolution, nonlinear and pooling layers, features are convolved with a 2D convolutional kernel (or filter or weights), a nonlinear function (typically a rectified linear unit - ReLU) is applied to the features, and feature maps are either downsampled or upsampled. CNNs having a large number of these layers create a deep network which is referred to as DCNNs. Most layers of a CNN include feature maps, in which each pixel is considered as a neuron. In each convolutional layer, each neuron is connected to feature maps of the previous layer using 2D filters or set of weights. Early layers of a CNN consist of convolutional, ReLU and pooling layer, the later layers are mostly fully connected layers. Input image is repeatedly convolded in layers of a CNN and with each layer the receptive field increases. Earlyer layers of a CNN are responsible for extracting low level features (e.g. color, edge) and later layers extract more complex features (e.g. shape) [285, 183].

DCNNs provide several advantages including: 1) learning features automatically and directly from input data with minimum domain knowledge, 2) ability to learn very complex functions and 3) learning representation of data within several levels of abstraction using hierarchical structure. It means that for the detection and classification tasks, it is not anymore needed to search for the best set of features which perfectly characterizes the objects of interest considering different environmental conditions and possible object variants.

The success of DCNNs mainly depends on the existence of large training sets. Availability of large scale labeled datasets such as PASCAL VOC [70], ImageNet [223], and MS COCO [147] play a key role in this success. Using these datasets, researchers could aim for more complex tasks with large inter-class similarity and intra-class variations [70, 223]. Moreover, access high-performance hardware such as GPUs has provided the required computational power for handling huge networks.

5.2 Localization and Classification Methods

There has been a clear change in object feature representations, localization and classification, from handcrafted features [268, 101, 53] to learned DCNN based features [87, 52, 215]. Approaches proposed since deep learning entered the field can be grouped into two categories:

- Two-stage detection frameworks, which form a coarse-to-fine process by including a preprocessing step for generating object proposals.
- One-stage detection frameworks, approaches that are completed in one step which do not separate the process of region proposal detection.

5.2.1 Two-Stage Frameworks

In these frameworks (region-based), category independent regions (ROIs), based on objectness measure (see Chapter 4, Section 2), are first generated from an

image. Then CNN [127] based features are extracted from these regions, and as last step extracted features are fed into classifiers to determine the category label of the ROIs. Methods such as RCNN [87], OverFeat [233], DetectorNet [249] and MultiBox [233] were developed almost simultaneously,and use CNN architectures for object detection. In the followings some of methods within this category are briefly described.

1. **RCNN** [87]: Inspired by high performance image classification results achieved by CNNs and also success of the selective search as a region proposal method [261], Girshick et al. proposed RCNN. It starts with extracting region proposals (ROIs) using selective search. ROIs with IOU ≥ 0.5 overlap with a ground truth are set as positive examples and the rest as negative ones. Then ROIs are rescaled to a fixed size and fed for fine-tuning into a CNN model, AlexNet [127], which is pre-trained on a large dataset (ImageNet). Features extracted from the CNN model are used to train a set of linear SVM classifiers to predict the existence of an object within each ROI, and also to recognize the category of the object. RCNN also uses a BB regression, learned for each class of object using CNN features.

   Despite high performance in object detection, RCNNs suffer from some constraints. Training is accomplished in multiple stages, and since each stage must be trained separately, it is slow and hard to optimize. Training SVM and BB regressors are computationally expensive and time consuming because CNN features have to be extracted from a large number of object proposals. All these issues lead to a slow detection speed.

   These constraints motivated proposing of other detection frameworks such as SPPNet, Fast RCNN, Faster RCNN and Mask RCNN to overcome these problems.

2. **SPPNet** [106]: In the testing process of RCNN, extraction of CNN features from a large number of region proposals in each image was the main bottle neck. To resolve this problem, He et al. [106] proposed spatial pyramid pooling (SPP) [135] in the CNN architecture. They added the SPP layer on top of the fully connected (FC) layer to enable a CNN to generate fixed length representation of features. It discarded the need of rescaling region proposals. RCNN with SPPNet needs to run the convolution layer only once over the entire image to generate fixed length representation of region proposals with different sizes, and avoid repeatedly computing the convolutional features. SPPNet speed up RCNN significantly without sacrificing detection accuracy.

   Although SPPNet has accelerated detection speed, training is still multi-stage, which result in a slow training process. Moreover, SPPNet does not fine-tune convolutional layers before the SPP layer. This limits its accuracy when deeper networks are used. Later Fast RCNN [85] was proposed to overcome these problems.

3. **Fast RCNN:** Were proposed by Girshick et al. [85] to address some limitations of RCNN and SPPNet. It improves their detection accuracy and speed. In Fast RCNN, the detector training is end-to-end and enables the model to train simultaneously the softmax classifier and class specific bounding box regressor under the same network configurations. Fast RCNN adds a region of interest (ROI) pooling layer between the last convolutional layer and the first fully connected layer, it provides the FC layers a fixed length feature vector for each region proposal. FC layers are divided into two output layers, a multiclass classifier (softmax) which computes the probabilities of object categories and class specific bounding box regressor to refine proposals. Fast RCNN improves the speed in both the training and testing processes. Furthermore, it provides higher detection accuracy, uses a single training process and is able to fine-tune all network layers.

   Although, compared to RCNN and SPPNet, Fast RCNN provides higher detection speed and accuracy, it is still dependent on the selective search for region proposal generation, which limits its detection speed. Therefore, the possible next level of progress was to generate region proposals using a CNN model, which resulted in development of Faster RCNN [215].

4. **Faster RCNN:** as it was shown in some research works such as [293, 292, 47], it is possible to localize objects in convolutional layers of CNNs, while fully connected layers have less ability for this purpose. Therefore, a CNN can take the place of selective search for generation of region proposals. Faster RCNN proposed region proposal network (RPN) for generating region proposals which is accomplished using features from the last convolutional layer. RPN firstly initializes k reference boxes (anchors) with different sizes and aspect ratios from each location in the convolutional feature map. Then reference boxes are mapped to a lower dimensional vector and simultaneously fed into two fully connected layers, an object category classifier and a bounding box regressor. Therefore, RPN can be considered as a fully convolutional network (FCN) [155]. Faster RCNN utilizes the same network as Fast RCNN, but replaced selective search with RPN which is applied to the last convolutional layer to generate region proposals faster and more accurately.

   From RCNN to Faster RCNN all individual blocks, proposal detection, feature extraction, bounding box regression and category classification are integrated into a unified process, making learning an end-to-end approach.

   Although Faster RCNN speed up the region proposal detection, resulting in near realtime detection, it is still computationally expensive.

5. **Mask RCNN:** was proposed by He et al. [105] to achieve pixel-wise object segmentation by extending Faster RCNN. Mask RCNN uses the same

two stage architecture in Faster RCNN, with the same first stage, utilizing RPN for generation of region proposals, however the second stage is divided into two branches. One stage predicts object category and BB regression (similar to the final stage of Faster RCNN), and one stage outputs a binary mask for each ROI. These two parts are run in parallel. The second branch consists of a CNN feature map and a fully convolutional network (FCN). Furthermore, since ROI pooling layer causes misalignment, a ROI alignment layer was added to maintain the spatial location of pixels. Mask RCNN is simple to train, adding a small overhead to the Faster RCN. With the backbone network of ResNeXt101-FPN [281] it achieved highest accuracy in COCO object instance segmentation and BB object detection. However, its speed is limited to 5 FPS [105], which indicates the need of further improvements.

It worth mentioning that other methods also have been developed based on two-stage frameworks to further improve the accuracy of detection and also speed up the detection process. Chained Cascade Network [188], is an end-to-end learning approach consisting of more than two cascade classifiers, and DCNNs for object detection. This method is further extended in Cascade RCNN [32], and applied for simultaneously detecting objects and instance segmentation. It won the detection challenge of COCO in 2018. Furthermore, the Light Head RCNN [145] method was developed to accelerate the detection speed of RFCN [52] method by reducing the ROI computation.

5.2.2 One-Stage Frameworks

The region based strategies, based on region proposal detection, are computationally expensive. Therefore, instead of proposing approaches for optimizing an individual components of two-stage frameworks, researchers developed one-stage (unified) strategies.

In one-stage framework architectures, prediction of class category and regression of bounding boxes are directly applied to the full image, with a single feedforward CNN. These frameworks does not include region proposal generation or post classification sections. Since the whole process uses a single network, it can be optimized end-to-end based on detection performance.

1. **DetectorNet:** proposed by Szegedy et al. [249], was one of the first approaches to use a CNN for object detection. It formulated the detection as a regression problem to BB masks of objects. DetectorNet uses Alexnet [127] as the backbone network, however the softmax classifier layer was replaced with a regression layer. To increase robustness of object mask localization, five networks are used, one to predict the object box mask, and four to predict four halves of the box, bottom, top, left and right halves. It specifically helps to separate objects which are located close to each other. These predicted masks are then converted to a bounding box using a grouping process.

The designed network need to be trained per object type and mask type. Also it does not scale to multiple classes. Since DetectorNet needs to take many crops of the image as input and run multiple networks for each, thus it a slow network.

2. **OverFeat:** proposed by Sermanet et al. [233], integrated object localization and classification using one CNN. It performs object classification at different locations of the image, using a sliding window fashion method on multiscales of the input image to generate object candidates. OverFeat uses a CNN such as AlexNet [127], which takes input images with fixed size, and models the network into a fully convolutional network by using fully connected layers as convolutions with kernels of size 1-by-1, to enable the model to take inputs of any size. Using multiscales, it improves performance by passing six enlarged scales of the input image trough the network. A classifier outputs a grid of predictions, including the class and confidence value, for each multiscale input.

   Once the object is identified, a localization regressor is applied to predict the location of the bounding box. The regressor and the classifier use the same convolutional layers (feature extractor). However, fully connected (FC) layers need to be trained to predict object bounding boxes at each spatial location and scale. The regressor is class specific.

   In OverFeat predicted bounding boxes are combined using a greedy merging strategy, in which individual BBs with sufficient overlap from localization and also confidence value of being the same object, coming from classifier, are merged.

   Since computation of convolution between overlapped regions are shared, which means that features are not needed to be computed repeatedly for those areas in the network, OverFeat has a speed advantage but it is not as accurate as the RCNN [87] method. OverFeat is similar to methods which were proposed afterwards, such as YOLO [212] and SSD [153], but different in training the classifier and the regressor, which in OverFeat is done in a sequential manner.

3. **YOLO (You Only Look Once):** was the first real one-stage detector, proposed by Redmon et al [212]. It cast the object detection task as a regression problem from image pixels to bounding boxes and class probabilities. As indicated by its name, YOLO drops the region proposal generation and verification steps, and directly predicts detection using features from an entire image based on a small set of candidate regions. YOLO divides an image into S x S grid, each predicting class probabilities, bounding box locations and a confidence score (objectness score), simultaneously. Dividing the image, it generates only 98 regions per image, which is far fewer than about 2000 regions generated using selective search method. Since YOLO uses the entire image for prediction,

it encodes contextual information about object classes, which results in reducing the prediction of false positives from the background.

YOLO is fast by discarding the region proposal generation step, running at 45 FPS, Fast YOLO can run at 155 FPS [212]. In spite of its speed, YOLO suffers from decrements in localization accuracy in comparison to two-stage framework methods such as Fast RCNN. The reason is the coarse division of the image, which results in less accurate locations, scales and aspect ratios of bounding boxes. Additionally, it may lead to failure in detecting small objects as each grid cell is only considered to contain one object type.

Redmone and Farhadi [210] later proposed YOLO V2, as an improved version of YOLO. In YOLO V2 the GoogleNet [250] in YOLO is replaced with a simpler network called DarkNet19, a batch normalization [104] method is added and also a better anchor box generation approach, which generates boxes with various sizes and aspect ratios, is used. Redmone and Farhadi [210] also introduced YOLO9000, which uses a joint optimization method to train on ImageNet and COCO datasets, detecting over 9000 object classes in real time. The next version of YOLO, YOLO V3 [211] was also proposed by Redmond and Farhadi, in which they improved the performance by using logistic regression to predict objectness score of each bounding box and a also replacing the DarkNet19 with a new network as a hybrid approach between the DarkNet19 and some perception from residual networks.

4. **SSD (Single Shot Detector):** proposed by Liu et al [153] to preserve real time detection without sacrificing detection accuracy, faster than YOLO [212] and accuracy comparable with two-stage framework methods such as Faster RCNN [215]. The main contribution of SSD is combining RPN idea from Faster RCNN, YOLO and multisclae convolutional features [100] to introduce multi reference and multi resolution detection techniques. Similar to YOLO, SSD generates a fixed number of BBs and scores for each, followed by a NMS to refine BBs and generate the final detection. The early layers of CNN network in SSD is based on VGG [238], followed by several convolutional layers with decreasing sizes towards the end of the network. SSD separates predictions by aspect ratio, making prediction of different scales from multiple convolution feature maps with different scales at the top layers of the network.

   SSD achieved high detection accuracy in real time, outperforming Faster RCNN in accuracy and YOLO in detection speed.

5. **CornerNet:** which is detecting objects as paired keypoints was proposed by Law et al. [134] by questioning the need of anchor boxes in one-stage framework methods, to detect objects. The authors discussed that to detect objects using anchor boxes, the system needs to generate a large number of anchors to ensure that at least one will sufficiently overlap with

the ground truth. This results in a significant imbalance between positive and negative examples, and slows down the training process. Additionally, using anchor boxes introduces extra hyper parameters including sizes and aspect ratios to the system. Therefore, CornerNet does not use anchor boxes for object detection, instead formulats object detection based on detecting and grouping keypoints. In this method top left and right bottom corners are detected and paired to form object bounding boxes. The idea of using keypoints for object detection was borrowed from the work on Associative Embedding for multiperson pose estimation [177]. The backbone network in CornerNet consists of two stacked Hourglass networks [178], followed by two corner pooling steps, one for top left corner and one for bottom right corner, with the purpose of improving corner localization.

CornerNet improved object detection performance, outperformed all previous unified framework based approaches and achieved competitive accuracy compared to the two-stage detectors on COCO dataset. However, the detection speed is limited to 4 FPS, which is slower than SSD and YOLO. Also, CornerNet may generate false BBs since it is difficult to determine which pairs of corners should be grouped into the same object.

Duan et al. [63] proposed CenterNet to further improve CornerNet, using triplets of key points, with an extra keypoint at the centre of a proposal, to detect objects. CenterNet improved the detection accuracy, but with slower speed than CornerNet.

5.3 Backbone Networks

One of the crucial steps in any detector is extracting features which can best describe an object of interest. Before the deep learning era, a great effort was devoted to extracting features manually, either based on visual cues of objects (see Chapter 3) or using local descriptors such as SIFT and HOG. Moreover, since an object could not be described using only one feature, approaches to group and abstract features such as Bag of Words [241] and Fisher Vector [196] were utilized. These feature representation algorithms demand high domain expertise and also careful engineering.

As opposed to traditional approaches for feature representation, deep CNNs are able to learn effective features from raw images [136]. It reduces the need of domain expertise and complex design of procedures. Therefore, in deep learning methods, the focus has been moved to design of networks with better architecture, and also to the training process, to achieve better feature representations.

In all detection methods presented in Section 5.2, the CNN architecture used to develop the network plays a critical role.The CNN architecture acts as the backbones of the detection network. Thus, a considerable amount of research in improving detection accuracy was dedicated to developing better

and deeper CNN networks. Therefore, in the followings we review some of the popular CNN architecture which are used in object detection tasks. The general trend in developing CNN architectures is to focus on greater depth. AlexNet [127] was the first CNN model developed, with eight layer deep and using 5x5 and 7x7 convolution filters. It started the deep learning revolution in the field. Then VGGNet [238] proposed by Oxford's Visual Geometry Group, increased the model depth to 16-19 layers and used smaller convolution filters (3x3), and in its time it achieved the highest performance on the ImageNet dataset. The next generation of CNN architectures was proposed by Google Inc. as GoogLeNet [250], and its following family known as Inception [250, 114, 251]. GoogLeNet increases both the width and depth (22 layers) of the network. The Inception family increased the number of layers up to 47 in Inception V3 [252], and also introduced the factorizing convolution and batch normalization in CNNs.
In the Deep Residual Networks (ResNet) [103], the number of layers increased up to 152, aiming at making the training process easier by reformulating its layers as learning residual functions. It shows the effectiveness of skip connections to learn extensive deep network. Network training was further accelerated in InceptionResNet [251] on the basis that shortcut connections (ResNet) can increase the training speed. It combines the Inception networks with short cut connections, inspired by ResNet. Extending the ResNet, DenseNet was proposed by Huang et al. [109]. DenseNet built from densely connected blocks, which connect each layer to every other layer in a feedforward fashion, such that each layer get supervision from other layers, using shorter connections. The next development of CNNs architectures was proposed by Hu et al. [103], developing Squeeze and Excitation (SE) blocks. Stacking a collection of these blocks a SE network (SENet) is constructed. Since structure of SE blocks are simple and computationally lightweight, components of any CNN architecture can be replaced by these blocks to enhance performance with minimal additional computational cost. The main contribution of SENet was in integration of global information in learning importance of features. The development of CNN architectures continues, with recently proposed methods such as Hourglass [178], Xception [43], DetNet [144] and GLoRe [39].
A considerable issue in developing a CNN architecture is the number of parameters, which directly affect training speed. While some earlier networks with few number of layers such as AlexNet, ZFNet [285], OverFeat [233], and VGGNet have huge number of parameters, most of these parameters come from the FC layers. Therefore, in newer proposed networks like ResNet and DenseNet, the FC layers are dropped, resulting in much fewer parameters.

Training a CNN needs a large scaled labeled dataset, with intraclass variety of all objects. It was shown [189] that using a pretrained deep model, trained on a large scale dataset with object level annotations, improves detection accuracy. The pretrained network is typically fine-tuned to fit the detection target objects, and increase detection performance. Since in the pretrained network

weights are optimized for the dataset which it was trained on, a fine-tuning step is required to update the network weights for the target task. Fine-tuning basically fills in the gap between the source and target datasets. Researchers also performed object detection and classification tasks on pretrained networks without fin-tuning them [61, 86], which showed that depending on the selected layer to extract features, detection accuracy differs. It is generally accepted that features from earlier layers are better in object localization while later ones work better for classification.

Despite the great success of DCNNs, still challenges remained to be addressed. Deep learning based methods required huge amounts of labeled data for training, extreme computational resources and expertise to select suitable hyper parameter, learning parameters and the network architectures. Moreover, it is usually difficult to fully understand and interpret trained networks, also DCNNs suffer from lack of robustness to degradation [152]. All these issues limit the use of deep neural network in real world application.

5.4 Bounding box refinement

In this section, bounding box regression and Hard Negative Mining (HNM) for bounding box refinement in deep learning based methods are briefly discussed.

1. **Bounding Box Regression:** before development of the Faster RCNN, bounding box regression was an individual post processing step in detection algorithms (see Section 4.3.1). After introduction of the Faster RCNN method, BB regression integrated with the detector, and also trained in an end-to-end fashion. To predict BBs more robustly, deep learning based BB regression often uses two main regression loss functions, either the smooth-L1 which is used in Faster RCNN [215], or the root-square, used in YOLO [212]. These two functions are less sensitive to outliers than the least square error which was used in DPM [73], which makes them a good candidate for BB regression. Furthermore, as another solution for BB regression, some researchers normalized the coordinates of bounding boxes to achieve more robust results [85, 153, 146].

2. **Hard Negative Mining:** in early times of developing deep learning based detectors, the bootstrap approach (see Section 4.3.1) for HNM was discarded, mainly due to increased computing power. Detection methods such as Faster RCNN and YOLO replaced bootstrapping with balancing the weights between the positive and negative windows. However, later it was shown that weight balancing does not resolve the problem [146]. Therefore, the bootstrap method was reintroduced into DCNN based detectors. As an example, the SSD method uses only the gradients of a small part of the negative samples, which have the highest loss value, to be back-propagated for training. As an alternate approach, researchers

reshaped the standard cross entropy loss functions in a fashion to put more focus on difficult and misclassified examples [294, 146].

5.5 Evaluation Methods

Performance of detection algorithms can be based on the detection speed (Frames Per Second (FPS)), precision and recall. Output of a detector is defined by three measures, bounding box location, predicted category and confidence that BB include a category of an object class. A detected region is considered TP if the overlap ratio (IOU) is greater than a predefined threshold and the predicted category of the object matches the ground truth label. Otherwise, it is a FP. Moreover, the confidence level is compared with a threshold to determine correctness of the predicted class label from the detected object. Based on TP and FP, precision and recall are computed as a function of the confidence threshold. By varying the confidence threshold, different pairs of precision and recall can be obtained, and therefore precision can be computed as a function of recall. From this *Average Precision*(AP) is derived. AP is the most common metric for performance evaluation of the detector and is computed separately for each class category. To compare performance over all object categories, the *mean AP* (mAP), average over all object categories, is used as the final measure of performance. It worth mentioning that in MS COCO dataset, more attention is given to BB localization, therefore instead of using a fixed IOU threshold, MS COCO AP is averaged over multiple IOU thresholds.

5.6 Deep Learning in Agricultural and Forestry Fields

DCNNs have been successfully applied to many machine learning based applications, including computer vision in agriculture (agrovision) and forestry. In the following, utilization of DCNNs in these fields is reviewed.

DCNNs are able to solve more complex problems by utilizing more complex models, which results in higher accuracy and lower regression errors. The hierarchical structure and large learning capacity of DCNNs result in high detection and classification performance, as well as flexibility and adaptability to a wide variety of challenges.

The convolution layers are different levels of feature representations, starting with lower level features (edge, color) from earlier layers, to high level features (more discriminative) in deeper layers. The convolutional layers act as automatic feature extractors, with dimensionality reduced by the pooling layers. The fully connected layers, which are usually located near the output of the network, act as classifiers that use higher level features (learned during the

training process) to classify objects into predefined classes or make numerical predictions.

Conventionally, researchers used hand engineered features to represent visual characteristics of the target fruit/tree, using them for localization and feeding them to a machine learning algorithm for classification. In this approach, extracting features was a time consuming and complex task which highly affected system performance. Moreover, selecting an effective set of feature required expertise and knowledge within the field. Additionally, manually crafted machine vision algorithm were not able to adaptively fit with environmental variations. In comparison, deep learning based methods provide deeper neural networks, a hierarchical feature representation using various convolutions, and most importantly drop the need of hand engineered feature extraction. This means larger learning capability, and results in higher detection and classification accuracy.

The reminder of this section is divided into sub-sections on agricultural and forestry areas where DCNNs are applied, CNN architectures researchers used, pre-processing approaches, different datasets, and methods for increasing the number of images in datasets.

1. **Areas of Use:** most applications of deep learning based algorithms can be categorized to plant or crop recognition, plant disease detection, identification of weeds, land cover classification, fruit counting [206] and classification and tree species detection. A large fraction of all research deal with object classification and identification, including obstacle detection [246].

2. **DCNNs Architectures:** researchers have utilized different structures of DCNNs to develop a network which are fit to the task. Generally it can be divided into approaches which use popular CNN architectures (such as AlexNet, VGGNet and ResNet), CNN models which were developed by the researchers, adopting first-order Differential Recurrent Neural Networks models (DRNN), and also using Long Short-Term Memory models (LSTM) [83]. In some of the works, the CNN model was combined with a classifier. In such approaches, CNNs are used to extract features using convolution layers and FC layers are replaced with classifiers like SVM [62, 186], logistic regression [40], large margin classifiers (LCM) [282] and linear regression [38]. Ostovar et al. [186], extracted features from a VGG-19 CNN and fed them to a classifier for detecting severity of rot in tree stumps [186].

   Furthermore, since it is difficult and time consuming to train a network from scratch with small datasets, some researchers, specifically those who used popular CNNs, took advantage of transfer learning (fine-tuning). In this approach, a pretrained CNN, which is already trained on large datasets such as ImageNet, COCO, is retrained to further leveraging the already existing knowledge to increase the learning efficiency. This approach was used in [15, 243, 45, 226, 246, 62, 157, 166, 186] for the

GoogLeNet, AlexNet, DenseNet and the VGG-16 architectures.

In the proposed methods for classification in the agricultural area, developed approaches could classified variable number of target objects, ranging from two [164, 157, 62] up to 1000 classes [62], which developed for plant identification including herb, tree and fern species. For example, Rebetes et al. [209], Lee et al. [138] and Xinshao and cheng [282] developed methods to recognize 44 plant species, classify weeds found in agricultural fields into 91 classes and identify 22 different crops, respectively. As a common procedure, the output of these CNNs were probability values, estimating the probability of the object belonging to a predefined class. The class with highest value was then selected as the predicted class. In some work, for counting fruits [206, 38] and localizing fruits in the scene [15, 226], the output are scalar values and multiple bounding boxes, respectively.

3. **Data Pre-processing:** similar to traditional approaches of detection and classification, also in deep learning based methods several pre-processing steps are utilized before feeding the image into the DCNNs with the aim of increasing detection accuracy. The most common procedure is image resizing, to adapt the image into the CNNs network requirement. Image segmentation is another practice, with the prupose of facilitating the learning process using ROIs [202, 226, 186, 94]or to increase the dataset size [283, 112]. Image segmentation could also make the annotation process easier [15]. In addition, foreground extraction [138] and background removal [164] are popular approaches. Conversion to other color spaces (HSV and grayscale) [5, 138] were also used as a pre-processing step. Also, generating bounding boxes [164, 38] and feeding them to a CNN architecture showed to be a useful approach to facilitate weed detection and fruit counting. Furthermore, some researchers used extracted features form images as input to the CNNs models,such as histogram [209], visual cues like shape [96], wavelet transformation [132], and Gray Level Co-occurrence Matrix (GLCM) features [228].

4. **Datasets:** The data used for training CNN models can be of three kinds: real images, which are collected by the researchers based on task needs [243, 15, 186], synthetically produced images [66, 206], or images from publicly available datasets such as LifeCLEF[1], MalayKew[2], Flavia[3] and Crop/Weed Field Image Dataset[4] [102].

   In general, more complicated tasks require larger datasets, for example object classification tasks involving a large number of classes [166], or

[1]https://www.imageclef.org/2014/lifeclef/plant
[2]http://web.fsktm.um.edu.my/∼ cschan/downloads_MKLeaf_dataset.html
[3]http://flavia.sourceforge.net/
[4]https://github.com/cwfid/dataset

with small variations between classes [175].

Since collecting a large number of images from environments such as orchards and forests is a difficult task and requires lots of considerations such as illumination conditions, occlusion, and distance to the object, some researchers employ augmentation techniques [127]. It helps to artificially enlarge the number of training images, improve the learning process and performance, and also enhance generalization ability of the network by providing varied data for training process. An augmentation process is particularly important when only small datasets for training is available [243, 175, 38], or when researchers train the network using synthetic images and test them on real data [206]. In these situations, augmentation provide networks with better generalization capabilities and also better adaptation to the real world difficulties. Data augmentation include approaches such as transformation, rotation, cropping, mirroring, translation, transposing, scaling, varying HSV channels and also adding shadows to images. One advantage with augmentation is that there is no need for extra manual labeling of newly generated images.

With the development of generative adversarial networks (GAN) [90], another method for image generation was introduced. In GANs two deep convolutional neural networks are trained simultaneously and adversarially: a generative model and a discriminative model. The aim of the generative model is to capture the feature distribution of a dataset by learning to generate images. The discriminative model evaluates generated images to determine how well they are similar to the dataset. Since both models are implemented as CNNs, the error can be back propagated to minimize the loss of both models simultaneously. The result after training is a generative model that can generate new images with high similarity to the learned dataset. Hence, GANs provide the same benefits as using augmentation techniques.

Some researchers compared DCNN based networks with other techniques, mostly conventional approaches for detection and classification, which were implemented only for comparison purposes within the same task. In most cases, CNN based methods outperform the other approaches. In [203, 92, 186] CNN achieved higher accuracy than a SVM classifier, work in [138] demonstrated that a CNN improved classification performance compared to ANN, and also a Random Forest (RF) classifier [161, 165]. Moreover, CNN showed higher performances than unsupervised feature learning [158], shape and color features [243, 65] and multilayer perceptrons [242, 131] based methods. Further comparisons with texture based regression models [38], Gaussian Mixture Models [228] and Naive Bayes classifiers [283] also proved superiority of CNNs. Furthermore, it is shown and supported by other work such as [175] that automatic feature extraction, using output of different convolutional layers, is more effective than both manually crafted features and also features extracted by

conventional methods such as SIFT, GLCM, texture, shape and color based algorithms, histogram and area based techniques (ABT).

To provide a better analysis of applications of deep learning in the agricultural and forestry fields, it is valuable to indicate advantages and disadvantages of utilizing DCNNs within the field.

5.6.1 Advantages of DCNNs within the Agricultural and Forestry Fields

Rather than improving the performance in classification or prediction tasks, reducing the effort for extracting features is demonstrated in many research works. Manually engineered feature extraction is a time consuming method that requires considerable effort, but is done automatically in DCNNs. It is worth mentioning that finding a reliable set of features or an appropriate feature extractor requires expertise, and also is a complex task.

Furthermore, DCNNs based approaches have proved to generalize well. For instance in [206] and [38], a developed system learned explicitly to count fruit, while also showing robustness to occlusion, illumination and scale variations. The same model could be used to identify several circular fruits such as citrus, peaches, and mangoes. Amara et al. [5] developed a DCNN based method to classify banana leaves. Their model was robust to environmental and imaging conditions such as varying illuminations, resolution, size and orientation, in addition to having the ability to handle complex backgrounds.

Despite the longer training time for DCNNs compared to conventional methods, it is much faster in detection and classification tasks, especially when one-stage framework approaches are utilized. It provides real time systems which is an essential need in agrovision and forestry. For example, the developed approach for detecting obstacles [45] based on a CNN architecture, took much longer time to train compared to SVM and KNN methods, but was extremely faster to use.

Additionally, deep learning based models are able to develop simulated datasets for training, which can be utilized for testing on real world images. This is a way to overcome the lack of training images, and the difficulties in collecting images in outdoor environments with different considerations such as varying illumination and different levels of occlusion.

5.6.2 Limitations and Disadvantages of DCNNs

A critical drawback of DCNN methods which limits their usage is the need of extremely large datasets for training. These problem can be serious, especialy when the researchers aim to develop the CNN model from scratch. Although simulation methods such as augmentation and GANs are available, depending on the complexity of the task (number of classes and expected performance), at least several hundred real images are required.

Problems with both collected and generated datasets are low variation among classes, existence of noise such as low resolution images, inaccurate imaging devices and occlusion. For example in [166, 226] authors determined and discussed that existence of more diverse training sets could increase classification performance.

Data annotation as a labour expensive task is a considerable problem, specifically within deep learning based method, with the need of large datasets. Depending on the complexity of the task, considerable time is needed for experts to accurately annotate the input images. This problem was experienced by Ostovar et al. [186], for detection of rot areas in tree stumps. For annotation, experts had to be hired or volunteered for the task. Additionally, annotators are always subject to errors and inaccuracies during the labeling process especially when it is a complex task.

DCNN based models can learn problems and even generalize them to some extent, but similar to conventional approaches, they cannot generalize the learned model beyond the dataset which was used for training them. However, in real outdoor environment, due to their dynamic and unstructured characteristics, unpredicted conditions such as different levels of occlusions, illuminations, variations in sizes and colors is a possible scenario. This problems was determined and discussed in some research works within agricultural field [206, 166, 138, 226, 45, 96].

Furthermore, within agriculture and forestry, there are not many publicly available datasets for researchers. This problem has existed from the early times of computer vision in these fields. The lack of available datasets brings negative effects to the field, such that it is not possible to create a benchmark to compare performance of developed systems for a specific category of problems. Additionally, researchers are required to spend considerable time on collecting images.

Chapter 6

Summary of Contributions and Thesis Conclusion

This chapter contains a summary of the contributed papers and also a brief conclusion of the thesis.

6.1 Summary of Contributions

6.1.1 Paper I: Detection of Trees Based on Quality Guided Image Segmentation

In this paper, a part of CROPS project, the aim is identifying trees in forestry environments using color images. The results may be used in autonomous forest machines for tree detection for harvesting operations, and also to prevent collisions by autonomous timber carriers in the forest.

We integrated a novel method for seed point generation with a segmentation method similar to region growing. Experiments showed that conversion to HSV results in highest performance compared to alternative colour-spaces. For seed point generation considering the vertical orientation of trees, the hue matrix is vertically compressed. This speeds up the process by decreasing the dimensionality of the analyzed matrix, and also improve the seed point generation. The outliers in the compressed hue matrix represent seed points. Each seed point in the image is then used to segment the entire image based on its neighbouring pixels. In order to increase the detection rate, a series of morphological operations are applied to the generated segments. The morphological operations are designed considering the vertical nature of trees. After applying a quality function to all segments, segments that cover the same seed point are compared and the one with highest quality value is assigned to it. It results in selecting the optimal segmentation for each seed point. The set of selected segments indicates trees in the images.

The method is evaluated using images from forest environments in northern Sweden. The proposed method successfully detected 171 of the 197 trees, giving the recall and precision rate of 86.8% and 81.4%.

I was contributing in formulation of the algorithm, and also to its implementation. I also performed the data analysis and system evaluation, and contributed in writing the draft and revising the manuscript.

6.1.2 Paper II: Adaptive Image Thresholding of Yellow Peppers for a Harvesting Robot

Due to varying illumination conditions causing overexposure and shadows and also unstructured environment of greenhouses, fixed thresholds do not work well for object detection. Therefore in this work, as part of the SWEEPER project, we proposed a novel method for adaptive image-dependent thresholding of yellow peppers based on reinforcement learning (RL).

First, RGB images are converted to HSV color-space, and then a set of features based on Hue and Saturation values are extracted. These features are used to define *states*. The agent then performs an *action* based on exploration-exploitation strategy for the *state*. Each *action* is a vector with four parameters: $\{H_{min}, H_{max}, S_{min}, S_{max}\}$ for image thresholding. Each parameter can get a value between 0 to 1 with 0.05 steps. In the next step, the image is segmented using the defined *action*, and the system computes the reward value. The *reward* value is the similarity between the segmented image, achieved by applying the *action*, and the labelled image. To compute the similarity, we introduced two measures, segmentation-overlap and segmentation-efficiency based on the ratio between the number of pixels in the overlap region, segmented region and the labelled region. This process continues until the agent meets the convergence threshold. It worth mentioning that in this work, the transition function is defined to be non-deterministic, returning a new representation of the image, a segmented image, instead of returning a new *state*.

We compared results from three exploration-exploitation strategies with a benchmark. Exploration-exploitation strategies included decaying epsilon-greedy, epsilon greedy with values 0.2, 0.5 and 0.7, and also a novel strategy denoted Q-value difference measurement, which switches between exploration and exploitation strategies adaptively. The benchmark is an exhaustive search on all actions to determine the best threshold for each image. Results showed that decaying epsilon greedy strategy achieved highest performance, 91.5% of the benchmark with 73% fewer iterations.

I was the main author of the paper. I proposed the framework to use Reinforcement Learning for adaptive thresholding, developed the algorithm and implemented it. I conducted the data analysis and system evaluation. I also prepared the draft and participated in reviewing and editing the manuscript.

6.1.3 Paper III: Integrating Kinect Depth Data with a Stochastic Object Classification Framework for Forestry Robots

As part of the CROPS project, in this work we developed techniques to detect trees, bushes, stones and humans in images from forestry environments. The solutions may be used for autonomous harvesters to detect surrounding objects and based on their categories performs the best action. For this purpose we integrated an RGB camera and a depth sensor with the hypothesis that adding extra data (depth) would increase classification performance.

The major contribution of this research is analysing the effect of adding depth data to RGB data for object detection and classification. Depth data is used for two purposes, labelling of the RGB images (extracting foregrounds) and classification. We extracted 12 color-based features from RGB images, and six depth-based features from depth images. These features were then fed into five classifiers, once using only RGB features and once using RGB and depth features. The classifiers were K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB) and Linear Discriminant Analysis (LDA). All classifiers outputs were then combined using the Weighted Majority Vote (WMV) method.

By using only RGB based features, the system achieved 93.04% performanc. By integrating depth based features, performance was increased to 96.20% which proves the hypothesis that adding depth data increases the classification performance.

I evaluated available 3D sensors to be used in the project and I also collected all the data. I contributed to algorithm development and programming. System evaluation was mainly done by me, and I also participated in drafting the manuscript, reviewing and editing it.

6.1.4 Paper IV: Detection and classification of Root and Butt-Rot (RBR) in Stumps of Norway Spruce Using RGB Images and Machine Learning

Rot significantly affect quality of timber, and consequently reduce economic outcome of harvesting operation. It is therefor essential to detect rot while harvesting, to increase the accuracy of timber sorting and also to determine which trees are affected, to be able to perform further treatments to prevent spread of rot to other areas in the forest. As part of the PRECISION project, we developed a system to automatically detect stumps and classify them based on the presence or absence of rot. Additionally, stumps were classified into three classes of infestation, rot $= 0\%$, $0\% <$ rot $< 50\%$ and rot $\geq 50\%$.

In this work to detect stumps we used a deep learning based method, Faster R-CNN. To classify rot in detected stumps, we proposed three methods using both conventional machine learning algorithms and deep learning approaches. In the first method, based on conventional learning approach, we used Bag

of Features (BoF) method to extract features and then fed them to a SVM classifier. In the other classifier we merged conventional machine learning and deep learning approaches, extracted features from FC-7 layer of VGG-19 network and fed them to a SVM classifier. In the pure deep learning classifier, we fine-tuned a VGG-19 network for the classification of rot. Moreover, to compute the segmentation quality, determining how well detected stumps matches with the manually labelled stumps, we used measures from the previous work, segmentation-overlap and segmentation-efficiency.

The results showed that tree stumps were detected with 95% precision and 80% recall rate. Using fine-tuned VGG-19, it achieved the highest classification performance. Stumps with and without RBR were correctly classified with accuracy of 83.5% and 77.5%. Also, classifying rot into three classes, stumps with rot = 0%, 0% $<$ rot $<$ 50% and rot $\geq$ 50% were classified with 79.4%, 72.4% and 74.1% accuracy, respectively.

I was the main author of the paper and participated in formulation of the problem. I developed algorithms, implemented them, and evaluated the system. Moreover, I lead a team of students for data collection and manual image labelling of the dataset. I drafted the manuscript and contributed in the reviewing and editing process.

6.1.5 Paper V: A Direct Method for 3D Hand Pose Recovery

Using 3D hand gesturs for human-computer/robot interaction is a common practice. Variability of hand gestures and flexibility of hand movements could results to developing robust gesture recognition systems in real-time. In this work, we address natural and immersive hand gesture interaction by combining two tasks, gesture recognition and gesture pose estimation using 3D data acquired by Kinect for 3D object manipulation. We introduced a method to reduce the complexity of hand pose recovery from 27 DOF to only 6 DOF global hand motion.

The system input is a sequence of depth images. In the first step, we reduce noise and smooth the images, then apply a head detection method followed by a segmentation approach to localize and track hands in the sequence. Then, we estimate hand poses using the new optical flow constraint equation. In the final stage, we utilized extracted hand pose parameters for 3D object manipulation. To localize hands, first we extracted human body as foreground and detect the head as the topmost body part, then based on its depth value we predict hand depth value, expecting to be located within a particular distance from the head. Afterwards, the hand center point is extracted and a region of interest (ROI) defined enclosing it. Using this approach, we just need to process ROI to segment the hand in sequence of next frames. When hands are localized and tracked in image sequences, we compute gesture pose parameters using optical flow techniques.

Since for the system evaluation we needed ground truth data, which was

not available, we used Active Motion Capture system by mounting a camera on the user's hand to measure three rotational parameters in each frame. In this system, SIFT features are extracted from the environment and tracked in the next frames. Therefore, we are able to find point correspondences between two frames and consecutively compute the relative motion parameters by analyzing point correspondences. Additionally, we used displacement of hand center point to compute translation motion parameters. These six parameters provide the ground truth data for the evaluation.

Two determine the system accuracy, two criterias, Absolute Rotation Error (ARE) and Absolute Displacement Error (ADE) are measured using actual and estimated values. The best performance achieved where mean ARE is 6.7° and mean ADE is 10.5 mm.

The problem formulation was jointly done by all authors. I designed and implemented the algorithm and performed data analysis. For the system evaluation I created a mockup to test the system in initial steps. I also participated in writing the draft and revising the manuscript.

6.1.6 Paper VI: Human Detection Based on Infrared Images in Forestry Environments

Since safety of humans is the main consideration for both manned and autonomous harvesters, in this work we used a thermal camera to detect the presence of humans in close range of the forestry machines to prevent any harm to humans. Detecting humans in forestry environment using RGB cameras is challenging due to the lighting conditions and occlusion from trees and bushes, therefore, we utilized the thermal camera. Two approaches were proposed and compared: shape-dependent and shape-independent.

To detect humans, we first extracted regions of interest (ROIs) by thresholding thermal images based on pixel intensity. Then based on the approach, features from ROIs were extracted and fed into either a heuristically designed decision rule or three classifiers, SVM, KNN and NB, to determine presence of a human in the ROIs. In the shape-dependent approach, extracted features were based on human characteristics such as shape, height, length and location of the head as the hottest spot in the body. Extracted features in the shape-independent approach included statistical characteristics of thermal images. To evaluate the overlap between extracted ROIs and manually labeled ones, we used two concepts, denoted side-accuracy and side-efficiency. They quantify how much of the manually labelled ROI is covered by the extracted ROI, and how much of the extracted ROI is covered by the labelled ROI. Definition of true positive, false positive and false negative for accuracy evaluation of the system was set based on the overlap threshold values and the class prediction.

The shape-dependent achieved 79% precision and 38% recall, while the shape-independent appraoch reached 80% precision and 43% recall.

I was the main author of this paper. I proposed the framework, developed the algorithms and implemented them. All data collection, analysis, curation

and evaluation was done by me. I also wrote the draft manuscript and contributed in reviewing and editing processes.

6.1.7 Paper VII: Natural Language Guided Object Retrieval in Images

In this work, we proposed a method for generating responses to natural language queries regarding objects and their spatial relations in images. We merged a computer vision based method for object detection and classification, YOLO, and natural language processing approaches for analysing input queries. The system response includes identification of objects in images based on the query, and also generation of an appropriate text that answers the query.

The proposed system contains three parts for processing the input image and the text query: image analysis, spatial relation analysis and text query analysis. In the image analysis part, utilizing YOLO, objects are detected and classified, output bounding boxes along with object labels. Bounding boxes and their categories are fed into spatial relation analysis part, to predict spatial relations between bounding boxes and objects. It consist of three classifiers, a Multi-Layer Perceptron (MLP), a KNN and a SVM. Output of these classifiers are integrated using a Weighted Average Probability network (WAP) and generate spatial relation words such as "right", "top", "front", for each pair of bounding boxes. The text query analysis part is responsible for analysing the input query, extracting tuples, describing objects and their spatial relations. Output of these three parts are used for natural language grounding, which map the tuples from the input text query to the bounding boxes, object categories, and spatial relations. To achieve the most probable grounding, we combine probabilistic measures of object classes and spatial relations with a measure of word semantic similarity between extracted tuples from the text query, object classes and their spatial relation. The designed algorithm can respond to three type of text questions: attention queries, e.g. "find the person to the right of the monitor", relation queries, e.g. "where is the person?" and classification queries, e.g. "what is to the right of the monitor?".

To evaluate system performance, we measure how well the generated answers match human assessment. For this purpose, 30 test users were involved. The users were asked to compose three questions, based on three type of queries the system can answer, for each input image and evaluate the generated answers. Results demonstrated that the system correctly answered 81.9% of all questions.

I was the main author of the paper. I contributed to the formulation of the problem statement, and to development of algorithms, in particular the parts on object detection and deep learning. I also contributed to the parts on query analysis based on NLP and the natural language grounding. I implemented all parts, performed data analysis and evaluated the system. Moreover, I drafted the manuscript and contributed in reviewing and editing processes.

6.2 Thesis Conclusion

This thesis contributes to analysis of three main criterion in developing autonomous systems in agricultural and forestry environments including imaging sensors, visual cues of target objects and the image analysis algorithms. These algorithms are divided into object detection (localization) and recognition (classification) methods and following the evolution of the computer vision and machine learning approaches, both conventional and deep learning based methods are described.

Object detection in uncontrolled outdoor environment is a complex task, suffers mainly from unstructured environments, illumination changes, occlusions, shadows and various size, shape and colour of target objects, which results to degraded performance. Furthermore, hand-engineered feature extraction and dealing with setting a large number of parameters in image analysis algorithms required expertise, which are considered as a common source of reduction in accuracy of autonomous systems in this field. Utilizing deep learning based methods improve the performance, by dropping the need of hand crafted feature extraction and reducing the number of parameters. However, these approaches require a large number of annotated images for training, are usually slow in the training process, and are also computationally expensive.

There are some other limitations in this field, including: 1) lack of datasets for objects like fruits and trees, which results in researchers not being able to compare their results with others using a benchmark, 2) data collection is a time consuming process in which many considerations should take into account, such as acquiring images with different illumination conditions and levels of occlusions to simulate real working environments, 3) Image annotation, which is needed for the training process, is a tedious process, and in some cases it needs expertise to achieve accurate data annotations. Moreover, annotated data usually contains errors which have to be taken into account when developing an autonomous system.

Therefore, despite extensive research works in automation of agricultural and forestry operations, the performance of these systems are still not sufficient, and developed systems are not completely ready for the market. In terms of Technology Readiness Levels (TRL), currently many agricultural and forestry robotics systems are available as prototypes, positioned between level 8, system development, and level 9, proving the system in operational environments. Regarding the CROPS project, it reached level 6 which is demonstrating the system in relevant environments. The SWEEPER project, with the aim of developing a sweet pepper harvesting robot, achieved one part of level 9 by testing the robot in real greenhouse environment during summer 2018. The PRECISION project is still in its initial steps, on the intersection between level 3, experimental proof of concept, and level 4, technology validation in lab.

In this thesis, describing each of three criterion, imaging sensors, visual cues and image analysis algorithms, we also discussed their limitations and possibil-

ity of overcoming any of aforementioned challenges. The thesis contributions also include scientific publications. In Papers I, VI (both based on the CROPS project) and III, the focus is on detecting and recognizing objects in forestry environments. In Paper I we developed a tree detection approach, as an initial step in developing an autonomous harvester. Since safety for humans is one of the most important considerations when developing autonomous forest machines, we address this issue by developing a human detection algorithm based on infrared images in Paper VI. Using such a system, an autonomous machine would either stop the cutting process or alert the machine operator to prevent any harm, if a human is detected close to the vehicle.

Autonomous vechicles should be able to detect and classify also other types of surrounding objects. Trees should be identified such that a harvester can get closer to be able to harvest. If a stone is detected in front of an autonomous machine, it sometimes should be considered as an obstacle and the vehicle should re-plan the path to avoid it. We address detection and classification of trees, bushes, humans and stones in forestry environment using both RGB images and depth data in Paper III.

Pre-selected thresholds for segmentation of images is problematic in unstructured environments such in greenhouses due to the similarity between foreground and background features. Therefore in Paper II (as part of the SWEEPER project) we address this problem by proposing a novel method for adaptive image thresholding based on reinforcement learning (RL) to localize sweet peppers in greenhouse environments. As part of the PRECISION project, in Paper IV, we developed an autonomous system using RGB images to detect tree stumps and classify them based on presence or absence of rot using Faster-RCNN for detection and three methods for classification based on both conventional and deep learning algorithms. This system, as part of the whole project, aims at increased performance when sorting timbers at cutting time.

Merging works in Papers I, III, IV and VI, we could have an autonomous forest vehicle, able to detect trees, determine obstacles in the way to the tree for harvesting, identifying stumps and classify the harvested trees based on the presence of rot. Additionally, it could be equipped with a safety module to detect the presence of humans in close range, and accordingly either change the direction of motion, stop the cutting process, or stop the entire vehicle.

In Papers V and VII, more attention have been put to developing systems to provide interaction with robots. In Paper V, we designed a system to estimate 3D hand gestures using depth data acquired from Kinect camera to enable users to interact with an augmented environment. Moreover, as an emerging methodology, in Paper VII, we have merged computer vision techniques for object detection with natural language processing (NLP) methods for analysing user queries, to develop a system to accurately answer queries regarding objects and their spatial relation in images. Further development of the work in Paper V, it is possible to use it as a training system for operators of forestry harvesting systems, to develop semi-autonomous harvesters, controlling the harvester

crane towards the tree and also the gripping process, using hand gestures. It brings some additional benefits such that the operators can control harvesting machineries from any location, so they do not have to be in forest with difficult environmental conditions such as extensive temperatures.

Bibliography

[1] Irshad Ahmad et al. "Weed classification based on Haar wavelet transform via k-nearest neighbor (k-NN) for real-time automatic sprayer control system". In: *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. ACM. 2011, p. 17.

[2] V Alchanatis et al. "Apple yield mapping using hyperspectral machine vision". In: *Precision agriculture'07. Proceedings of the 6th European Conference on Precision Agriculture*. 2007, pp. 555–562.

[3] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. "Measuring the objectness of image windows". In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2189–2202.

[4] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. "What is an object?" In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 73–80.

[5] Jihen Amara, Bassem Bouaziz, Alsayed Algergawy, et al. "A Deep Learning-based Approach for Banana Leaf Diseases Classification." In: *BTW (Workshops)*. 2017, pp. 79–88.

[6] Zhao De-An et al. "Design and control of an apple harvesting robot". In: *Biosystems engineering* 110.2 (2011), pp. 112–122.

[7] Boaz Arad et al. "Controlled Lighting and Illumination-Independent Target Detection for Real-Time Cost-Efficient Applications. The Case Study of Sweet Pepper Robotic Harvesting". In: *Sensors* 19.6 (2019), p. 1390.

[8] Pablo Arbelaez et al. "Contour detection and hierarchical image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2010), pp. 898–916.

[9] Pablo Arbeláez et al. "Multiscale combinatorial grouping". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 328–335.

[10] C Wouter Bac et al. “Harvesting robots for high-value crops: State-of-the-art review and challenges ahead”. In: *Journal of Field Robotics* 31.6 (2014), pp. 888–911.

[11] Johan Baeten et al. “Autonomous fruit picking machine: A robotic apple harvester”. In: *Field and service robotics*. Springer. 2008, pp. 531–539.

[12] XD Bai et al. “Crop segmentation from images by morphology modeling in the CIE L* a* b* color space”. In: *Computers and electronics in agriculture* 99 (2013), pp. 21–34.

[13] Qadeer Baig et al. “Fusion between laser and stereo vision data for moving objects tracking in intersection like scenario”. In: *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2011, pp. 362–367.

[14] Sudheer Reddy Bandi, A Varadharajan, and A Chinnasamy. “Performance evaluation of various statistical classifiers in detecting the diseased citrus leaves”. In: *International Journal of Engineering Science and Technology* 5.2 (2013), pp. 298–307.

[15] Suchet Bargoti and James Underwood. “Deep fruit detection in orchards”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 3626–3633.

[16] Sabine D Bauer, Filip Korč, and Wolfgang Förstner. “The potential of automatic methods of classification to identify leaf diseases from multispectral images”. In: *Precision Agriculture* 12.3 (2011), pp. 361–377.

[17] Avital Bechar and Clément Vigneault. “Agricultural robots for field operations: Concepts and components”. In: *Biosystems Engineering* 149 (2016), pp. 94–111.

[18] Christian Beder, Bogumil Bartczak, and Reinhard Koch. “A comparison of PMD-cameras and stereo-vision for the task of surface reconstruction using patchlets”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[19] Bir Bhanu and Ioannis Pavlidis. *Computer vision beyond the visible spectrum*. Springer Science & Business Media, 2006.

[20] Ashutosh Kumar Bhatt, Durgesh Pant, and Richa Singh. “An analysis of the performance of Artificial Neural Network technique for apple classification”. In: *AI & society* 29.1 (2014), pp. 103–111.

[21] Wei Bian and Dacheng Tao. “Max-min distance analysis by using sequential SDP relaxation for dimension reduction”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.5 (2010), pp. 1037–1050.

[22] Phil Blunsom. “Hidden markov models”. In: *Lecture notes, August* 15.18-19 (2004), p. 48.

[23] Derek Bradley and Gerhard Roth. “Adaptive thresholding using the integral image”. In: *Journal of graphics tools* 12.2 (2007), pp. 13–21.

[24] D M. Bulanon and T Kataoka. “Fruit detection system and an end effector for robotic harvesting of Fuji apples”. In: *Agricultural Engineering International: CIGR Journal* 12.1 (2010).

[25] DM Bulanon, TF Burks, and V Alchanatis. “Image fusion of visible and thermal images for fruit detection”. In: *Biosystems engineering* 103.1 (2009), pp. 12–22.

[26] DM Bulanon, TF Burks, and V Alchanatis. “Study on temporal variation in citrus canopy using thermal imaging for citrus fruit detection”. In: *Biosystems Engineering* 101.2 (2008), pp. 161–171.

[27] Duke M Bulanon, Thomas F Burks, and Victor Alchanatis. “A multispectral imaging analysis for enhancing citrus fruit detection”. In: *Environmental Control in Biology* 48.2 (2010), pp. 81–91.

[28] Duke M Bulanon, Hiroshi Okamoto, and Shun-Ichi Hata. “Feedback control of manipulator using machine vision for robotic apple harvesting”. In: *2005 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2005, p. 1.

[29] Duke M Bulanon et al. “Development of a real-time machine vision system for the apple harvesting robot”. In: *SICE 2004 Annual Conference*. Vol. 1. IEEE. 2004, pp. 595–598.

[30] Christopher JC Burges. “A tutorial on support vector machines for pattern recognition”. In: *Data mining and knowledge discovery* 2.2 (1998), pp. 121–167.

[31] Ali Caglayan, Oguzhan Guclu, and Ahmet Burak Can. “A plant recognition approach using shape and color features in leaf images”. In: *International Conference on Image Analysis and Processing*. Springer. 2013, pp. 161–170.

[32] Zhaowei Cai and Nuno Vasconcelos. “Cascade r-cnn: Delving into high quality object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6154–6162.

[33] Yüksel Çakır et al. “Detection of oranges in outdoor conditions”. In: *2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*. IEEE. 2013, pp. 500–503.

[34] Martha Cardenas-Weber, Amots Hetzroni, and Gaines E Miles. “Machine vision to locate melons and guide robotic harvesting”. In: *Paper-American Society of Agricultural Engineers (USA)* (1991).

[35] Joao Carreira and Cristian Sminchisescu. “Constrained parametric min-cuts for automatic object segmentation”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 3241–3248.

[36] Joao Carreira and Cristian Sminchisescu. "CPMC: Automatic object segmentation using constrained parametric min-cuts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.7 (2011), pp. 1312–1328.

[37] Supawadee Chaivivatrakul and Matthew N Dailey. "Texture-based fruit detection". In: *Precision Agriculture* 15.6 (2014), pp. 662–683.

[38] Steven W Chen et al. "Counting apples and oranges with deep learning: A data-driven approach". In: *IEEE Robotics and Automation Letters* 2.2 (2017), pp. 781–788.

[39] Yunpeng Chen et al. "Graph-based global reasoning networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 433–442.

[40] Yushi Chen et al. "Deep learning-based classification of hyperspectral data". In: *IEEE Journal of Selected topics in applied earth observations and remote sensing* 7.6 (2014), pp. 2094–2107.

[41] Ming-Ming Cheng et al. "BING: Binarized normed gradients for objectness estimation at 300fps". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 3286–3293.

[42] Yu-Tseh Chi and Peter P Ling. "Fast fruit identification for robotic tomato picker". In: *2004 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2004, p. 1.

[43] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[44] Sujan Chowdhury, Brijesh Verma, and David Stockwell. "A novel texture feature based multiple classifier technique for roadside vegetation classification". In: *Expert Systems with Applications* 42.12 (2015), pp. 5047–5055.

[45] Peter Christiansen et al. "DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field". In: *Sensors* 16.11 (2016), p. 1904.

[46] Chia-Lin Chung et al. "Detecting Bakanae disease in rice seedlings by machine vision". In: *Computers and Electronics in Agriculture* 121 (2016), pp. 404–411.

[47] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. "Weakly supervised object localization with multi-fold multiple instance learning". In: *IEEE transactions on pattern analysis and machine intelligence* 39.1 (2016), pp. 189–203.

[48] Oded Cohen, Raphael Linker, and Amos Naor. "Estimation of the number of apples in color images recorded in orchards". In: *International Conference on Computer and Computing Technologies in Agriculture*. Springer. 2010, pp. 630–642.

[49] S.L. Eddins C.R. Gonzalez R.E. Woods. *Digital Image Processing Using Matlab (second ed.)* McGraw Hill Companies, New Delhi, 2010.

[50] Gordana Dodig Crnkovic. "Constructive research and info-computational knowledge generation". In: *Model-Based Reasoning in Science and Technology*. Springer, 2010, pp. 359–380.

[51] Gabriella Csurka et al. "Visual categorization with bags of keypoints". In: *Workshop on statistical learning in computer vision, ECCV*. Vol. 1. 1-22. Prague. 2004, pp. 1–2.

[52] Jifeng Dai et al. "R-fcn: Object detection via region-based fully convolutional networks". In: *Advances in neural information processing systems*. 2016, pp. 379–387.

[53] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: 2005.

[54] Hongshe Dang, Jinguo Song, and Qin Guo. "A fruit size detecting and grading system based on image processing". In: *2010 Second International Conference on Intelligent Human-Machine Systems and Cybernetics*. Vol. 2. IEEE. 2010, pp. 83–86.

[55] Jnaneshwar Das et al. "Devices, systems, and methods for automated monitoring enabling precision agriculture". In: *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE. 2015, pp. 462–469.

[56] François-Michel De Rainville et al. "Bayesian classification and unsupervised learning for isolating weeds in row crops". In: *Pattern Analysis and Applications* 17.2 (2014), pp. 401–414.

[57] Debadeepta Dey, Lily Mummert, and Rahul Sukthankar. "Classification of plant structures from uncalibrated image sequences". In: *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*. IEEE. 2012, pp. 329–336.

[58] E d'Grand et al. "J. Magali: A self-propelled robot to pick apples". In: *American Society of Agricultural Engineering Paper* 46 (1987), pp. 353–358.

[59] Y Dobrusin et al. "Real-time image processing for robotic melon harvesting". In: *Paper-American Society of Agricultural Engineers (USA)* (1992).

[60] Piotr Dollár and C Lawrence Zitnick. "Fast edge detection using structured forests". In: *IEEE transactions on pattern analysis and machine intelligence* 37.8 (2014), pp. 1558–1570.

[61] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[62] Clement DOUARRE et al. “Deep learning based root-soil segmentation from X-ray tomography”. In: *bioRxiv* (2016), p. 071662.

[63] Kaiwen Duan et al. “CenterNet: Keypoint Triplets for Object Detection”. In: *arXiv preprint arXiv:1904.08189* (2019).

[64] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[65] Mads Dyrmann, Henrik Karstoft, and Henrik Skov Midtiby. “Plant species classification using deep convolutional neural network”. In: *Biosystems Engineering* 151 (2016), pp. 72–80.

[66] Mads Dyrmann et al. “Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network”. In: *Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark*. 2016, pp. 26–29.

[67] Yael Edan et al. “Robotic melon harvesting”. In: *IEEE Transactions on Robotics and Automation* 16.6 (2000), pp. 831–835.

[68] Ian Endres and Derek Hoiem. “Category independent object proposals”. In: *European Conference on Computer Vision*. Springer. 2010, pp. 575–588.

[69] Dumitru Erhan et al. “Scalable object detection using deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 2147–2154.

[70] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[71] Yajun Fang et al. “A shape-independent method for pedestrian detection with far-infrared images”. In: *IEEE Transactions on Vehicular Technology* 53.6 (2004), pp. 1679–1697.

[72] Pedro F Felzenszwalb et al. “Object detection with discriminatively trained part-based models”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2009), pp. 1627–1645.

[73] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. “A discriminatively trained, multiscale, deformable part model”. In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[74] Guo Feng, Cao Qixin, and Nagata Masateru. “Fruit detachment and classification method for strawberry harvesting robot”. In: *International Journal of Advanced Robotic Systems* 5.1 (2008), p. 4.

[75] Jie Feng et al. “Salient object detection by composition”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 1028–1035.

[76] Juan Feng et al. "A novel 3D laser vision system for robotic apple harvesting". In: *2012 Dallas, Texas, July 29-August 1, 2012*. American Society of Agricultural and Biological Engineers. 2012, p. 1.

[77] Steven A Fennimore and Douglas J Doohan. "The challenges of specialty crop weed control, future directions". In: *Weed Technology* 22.2 (2008), pp. 364–372.

[78] Martin A Fischler and Robert A Elschlager. "The representation and matching of pictorial structures". In: *IEEE Transactions on computers* 1 (1973), pp. 67–92.

[79] Davinia Font et al. "A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm". In: *Sensors* 14.7 (2014), pp. 11557–11579.

[80] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[81] H Gan et al. "Immature green citrus fruit detection using color and thermal images". In: *Computers and electronics in agriculture* 152 (2018), pp. 117–125.

[82] Jordi Gené-Mola et al. "Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities". In: *Computers and Electronics in Agriculture* 162 (2019), pp. 689–698. ISSN: 0168-1699.

[83] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM". In: (1999).

[84] Soumi Ghosh and Sanjay Kumar Dubey. "Comparative analysis of k-means and fuzzy c-means algorithms". In: *International Journal of Advanced Computer Science and Applications* 4.4 (2013).

[85] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[86] Ross Girshick et al. "Region-based convolutional networks for accurate object detection and segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2015), pp. 142–158.

[87] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[88] Aleana Gongal, Suraj Amatya, and Manoj Karkee. "Identification of repetitive apples for improved crop-load estimation with dual-side imaging". In: *2014 Montreal, Quebec Canada July 13–July 16, 2014*. American Society of Agricultural and Biological Engineers. 2014, p. 1.

[89] A Gongal et al. "Sensors and systems for fruit detection and localization: A review". In: *Computers and Electronics in Agriculture* 116 (2015), pp. 8–19.

[90] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[91] Tony Grift et al. "A review of automation and robotics for the bioindustry". In: *Journal of Biomechatronics Engineering* 1.1 (2008), pp. 37–54.

[92] Guillermo L Grinblat et al. "Deep learning for plant identification using vein morphological patterns". In: *Computers and Electronics in Agriculture* 127 (2016), pp. 418–424.

[93] Chunhui Gu et al. "Recognition using regions". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 1030–1037.

[94] Haiyan Guan et al. "Deep learning-based tree classification using mobile LiDAR data". In: *Remote Sensing Letters* 6.11 (2015), pp. 864–873.

[95] Maria Guijarro et al. "Automatic segmentation of relevant textures in agricultural images". In: *Computers and Electronics in Agriculture* 75.1 (2011), pp. 75–83.

[96] David Hall et al. "Evaluation of features for leaf classification in challenging conditions". In: *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE. 2015, pp. 797–804.

[97] Michael W Hannan and Thomas F Burks. "Current developments in automated citrus harvesting". In: *2004 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2004, p. 1.

[98] MW Hannan, TF Burks, and Duke M Bulanon. "A machine vision algorithm combining adaptive segmentation and shape analysis for orange fruit detection". In: *Agricultural Engineering International: CIGR Journal* (2010).

[99] MW Hannan, TF Burks, DM Bulanon, et al. "A real-time machine vision algorithm for robotic citrus harvesting". In: *2007 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2007, p. 1.

[100] Bharath Hariharan et al. "Object instance segmentation and fine-grained localization using hypercolumns". In: *IEEE transactions on pattern analysis and machine intelligence* 39.4 (2016), pp. 627–639.

[101] Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. "Combining efficient object localization and image classification". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 237–244.

[102] Sebastian Haug and Jörn Ostermann. "A Crop/Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks". In: *Computer Vision - ECCV 2014 Workshops*. 2015, pp. 105–116. DOI: 10.1007/978-3-319-16220-1_8. URL: http://dx.doi.org/10.1007/978-3-319-16220-1_8.

[103] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[104] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[105] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[106] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[107] Thomas Hellström and Ahmad Ostovar. “Detection of trees based on quality guided image segmentation”. In: *Proceedings of the Second International RHEA Conference, Madrid, Spain*. 2014, pp. 21–23.

[108] Jan Hosang et al. “What makes for effective detection proposals?” In: *IEEE transactions on pattern analysis and machine intelligence* 38.4 (2015), pp. 814–830.

[109] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[110] Soonmin Hwang et al. “Multispectral pedestrian detection: Benchmark dataset and baseline”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1037–1045.

[111] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. “Independent component analysis, adaptive and learning systems for signal processing, communications, and control”. In: *John Wiley & Sons, Inc* 1 (2001), pp. 11–14.

[112] Dino Ienco et al. “Land cover classification via multitemporal spatial data by deep recurrent neural networks”. In: *IEEE Geoscience and Remote Sensing Letters* 14.10 (2017), pp. 1685–1689.

[113] John Illingworth and Josef Kittler. “A survey of the Hough transform”. In: *Computer vision, graphics, and image processing* 44.1 (1988), pp. 87–116.

[114] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015).

[115] Wei Ji et al. “Automatic recognition vision system guided for apple harvesting robot”. In: *Computers & Electrical Engineering* 38.5 (2012), pp. 1186–1195.

[116] Antonio Ramón Jiménez, R Ceres, and Jose L Pons. "A vision system based on a laser range-finder applied to robotic fruit harvesting". In: *Machine Vision and Applications* 11.6 (2000), pp. 321–329.

[117] Xing-Jian Jing. "Behavior dynamics based motion planning of mobile robots in uncertain dynamic environments". In: *Robotics and autonomous systems* 53.2 (2005), pp. 99–123.

[118] Vijay John et al. "Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks". In: *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE. 2015, pp. 246–249.

[119] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[120] F Juste et al. "Primeros resultados en campo de un prototipo de brazo robotizado para la recolección de cıtricos". In: *CIMA91, 23 Conf. Int. Maquinaria Agrıcola, Zaragoza*. 1991, pp. 433–440.

[121] Kevin E Kane and Won Suk Lee. "Spectral sensing of different citrus varieties for precision agriculture". In: *2006 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2006, p. 1.

[122] Kevin Edward Kane and Won Suk Lee. "Multispectral imaging for in-field green citrus identification". In: *2007 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2007, p. 1.

[123] Manoj Karkee et al. "Identification of pruning branches in tall spindle apple trees for automated pruning". In: *Computers and Electronics in Agriculture* 103 (2014), pp. 127–135.

[124] Y Kim. "Apple yield mapping using a multispectral imaging sensor". In: *Intl. Scientific Conference on Agricultural Engineering (AgEng'04)*. Katolieke Universiteit Leuven. 2004.

[125] Denis Klimentjew, Norman Hendrich, and Jianwei Zhang. "Multi sensor fusion of camera and 3d laser range finder for object recognition". In: *2010 IEEE Conference on Multisensor Fusion and Integration*. IEEE. 2010, pp. 236–241.

[126] De-yuan Kong et al. "Research of apple harvesting robot based on least square support vector machine". In: *2010 International Conference on Electrical and Control Engineering*. IEEE. 2010, pp. 1590–1593.

[127] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[128] Shailesh Kumar, Joydeep Ghosh, and Melba M. Crawford. "Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis". In: *Pattern Analysis & Applications* 5.2 (2002), pp. 210–220.

[129] Ferhat Kurtulmus, Won Suk Lee, and Ali Vardar. "Green citrus detection using 'eigenfruit', color and circular Gabor texture features under natural outdoor conditions". In: *Computers and Electronics in Agriculture* 78.2 (2011), pp. 140–149.

[130] Ferhat Kurtulmus, Won Suk Lee, and Ali Vardar. "Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network". In: *Precision agriculture* 15.1 (2014), pp. 57–79.

[131] Nataliia Kussul et al. "Deep learning classification of land cover and crop types using remote sensing data". In: *IEEE Geoscience and Remote Sensing Letters* 14.5 (2017), pp. 778–782.

[132] Kentaro Kuwata and Ryosuke Shibasaki. "Estimating crop yields with deep learning and remotely sensed data". In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2015, pp. 858–861.

[133] Raphaël Labayrade et al. "Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner". In: *Autonomous Robots* 19.2 (2005), pp. 117–140.

[134] Hei Law and Jia Deng. "Cornernet: Detecting objects as paired keypoints". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 734–750.

[135] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 2169–2178.

[136] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[137] BongKi Lee et al. "A Vision Servo System for Automated Harvest of Sweet Pepper in Korean Greenhouse Environment". In: *Applied Sciences* 9.12 (2019), p. 2395.

[138] Sue Han Lee et al. "Deep-plant: Plant identification with convolutional neural networks". In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2015, pp. 452–456.

[139] Won Suk Lee. "Citrus Yield Mapping System in Natural Outdoor Scenes using the Watershed Transform". In: *Proceedings of the 2006 ASABE Annual International Meeting Sponsored, ASABE Oregon Convention Center, Portland, OR, USA*. 2006, pp. 9–12.

[140] PBC Leite et al. "Hidden Markov models applied in agricultural crops classification". In: *Proceeding of GEOBIA (GEOgraphic Object-Based Image Analysis for the 21St Century)* (2008).

[141] BIN LI, MAOHUA WANG, and NING WANG. "Development of a real-time fruit recognition system for pineapple harvesting robots". In: *2010 Pittsburgh, Pennsylvania, June 20-June 23, 2010*. American Society of Agricultural and Biological Engineers. 2010, p. 1.

[142] Han Li, Won Suk Lee, and Ku Wang. "Identifying blueberry fruit of different growth stages using natural outdoor color images". In: *Computers and electronics in agriculture* 106 (2014), pp. 91–101.

[143] Yali Li et al. "A survey of recent advances in visual feature detection". In: *Neurocomputing* 149 (2015), pp. 736–751.

[144] Zeming Li et al. "Detnet: A backbone network for object detection". In: *arXiv preprint arXiv:1804.06215* (2018).

[145] Zeming Li et al. "Light-head r-cnn: In defense of two-stage object detector". In: *arXiv preprint arXiv:1711.07264* (2017).

[146] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[147] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[148] Marvin Lindner, Andreas Kolb, and Klaus Hartmann. "Data-fusion of PMD-based distance-information and high-resolution RGB-images". In: *2007 International Symposium on Signals, Circuits and Systems*. Vol. 1. IEEE. 2007, pp. 1–4.

[149] Ola Lindroos et al. "Estimating the position of the harvester head–a key step towards the precision forestry of the future?" In: *Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering* 36.2 (2015), pp. 147–164.

[150] Peter P Ling et al. "Sensing and end-effector for a robotic tomato harvester". In: *2004 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2004, p. 1.

[151] Raphael Linker, Oded Cohen, and Amos Naor. "Determination of the number of green apples in RGB images recorded in orchards". In: *Computers and Electronics in Agriculture* 81 (2012), pp. 45–57.

[152] Li Liu et al. "Deep Learning for Generic Object Detection: A Survey". In: *International Journal of Computer Vision* (2019). ISSN: 1573-1405. DOI: 10.1007/s11263-019-01247-4. URL: https://doi.org/10.1007/s11263-019-01247-4.

[153] Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[154] Y Liu, B Chen, and J Qiao. "Development of a machine vision algorithm for recognition of peach fruit in a natural scene". In: *Transactions of the ASABE* 54.2 (2011), pp. 695–702.

[155] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[156] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[157] Heng Lu et al. "Cultivated land information extraction in UAV imagery based on deep convolutional neural network and transfer learning". In: *Journal of Mountain Science* 14.4 (2017), pp. 731–741.

[158] Francois PS Luus et al. "Multiview deep learning for land-use classification". In: *IEEE Geoscience and Remote Sensing Letters* 12.12 (2015), pp. 2448–2452.

[159] Diptesh Majumdar et al. "Application of fuzzy c-means clustering method to classify wheat leaf images based on the presence of rust disease". In: *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*. Springer. 2015, pp. 277–284.

[160] Santiago Manen, Matthieu Guillaumin, and Luc Van Gool. "Prime object proposals with randomized prim's algorithm". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 2536–2543.

[161] Chris McCool, Tristan Perez, and Ben Upcroft. "Mixtures of lightweight deep convolutional neural networks: Applied to agricultural robotics". In: *IEEE Robotics and Automation Letters* 2.3 (2017), pp. 1344–1351.

[162] SS Mehta and TF Burks. "Vision-based control of robotic manipulator for citrus harvesting". In: *Computers and Electronics in Agriculture* 102 (2014), pp. 146–158.

[163] Krystian Mikolajczyk et al. "A comparison of affine region detectors". In: *International journal of computer vision* 65.1-2 (2005), pp. 43–72.

[164] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. "Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 4 (2017), p. 41.

[165] Dinh Ho Tong Minh et al. "Deep Recurrent Neural Networks for mapping winter vegetation quality coverage via multi-temporal SAR Sentinel-1". In: *arXiv preprint arXiv:1708.03694* (2017).

[166] Sharada P Mohanty, David P Hughes, and Marcel Salathé. "Using deep learning for image-based plant disease detection". In: *Frontiers in plant science* 7 (2016), p. 1419.

[167] Dhiman Mondal and Dipak Kumar Kole. "A time efficient leaf rust disease detection technique of wheat leaf images using pearson correlation coefficient and rough fuzzy C-means". In: *Information Systems Design and Intelligent Applications*. Springer, 2016, pp. 609–618.

[168] Dimitrios Moshou et al. "Automatic detection of 'yellow rust'in wheat using reflectance measurements and neural networks". In: *Computers and electronics in agriculture* 44.3 (2004), pp. 173–188.

[169] Dimitrios Moshou et al. "Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organising maps". In: *Precision Agriculture* 7.3 (2006), pp. 149–164.

[170] Joseph L Mundy. "Object recognition in the geometric era: A retrospective". In: *Toward category-level object recognition*. Springer, 2006, pp. 3–28.

[171] Hiroshi Murase and Shree K Nayar. "Visual learning and recognition of 3-D objects from appearance". In: *International journal of computer vision* 14.1 (1995), pp. 5–24.

[172] Md Mursalin and Md Mesbah-Ul-Awal. "Towards classification of weeds through digital image". In: *2014 Fourth International Conference on Advanced Computing & Communication Technologies*. IEEE. 2014, pp. 1–4.

[173] AD Nakarmi and L Tang. "Automatic inter-plant spacing sensing at early growth stages using a 3D vision sensor". In: *Computers and electronics in agriculture* 82 (2012), pp. 23–31.

[174] Vishvjit S Nalwa. *A guided tour of computer vision*. Addison-Wesley Longman Publishing Co., Inc., 1994.

[175] Sarah Taghavi Namin et al. "Deep phenotyping: deep learning for temporal phenotype/genotype classification". In: *Plant methods* 14.1 (2018), p. 66.

[176] Yasasvy Nanyam et al. "A decision-fusion strategy for fruit quality inspection using hyperspectral imaging". In: *Biosystems engineering* 111.1 (2012), pp. 118–125.

[177] Alejandro Newell, Zhiao Huang, and Jia Deng. "Associative embedding: End-to-end learning for joint detection and grouping". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2277–2287.

[178] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[179] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 7 (2002), pp. 971–987.

[180] Hiroshi Okamoto and Won Suk Lee. "Green citrus detection using hyperspectral imaging". In: *Computers and electronics in agriculture* 66.2 (2009), pp. 201–208.

[181] Hiroshi Okamoto and Won Suk Lee. "Machine vision for green citrus detection in tree images". In: *Environmental Control in Biology* 48.2 (2010), pp. 93–99.

[182] Elham Omrani et al. "Potential of radial basis function-based support vector regression for apple disease detection". In: *Measurement* 55 (2014), pp. 512–519.

[183] Maxime Oquab et al. "Learning and transferring mid-level image representations using convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724.

[184] Ahmad Ostovar, Thomas Hellström, and Ola Ringdahl. "Human detection based on infrared images in forestry environments". In: *International Conference on Image Analysis and Recognition*. Springer. 2016, pp. 175–182.

[185] Ahmad Ostovar, Ola Ringdahl, and Thomas Hellström. "Adaptive Image Thresholding of Yellow Peppers for a Harvesting Robot". In: *Robotics* 7.1 (2018), p. 11.

[186] Ahmad Ostovar et al. "Detection and classification of Root and Butt-Rot (RBR) in Stumps of Norway Spruce Using RGB Images and Machine Learning". In: *Sensors* 19.7 (2019), p. 1579.

[187] Edgar Osuna, Robert Freund, and Federico Girosit. "Training support vector machines: an application to face detection". In: *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE. 1997, pp. 130–136.

[188] Wanli Ouyang et al. "Chained cascade network for object detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1938–1946.

[189] Wanli Ouyang et al. "DeepID-Net: Object detection with deformable part based convolutional neural networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.7 (2016), pp. 1320–1334.

[190] Xanthoula-Eirini Pantazi, Dimitrios Moshou, and Cedric Bravo. "Active learning system for weed species recognition based on hyperspectral sensing". In: *Biosystems Engineering* 146 (2016), pp. 193–202.

[191] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. "A general framework for object detection". In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE. 1998, pp. 555–562.

[192] Hetal N Patel, RK Jain, and Manjunath V Joshi. "Fruit detection using improved multiple features based algorithm". In: *International journal of computer applications* 13.2 (2011), pp. 1–5.

[193] Alison B Payne et al. "Estimation of mango crop yield using image analysis–segmentation method". In: *Computers and electronics in agriculture* 91 (2013), pp. 57–64.

[194] A Payne et al. "Estimating mango crop yield using image analysis using fruit at 'stone hardening'stage and night time imaging". In: *Computers and Electronics in Agriculture* 100 (2014), pp. 160–167.

[195] Jane H Pejsa and James E Orrock. "Intelligent robot systems: potential agricultural applications". In: (1984).

[196] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification". In: *European conference on computer vision*. Springer. 2010, pp. 143–156.

[197] Georg Petschnigg et al. "Digital photography with flash and no-flash image pairs". In: *ACM transactions on graphics (TOG)* 23.3 (2004), pp. 664–672.

[198] Alessio Plebe and Giorgio Grasso. "Localization of spherical fruits for robotic harvesting". In: *Machine Vision and Applications* 13.2 (2001), pp. 70–79.

[199] Jean Ponce et al. *Toward category-level object recognition*. Vol. 4170. Springer, 2007.

[200] Moacir Ponti et al. "Precision agriculture: Using low-cost systems to acquire low-altitude images". In: *IEEE computer graphics and applications* 36.4 (2016), pp. 14–20.

[201] Mostafa Pordel, Thomas Hellström, and Ahmad Ostovar. "Integrating Kinect Depth Data with a Stochastic Object Classification Framework for Forestry Robots." In: *ICINCO (2)*. 2012, pp. 314–320.

[202] Ciro Potena, Daniele Nardi, and Alberto Pretto. "Fast and accurate crop and weed identification with summarized train sets for precision agriculture". In: *International Conference on Intelligent Autonomous Systems*. Springer. 2016, pp. 105–121.

[203] Michael P Pound et al. "Deep machine learning provides state-of-the-art performance in image-based plant phenotyping". In: *Gigascience* 6.10 (2017), gix083.

[204] Lü Qiang et al. "Identification of fruit and branch in natural scenes for citrus harvesting robot using machine vision and support vector machine". In: *International Journal of Agricultural and Biological Engineering* 7.2 (2014), pp. 115–121.

[205] Md Khurram Monir Rabby, Brinta Chowdhury, and Jung H Kim. "A Modified Canny Edge Detection Algorithm for Fruit Detection & Classification". In: *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE. 2018, pp. 237–240.

[206] Maryam Rahnemoonfar and Clay Sheppard. "Deep count: fruit counting based on deep simulated learning". In: *Sensors* 17.4 (2017), p. 905.

[207] Esa Rahtu, Juho Kannala, and Matthew Blaschko. "Learning a category independent object detection cascade". In: *2011 international conference on Computer Vision*. IEEE. 2011, pp. 1052–1059.

[208] Jurij Rakun, Denis Stajnko, and Damjan Zazula. "Detecting fruits in natural scenes by using spatial-frequency based texture analysis and multiview geometry". In: *Computers and Electronics in Agriculture* 76.1 (2011), pp. 80–88.

[209] Julien Rebetez et al. "Augmenting a convolutional neural network with local histograms-A case study in crop classification from high-resolution UAV imagery." In: *ESANN*. 2016.

[210] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[211] Joseph Redmon and Ali Farhadi. "Yolov3: An incremental improvement". In: *arXiv preprint arXiv:1804.02767* (2018).

[212] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[213] Murali Regunathan and Won Suk Lee. "Citrus fruit identification and size determination using machine vision and ultrasonic sensors". In: *2005 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2005, p. 1.

[214] J Reid and S Searcy. "Vision-based guidance of an agriculture tractor". In: *IEEE Control Systems Magazine* 7.2 (1987), pp. 39–43.

[215] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[216] Jason D Rennie et al. "Tackling the poor assumptions of naive bayes text classifiers". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 616–623.

[217] Ola Ringdahl, Polina Kurtser, and Yael Edan. "Strategies for selecting best approach direction for a sweet-pepper harvesting robot". In: *Annual Conference Towards Autonomous Robotic Systems*. Springer. 2017, pp. 516–525.

[218] Ola Ringdahl et al. “Enhanced Algorithms for Estimating Tree Trunk Diameter Using 2D Laser Scanner”. In: *Remote Sensing* 5.10 (2013), pp. 4839–4856. DOI: 10.3390/rs5104839.

[219] Mohamed Rizon et al. “Object detection using circular Hough transform”. In: (2005).

[220] Anderson Rocha et al. “Automatic fruit and vegetable classification from images”. In: *Computers and Electronics in Agriculture* 70.1 (2010), pp. 96–104.

[221] Juan Romeo et al. “A new Expert System for greenness identification in agricultural images”. In: *Expert Systems with Applications* 40.6 (2013), pp. 2275–2286.

[222] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. “Neural network-based face detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* 20.1 (1998), pp. 23–38.

[223] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[224] Bryan C Russell et al. “LabelMe: a database and web-based tool for image annotation”. In: *International journal of computer vision* 77.1-3 (2008), pp. 157–173.

[225] Dymitr Ruta and Bogdan Gabrys. “An overview of classifier fusion methods”. In: *Computing and Information systems* 7.1 (2000), pp. 1–10.

[226] Inkyu Sa et al. “Deepfruits: A fruit detection system using deep neural networks”. In: *Sensors* 16.8 (2016), p. 1222.

[227] Omri Safren et al. “Detection of green apples in hyperspectral images of apple-tree foliage using machine vision”. In: *Transactions of the ASABE* 50.6 (2007), pp. 2303–2313.

[228] Mayanda Mega Santoni et al. “Cattle race classification using gray level co-occurrence matrix convolutional neural networks”. In: *Procedia Computer Science* 59 (2015), pp. 493–502.

[229] Alistair J Scarfe et al. “Development of an autonomous kiwifruit picking robot”. In: *2009 4th International Conference on Autonomous Robots and Agents*. IEEE. 2009, pp. 380–384.

[230] Cordelia Schmid and Roger Mohr. “Local grayvalue invariants for image retrieval”. In: *IEEE transactions on pattern analysis and machine intelligence* 19.5 (1997), pp. 530–535.

[231] Woo Chaw Seng and Seyed Hadi Mirisaee. “A new method for fruits recognition system”. In: *2009 International Conference on Electrical Engineering and Informatics*. Vol. 1. IEEE. 2009, pp. 130–134.

[232] Subhajit Sengupta and Won Suk Lee. "Identification and determination of the number of green citrus fruit under different ambient light conditions". In: *International Conference of Agricultural Engineering CIGR-AgEng2012*. 2012.

[233] Pierre Sermanet et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks". In: *arXiv preprint arXiv:1312.6229* (2013).

[234] LG Shapiro and GC Stockman. *Computer Vision Prentice-Hall Upper Saddle River*. 2001.

[235] Linda Shapiro. *Computer vision and image processing*. Academic Press, 1992.

[236] Yong Shi and Marcus Judd. "Finding nearest neighbors for multi-dimensional data". In: *DBKDA 2013* (2013), p. 60.

[237] Abhisesh Silwal, Aleana Gongal, and Manoj Karkee. "Apple identification in field environment with over the row machine vision system". In: *Agricultural Engineering International: CIGR Journal* 16.4 (2014), pp. 66–75.

[238] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[239] F Sistler. "Robotics and intelligent machines in agriculture". In: *IEEE Journal on Robotics and Automation* 3.1 (1987), pp. 3–6.

[240] Peter W Sites and Michael J Delwiche. "Computer vision to locate fruit on a tree". In: *Transactions of the ASAE* 31.1 (1988), pp. 257–0265.

[241] Josef Sivic and Andrew Zisserman. "Video Google: A text retrieval approach to object matching in videos". In: *null*. IEEE. 2003, p. 1470.

[242] Xiaodong Song et al. "Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model". In: *Journal of Arid Land* 8.5 (2016), pp. 734–748.

[243] René A Sørensen et al. "Thistle detection using convolutional neural networks". In: *2017 EFITA WCCA CONGRESS*. 2017, p. 161.

[244] Denis Stajnko, Miran Lakota, and Marko Hočevar. "Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging". In: *Computers and Electronics in Agriculture* 42.1 (2004), pp. 31–42.

[245] Denis Stajnko, Jurij Rakun, Michael Blanke, et al. "Modelling apple fruit yield using image analysis for fruit colour, shape and texture". In: *European journal of horticultural science* 74.6 (2009), p. 260.

[246] Kim Steen et al. "Using deep learning to challenge safety standard for highly autonomous machines in agriculture". In: *Journal of Imaging* 2.1 (2016), p. 6.

[247] Georgina Stegmayer et al. “Automatic recognition of quarantine citrus diseases”. In: *Expert Systems with Applications* 40.9 (2013), pp. 3512–3517.

[248] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[249] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. “Deep neural networks for object detection”. In: *Advances in neural information processing systems*. 2013, pp. 2553–2561.

[250] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[251] Christian Szegedy et al. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[252] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[253] Christian Szegedy et al. “Scalable, high-quality object detection”. In: *arXiv preprint arXiv:1412.1441* (2014).

[254] Amy L Tabb, Donald L Peterson, and Johnny Park. “Segmentation of apple fruit from video via background modeling”. In: *2006 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2006, p. 1.

[255] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. “Textural features corresponding to visual perception”. In: *IEEE Transactions on Systems, man, and cybernetics* 8.6 (1978), pp. 460–473.

[256] Kanae Tanigaki et al. “Cherry-harvesting robot”. In: *Computers and electronics in agriculture* 63.1 (2008), pp. 65–72.

[257] Alberto Tellaeche et al. “A vision-based classifier in precision agriculture combining Bayes and Support Vector Machines”. In: *2007 IEEE International Symposium on Intelligent Signal Processing*. IEEE. 2007, pp. 1–6.

[258] Alberto Tellaeche et al. “A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms”. In: *Innovations in Hybrid Intelligent Systems*. Springer, 2007, pp. 72–79.

[259] John Triantafilis and SM Lesch. “Mapping clay content variation using electromagnetic induction techniques”. In: *Computers and Electronics in Agriculture* 46.1-3 (2005), pp. 203–237.

[260] Tinne Tuytelaars, Krystian Mikolajczyk, et al. "Local invariant feature detectors: a survey". In: *Foundations and trends® in computer graphics and vision* 3.3 (2008), pp. 177–280.

[261] Jasper RR Uijlings et al. "Selective search for object recognition". In: *International journal of computer vision* 104.2 (2013), pp. 154–171.

[262] James P. Underwood et al. "Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors". In: *Computers and Electronics in Agriculture* 130 (2016), pp. 83–96. ISSN: 0168-1699.

[263] Régis Vaillant, Christophe Monrocq, and Yann Le Cun. "Original approach for the localisation of objects in images". In: *IEE Proceedings-Vision, Image and Signal Processing* 141.4 (1994), pp. 245–250.

[264] Koen EA Van de Sande et al. "Segmentation as selective search for object recognition." In: *ICCV*. Vol. 1. 2. 2011, p. 7.

[265] Eldert J Van Henten et al. "An autonomous robot for harvesting cucumbers in greenhouses". In: *Autonomous Robots* 13.3 (2002), pp. 241–258.

[266] Andrea Vedaldi et al. "Multiple kernels for object detection". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 606–613.

[267] Balasubramanian VijayaLakshmi and Vasudev Mohan. "Kernel-based PSO and FRVM: An automatic plant leaf type detection using texture, shape, and color features". In: *Computers and Electronics in Agriculture* 125 (2016), pp. 99–112.

[268] Paul Viola, Michael Jones, et al. "Rapid object detection using a boosted cascade of simple features". In: *CVPR (1)* 1.511-518 (2001), p. 3.

[269] Efi Vitzrabin and Yael Edan. "Adaptive thresholding with fusion using a RGBD sensor for red sweet-pepper detection". In: *Biosystems Engineering* 146 (2016), pp. 45–56.

[270] E Vrindts et al. "Management zones based on correlation between soil compaction, yield and crop data". In: *Biosystems Engineering* 92.4 (2005), pp. 419–428.

[271] Juan P Wachs et al. "Low and high-level visual feature-based apple detection from multi-modal images". In: *Precision Agriculture* 11.6 (2010), pp. 717–735.

[272] J Wachs et al. "Apple detection in natural tree canopies from multimodal images". In: *Proceedings of the 7th European Conference on Precision Agriculture, Wageningen, The Netherlands*. Vol. 68. 2009, pp. 293–302.

[273] Mirwaes Wahabzada et al. "Automated interpretation of 3D laserscanned point clouds for plant organ segmentation". In: *BMC bioinformatics* 16.1 (2015), p. 248.

[274] Jianlun Wang et al. “An adaptive thresholding algorithm of field leaf image”. In: *Computers and electronics in agriculture* 96 (2013), pp. 23–39.

[275] Jin-jing Wang et al. “Application of support vector machine to apple recognition using in apple harvesting robot”. In: *2009 International Conference on Information and Automation*. IEEE. 2009, pp. 1110–1115.

[276] Qi Wang et al. “Automated crop yield estimation for apple orchards”. In: *Experimental robotics*. Springer. 2013, pp. 745–758.

[277] Dale Whittaker et al. “Fruit location in a partially occluded image”. In: *Transactions of the ASAE* 30.3 (1987), pp. 591–0596.

[278] RR Wolfe and M Swaminathan. *Determining orientation and shape of bell peppers by machine vision*. ASAE, 1986.

[279] Gunter Wyszecki and Walter Stanley Stiles. *Color science*. Vol. 8. Wiley New York, 1982.

[280] Rong Xiao, Long Zhu, and Hongjiang Zhang. “Boosting Chain Learning for Object Detection.” In: *ICCV*. Vol. 3. 2003, p. 709.

[281] Saining Xie et al. “Aggregated residual transformations for deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.

[282] Wang Xinshao and Cai Cheng. “Weed seeds classification based on PCANet deep learning baseline”. In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 408–415.

[283] Hulya Yalcin. “Plant phenology recognition using deep learning: DeepPheno”. In: *2017 6th International Conference on Agro-Geoinformatics*. IEEE. 2017, pp. 1–5.

[284] Ting Yuan et al. “Spectral imaging for greenhouse cucumber fruit detection based on binocular stereovision”. In: *2010 Pittsburgh, Pennsylvania, June 20-June 23, 2010*. American Society of Agricultural and Biological Engineers. 2010, p. 1.

[285] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[286] Fangming Zhang and Naiqian Zhang. “Applying joint transform correlator in tomato recognition”. In: *ASABE Annual International Meeting*. 2008.

[287] Jainguo Zhang, Tieniu Tan, and Li Ma. “Invariant texture segmentation via circular Gabor filters”. In: *Object recognition supported by user interaction for service robots*. Vol. 2. IEEE. 2002, pp. 901–904.

[288] Xin Zhang et al. “Object class detection: A survey”. In: *ACM Computing Surveys (CSUR)* 46.1 (2013), p. 10.

[289] Jun Zhao, Joel Tow, and Jayantha Katupitiya. "On-tree fruit recognition using texture properties and color data". In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2005, pp. 263–268.

[290] LI Zhen and Tian-sheng HONG. "Automatic detection of growing orange fruits by machine vision". In: *2007 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers. 2007, p. 1.

[291] Nanning Zheng and Jianru Xue. *Statistical learning and pattern analysis for image and video processing*. Springer Science & Business Media, 2009.

[292] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[293] Bolei Zhou et al. "Object detectors emerge in deep scene cnns". In: *arXiv preprint arXiv:1412.6856* (2014).

[294] Min Zhou et al. "Multi-resolution networks for ship detection in infrared remote sensing images". In: *Infrared Physics & Technology* 92 (2018), pp. 183–189.

[295] Rong Zhou et al. "Using colour features of cv.'Gala'apple fruits in an orchard in image processing to predict yield". In: *Precision Agriculture* 13.5 (2012), pp. 568–580.

[296] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC, 2012.

[297] C Lawrence Zitnick and Piotr Dollár. "Edge boxes: Locating object proposals from edges". In: *European conference on computer vision*. Springer. 2014, pp. 391–405.