Chapter 1

Introduction

As far back as the 16th century Leonardo da Vinci was intrigued with how the outside world could be projected onto a wall of a dark room via a tiny opposite hole. His description of camera obscura (the pin-hole camera) was among the earliest works in the field which is now called optics. Further advancements in optics and camera technology allowed the scientific study of visual data.

In the 1950s Hubel and Wiesel (1959) performed experiments on the visual cortex of the brain and found how the early layers of mammal brains process visual data. They discovered that specific neurons responded to specific orientations of projected lines. This discovery has been fundamental to our understanding of how visual pathways extract increasingly complex information from the external stimuli presented to the retina. In the book, concisely named, “Vision” by Marr (1982), a more formal, somewhat idealized, visual image decomposition framework was described. Marr argued that an input image is first represented by lines, edges and curves, somewhat related to a pencil sketch. Subsequently, higher abstractions like textured surfaces and discontinuous segments are formed, analogue to shading. Finally the concept of an object as a whole is formed in 3D. This general hierarchical image decomposition logic is still at the basis of most modern-day computer vision approaches. With the advent of modern-day computers many traditional computer vision algorithms were developed to extract all kinds of information from images. Most notable are, edge detectors Lowe (1987), image segmentation algorithms (Shi and Malik, 2000), corner detectors (Lowe, 1999) and face detectors (Viola and Jones, 2001).

When sight evolved in early organisms it gave them a significant
advantage. From that time on, visual data could be processed by the brain to interpret the world in ways never seen before. Even in modern humans a large part of the brain is dedicated to processing visual stimuli. It is therefore not surprising that the fields computer vision and machine learning co-evolved.

In 1958 the American psychologist Frank Rosenblatt introduced the Perceptron (Rosenblatt, 1958), an electronic device that was developed in accordance with biological principles and which showed initial learning abilities. The Rosenblatt perceptron contained a 400 pixel camera that was able to detect certain geometric objects presented to it. During this era the New York Times (1958) picked up on the great potential of this machine that arguably “learns by doing”. This invention sparked great interest by both the general public and the scientific community. The idea of perceptrons and neural networks was however largely abandoned after Marvin Minsky and Seymour Papert published Perceptrons (Minsky and Papert, 1969). They showed that these perceptrons are only able to learn simple linear functions and were unable to learn a simple binary XOR function. This triggered the “Thought of Disillusionment” of the Gartner hype cycle, or like it was more famously coined: The first A.I. winter. After this, the connectionist research rose again with Hopfield networks (Hopfield, 1982). These are content-addressable recurrent neural networks with binary thresholds that converge to a local minimum. The Lighthill report (Lighthill, 1973) painted a rather pessimistic prognosis for the core technology of A.I. and was the main trigger for another A.I. winter. Although it is debatable on how many A.I. winters there have exactly been, from the early 1990s onward, neural network technology was mainly integrated into larger (vision) systems and the rise to widespread use was more gradual.

Fast forward to 2015, the researchers Yann LeCun, Yoshua Bengio and Geoffrey Hinton published an article about deep learning in Nature (LeCun et al., 2015). In this article they discussed convolutional neural networks, the latest innovation in neural networks. Deep learning has caused an unprecedented improvement on the state-of-the-art in a wide variety of applications and at this moment still keeps improving with an impressive pace (Statt, 2018). Again this has attracted major interest from scientists, companies and the general public alike. Deep learning is for a large part responsible for the innovations in the areas of self-driving cars
Chapter 1. Introduction

(Grigorescu et al., 2020), speech-controlled smart assistants (Deng et al., 2013) and medical image analysis (Ronneberger et al., 2015).

Why is deep learning so successful? In the traditional computer vision and machine learning tandem, the vision algorithms extract image features and the machine learning algorithms use these low-level features to learn high-level concepts like persons (Suleiman and Sze, 2014) and cars (Bougharriou et al., 2017). A successful example of this tandem is the Histogram of Oriented Gradients (HoG) (Dalal and Triggs, 2005) coupled with a Support Vector Machine (SVM) (Cortes and Vapnik, 1995). HoG calculates the histograms of local edge information and an SVM finds the grouping of features which determines the class that the image belongs to. For decades these types of combined approaches have solved many computer vision problems to a satisfactory degree. This has resulted in many practical applications for quality inspection, 3D reconstruction and geometric pattern matching. However, in deceptively difficult cases where there is much variation in illumination, appearance, shape and perspective the use of traditional computer vision approaches has been a mixed success. One of the reasons for this is that traditional feature extraction algorithms are manually designed, which has proven to be quite difficult for complex scenes (O'Mahony et al., 2019).

With deep learning or more specifically Convolutional Neural Networks (CNNs) (Schmidhuber, 2015; LeCun et al., 2015), end-to-end trainable machine learning models are possible. Where in traditional computer vision, feature extractors were handcrafted, in deep learning these feature extractors are trained using a large set of training data. Deep learning has been shown to perform well on many data modalities but, this dissertation will focus mainly on image data. Several feature extraction layers, like convolutional layers, learn to represent an image as increasingly abstract intermediate representations. The final full-connected layer is usually trained to recognize high-level concepts like the object class. This can already be regarded as the classic deep learning architecture (LeCun et al., 1998). More modern architectures use any combination of deep learning concepts to solve a multitude of tasks (image classification, object detection, object counting, image generation, etc.) The taxonomy of deep learning architectures still grows rapidly and in this dissertation also some new architectures for solving specific tasks are introduced.
Lee et al. (2009) and Krizhevsky et al. (2012) analyzed what CNNs actually learn when trained with a very large set of images. The authors showed that the initial layer learned to detect specifically oriented edges, corners and flat surfaces. Subsequent, deeper, layers of the network learned object hierarchies like wheels and cars or eyes and faces. This showed that CNNs were able to learn feature detectors that responded similarly to stimuli as those found in cat brains by Hubel and Wiesel around 50 years earlier (Hubel and Wiesel, 1959). Deep learning models have also been able to learn to decompose an image in a way similar to how it was formulated by David Marr in 1970 (Marr, 1982).

One of the earliest architectures that can be called a deep convolutional neural network is the Neocognitron (Fukushima, 1980) although more recently, the work of LeCun et al. (1998) is seen as modern parent-work of CNN architectures for image based pattern recognition. The reason for their recent success has mainly to do with the availability of high performance hardware in the form of Graphical Processing Units (GPUs). These highly parallel processing units were originally intended for gaming and other graphical applications, but it was found that they are also very suitable for deep learning (Luebke et al., 2006).

1.1 Research questions

This section introduces the central theme of this dissertation, the trinity of the technologies: deep learning, hyperspectral imaging and Unmanned Aerial Vehicles (UAVs). It gives a background on the reasoning behind this theme and its relation to this dissertation and the research questions.

Typical Red Green Blue (RGB) cameras are mainly inspired by the colors perceived by human vision. The retina has two types of photo-sensitive structures: cones and bars. In human eyes there are three types of cones and each type is sensitive to a specific part of the visible spectrum, namely: red, green and blue. Digital cameras record images in these three channels and from combinations of these three primary colors every human perceivable color can be created. This is ideal for creating realistic images of the world as humans perceive it.

However, the physical world itself does not consist of only these three primary colors but of a much richer frequency spectrum of photons called the electromagnetic spectrum. Many of these photons with specific
wavelengths are not visible to the human eye. Certain wavelengths like near infrared and ultraviolet are partly visible to dogs, cats and snakes for example. But many interesting wavelengths, like thermal infrared, short-wave infrared and x-ray, are not directly perceivable. Multispectral cameras can record more than the three primary colors and augment the image with additional image channels like near infrared and red edge for vegetation inspection (Berra et al., 2017). Hyperspectral cameras aim to perceive consecutive narrow bands of wavelengths of the electromagnetic spectrum. This allows for exact color measurement and chemical fingerprinting. The difference between multispectral and hyperspectral cameras is not clearly defined and, in this work, all cameras which produce images with more than the traditional three channels are referred to as hyperspectral cameras. In analogy with that, the images taken by these cameras are referred to as hyperspectral cubes.

The platform from which a camera takes images can be anything from hand-held to a tripod to a car. Due to recent advancements in aerial and battery technology, UAVs can bring cameras to places where images can be taken from unique aerial perspectives. This notion has received great interest from fields like precision agriculture and large structure inspection (Radoglou-Grammatikis et al., 2020; Hallermann and Morgenthal, 2013). Battery-powered small aerial platforms with three to eight rotors called multicopters are often used for these applications. Unfortunately these platforms have limited carrying capacity and can only stay airborne for up to 20 minutes, depending on the payload weight and size of the platform. Heavy multicopter platforms contain much kinetic energy during operation and can therefore become dangerous. For this reason the legislation for utilization of UAVs is strict and the usage of these systems is regulated 1. The development bottleneck of sophisticated computer vision applications with UAVs is mainly limited by these inherent limitations of the multicopter platforms. For example, high-quality images enable a more detailed analysis of the aerial images but these camera are usually too heavy to be fitted on small UAVs that can

---

1Regeling van de Staatssecretaris van Infrastructuur en Milieu, van 23 april 2015, IENM/BSK-2015/11533, houdende de vaststelling van regels voor op afstand bestuurde luchtvaartuigen.
be operated with a, in Dutch so called, ROC-light license \(^2\) (this type of UAV is used for the experiments of Chapter 4 and Chapter 5.)

Hyperspectral imaging in conjunction with UAVs is particularly helpful in precision agriculture. At the core of hyperspectral imaging devices there are regular 2D sensors. The limitations of hyperspectral imaging becomes apparent when knowing the methods by which these cameras actually create a 3D hyperspectral cube. Generally, either the spatial or temporal resolution is reduced to capture the extra spectral dimension (Chapter 3). Hyperspectral cameras are therefore mostly heavy due to extra optics, have low-resolution or they are sensitive to movement.

The interesting paradox here is that hyperspectral cameras seem ideal for the application (precision agriculture) but are not always suitable for the platform (UAV). The instability of aerial platforms conflicts with the sensitivity to movement of push-broom hyperspectral cameras, the low-resolution of light-weight hyperspectral cameras conflicts with the need to detect small image features and the heavy optics of some hyperspectral cameras conflict with the payload constraints of multicopters. Can state-of-the-art algorithms from the field of deep learning mitigate this, seemingly, incompatibility between hyperspectral imaging and UAVs? This leads to the following central research question that is addressed in this dissertation:

“\textit{How can deep learning be utilized to mitigate the limitations imposed by small aerial platforms employing hyperspectral imaging technology?}”

Answers to this broad question could be searched for in several directions. One could optimize the physical design of the devices with deep learning or one could use this technology to optimize the movement of the aerial platform so it can navigate more effectively and efficiently. This dissertation is not about optimizing physical aspects of the technologies, but instead, seeks to apply machine learning methods to perform image processing in such a way that the limitations are mitigated.

This trinity of the technologies deep learning, hyperspectral imaging and UAVs can be used as a framework to reason about applications in a

\(^2\)Beleidsregel van de Staatssecretaris van Infrastructuur en Milieu, van 30 mei 2016, nr. ILT-2016/16814, voor het verlenen van ontheffingen voor micro- en minidrones.
1.2. Dissertation overview

coherent way. Difficult challenges lie in the applications that require a combination of these three technologies. The first step is to identify applications within this trinity and the second step is to do research on several of these applications to gain scientific insights that can be used to make general conclusions. Solving all aspects of this trinity of technologies is not the goal, but instead it will be used as a vehicle to gain insights into the fields of deep learning and hyperspectral imaging from a perspective where the limits are imposed by the physical platform.

For some applications the full hyperspectral cube might not be needed and a problem might be solvable with less spectral bands. In that case cameras with less limitations could be used. From here follow the questions: “How can machine learning be used to automatically select important spectral bands?” and “What is the subsequent effect of using less bands on the performance of the posed problem?” These questions are addressed in Chapter 2.

The class of light-weight hyperspectral cameras that trade spatial resolution for spectral resolution are less sensitive to movement compared to alternative hyperspectral technologies. Therefore, they are more suitable to use with a UAV. This leads to the following question: “How can the quality and resolution of hyperspectral images be improved using deep learning?” This question is addressed in Chapter 3.

On most of the small commodity UAVs, low-cost and light-weight cameras are mounted. The images from these cameras often produce highly-compressed video with low image quality. The optics are often of low quality and provide a wide angle of view that causes all kinds of image distortions. This leads to the question: “How can deep learning be used to solve a challenging image-processing task using images produced by low-cost commodity UAVs?” This question is addressed in Chapter 4 and partly in Chapter 5.

1.2 Dissertation overview

Several narratives run though this dissertation. The first narrative is from an application point of view and an overview is discussed in the first part of this section. The remaining part gives an overview of the technical narrative to deep learning and designing CNNs for specific applications. The final narrative is in the arc which starts with the trinity of
technologies and its application framework, introduced in this chapter, and ends with the general technical and scientific contributions of this work in the conclusion.

The Netherlands has a leading role in the cultivation and export of seed potatoes, mainly due to their high quality. Monitoring quality and optimizing crop yield are therefore important. This agricultural output should be maximized while the ecological impact should be minimized and the state of crops needs to be monitored constantly and timely interventions for optimizing crop growth should be applied. Diseases are prevented or counteracted using pesticides. Many types of leaf damages can occur, varying from environmental effects (Wilkinson et al., 2012) to fungal infections (Turkensteen et al., 2010). When a disease is incorrectly diagnosed, a wrong treatment could be chosen. For example, pesticides are applied while there is no fungal infection. This could have a negative impact on the environment or even promote resistance to pesticides.

Two commonly-confused damages on potato plants are caused by either an Alternaria fungal infection or by exposure to ozone ($O_3$) (Turkensteen et al., 2010). Both produce similar brownish lesions on the leaf (Figure 1.1). Alternaria should be treated with pesticides while ozone damage should not. This makes a correct diagnosis important.

The properties of UAV-based camera systems are potentially ideal for the inspection of agricultural fields. UAVs are non invasive (they do not interact directly with the vegetation), they provide an aerial perspective at various heights and resolutions, they can use Global Positioning System (GPS) way points, visual features (Wang et al., 2016) or sensor fusion (de Boer et al., 2015) to navigate automatically and the inspection of agricultural fields using (hyperspectral) cameras and UAVs is gaining popularity (Paredes et al., 2017; van de Loosdrecht et al., 2014). Examples using regular RGB images are crop recognition (Rebetez et al., 2016) and disease detection (Mohanty et al., 2016). By expanding these measurements to near-infrared and the red-edge spectral ranges, Chlorophyll can be estimated to quantify overall vegetation health (Rasmussen et al., 2016; Berra et al., 2017; Simic Milas et al., 2018) using UAVs, which is used by farmers around the world (Mazur, 2016).

The cameras that are used for these applications record images in up to five color channels specifically chosen for Chlorophyll estimation (using five separate cameras). These sensors are not particularly suitable
for detecting subtle color differences in potato-leaf lesions because of their broad spectral sensitivity and limited spectral resolution. By further increasing the spectral resolution, diseases (Dijkstra et al., 2017) and soil concentrations can be determined (Pullanagari et al., 2016). These are typical application in which UAVs and hyperspectral imaging are combined. In Chapter 2 hyperspectral imaging and machine learning are combined to select appropriate spectral bands to distinguish between two important potato diseases. Hyperspectral images of leaves that are mapped to RGB are shown in Figure 1.1. Several band selection and machine learning methods are tested with a 28 band hyperspectral camera system. This gives insight in how machine learning can be used to reduce the amount of spectral bands while trying to maintain a high classification performance.

![Figure 1.1: Alternaria damage (left three images) and ozone damage (right three images)](image)

One of the downsides of UAVs is their limited payload capacity and the associated flight time. When aiming to utilize UAVs for precision agriculture efficient hyperspectral imaging devices are required. Light-weight cameras with a Multispectral Color Filter Array (MCFA) can produce a hyperspectral cube with up to 25 channels in one snapshot. These cameras contain a small filter mosaic that is repeated multiple times. For example, a $4 \times 4$ mosaic gives 16 spectral bands for each superpixel. An example of an image produces by this type of sensor is shown in Figure 1.2. These types of sensors are ideal for the temporal stability needed for UAVs, however, spatial resolution is greatly reduced and these sensors exhibit crosstalk which limits its direct applicability. In Chapter 3 a CNN specifically designed for upscaling hyperspectral images produced by this type of sensor is discussed. The research described in that chapter is an example of how the trinity of technologies introduced in this dissertation can be combined to benefit each other.
Chapter 1. Introduction

In most visual inspection tasks where UAVs are used, human experts manually inspect the images and usually no automatic inspection is performed. In these types of applications there lies great potential for deep learning in automating the inspection. One important indicator for predicting crop yield if the number of plants and for a better indication of crop-yield the plants should be localized and counted during the growth season. UAVs and deep learning are combined in Chapter 4 to propose a solution for the challenging task of counting potato crops from aerial images using commodity hardware and our novel CentroidNet algorithm. An example of an image of potato crops as viewed from the the UAV is shown in Figure 1.3.

Counting many small and connected objects is generally a challenging image analysis task (Cohen et al., 2017; Baygin et al., 2018) and solutions are mostly based on either traditional machine vision (Khan et al., 2018) or based on deep learning (Ferrari et al., 2015). Apart from counting potato crops, many other applications for counting objects exist. In Chapter 5 additional experiments for potato crop counting, bacterial colony counting
in water samples and cell nuclei counting are discussed. In that chapter the second version of CentroidNet is also introduced.

CentroidNet is a hybrid CNN specifically designed for counting. Designing methods that are hybrids between deep learning and computer vision are interesting because prior knowledge about the problem can be used to design specific solutions that are still trainable. This is more elaborately discussed in Chapter 6.

The technical narrative of this dissertation is as follows. Each chapter contains the technical and scientific details of this work. In Chapter 2 algorithms from traditional machine learning with paradigms somewhat borrowed from deep learning (e.g. using the Rectified Linear Unit (ReLU) (Krizhevsky et al., 2012)) are shown. A more detailed description of the deep learning methodology is described in Chapter 3. In that chapter the design of a minimal CNN is introduced for solving a specific problem namely hyperspectral demosaicking and crosstalk correction. In Chapter 4 a more general and existing CNN from literature is used as basis for the novel deep learning algorithm called CentroidNet. This algorithm uses a CNN and traditional computer vision to accurately detect centroids of objects. The more general principles underlying the different CNN architectures are discussed as deep design patterns in Chapter 5. In that chapter CentroidNetV2, a more comprehensive hybrid deep learning algorithm for detecting centroids as well as delineations of objects is discussed in detail.

In Chapter 6 a summary of the most important results is given and the research questions are answered. Subsequently the technical and scientific insights that were gained during this research are discussed. In the final part, future-research directions are proposed.