Deep learning and hyperspectral imaging for unmanned aerial vehicles

Dijkstra, Klaas

DOI:
10.33612/diss.131754011

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 6

Discussion and conclusion

In the first part of this chapter the answers to the research questions are discussed. In the second part the usefulness of combining deep learning with traditional computer vision paradigms is discussed. The final part will discuss the envisioned path for the future based on the conclusions from this dissertation.

The main research question addressed in this dissertation is: “How can deep learning be utilized to mitigate the limitations imposed by small aerial platforms employing hyperspectral imaging technology?” It should be clear that it is virtually impossible to overcome all limitations imposed by small aerial platforms when using hyperspectral imaging devices. Much of the limitations are a result of the underlying physics or are caused by theoretical limits. However, in this dissertation, solutions to mitigate several of the limitations have been demonstrated and discussed.

6.1 Research questions

A methodology for selecting the optimal spectral bands from a hyperspectral cube has been proposed for distinguishing two potato diseases on leaves in Chapter 2. A per-pixel classification of potato leaves was performed using several classifiers, and it was demonstrated, using feature selection and extraction, that a subset of three bands still provides useful results. This showed how the laboratory set up of a 28-band imaging system could be reduced to a three-band system while still retaining sufficient accuracy. The Liquid Crystal Tunable Filter (LCFT) used for collecting the hyperspectral images is difficult to use on a Unmanned Aerial Vehicle (UAV) because of its temporal instability (each
hyperspectral plane is taken at a different time). However, when reduced
to a three-band camera system, it is suitable for usage on a UAV. For
example, by using three separate cameras.

Prior to this research, distinguishing *Alteraria* and ozone damage was
difficult to perform because of visual similarities between the diseases (Turkensteen et al., 2010). This research contributes by providing a
machine-learning-based methodology for distinguishing *Alternaria* and
ozone damage on potato plant leaves in laboratory conditions using
hyperspectral imaging. Other studies used commodity multi-spectral
cameras or regular RGB cameras for crop-health monitoring on a UAVs
(Mohanty et al., 2016; Rebetez et al., 2016). These cameras are limited with
respect to the spectral resolution and the spectral bands that are captured
are intended for measuring Chlorophyll of vegetation (Berra et al., 2017).
This research contributes by demonstrating various methods for selecting
important subsets of spectral bands from a hyperspectral image cube.
This gives information on which bands need to be captured for detecting
specific diseases like *Alternaria* and ozone. This methodology provides an
answer to 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?”. 

The usage of a 16-band light-weight hyperspectral camera system
with a Multispectral Color Filter Array (MCFA) (Geelen et al., 2015),
mounted on a UAV, has been proposed in this dissertation. The
hyperspectral images produced by this system suffer from crosstalk and
their spatial resolution is reduced sixteen fold to accommodate for the
increase in spectral resolution (Keren and Osadchy, 1999). Several
methods for demosaicking images have been proposed in literature: using
edge information, neural networks, inpainting and linear models (Monno
et al., 2012; Wang, 2014; Wang et al., 2017; Aggarwal and Majumdar,
2014). This research contributes by proposing a deep-learning-based
method for demosaicking a $4 \times 4$ image mosaic and simultaneously
reducing crosstalk between spectral bands. A custom Convolutional
Neural Network (CNN) has specifically been designed to reduce crosstalk
and to increase the spatial resolution of the images created by this type of
hyperspectral camera system. Usually crosstalk (Hirakawa, 2008) is
considered detrimental. However, this research contributes by observing
that the spatial and spectral correlations in the hyperspectral information
6.2 Computer vision and deep learning

actually benefit the upscaling process. This way of using deep learning as a signal processing method demonstrates an answer to the question: *How can the quality and resolution of hyperspectral images be improved using deep learning?*

A notoriously difficult challenge in computer vision is to detect objects, which are connected in the image, as separate objects. In this dissertation two versions of the deep learning architecture CentroidNet have been introduced specifically for this task. The architecture was tested on aerial images of potato crops that were collected using a low-cost commodity UAV. This research contributes by showing that our approach achieves a higher F1 score for counting potato-crops compared to other one-stage object detection models in literature (Redmon and Farhadi, 2017; Lin et al., 2017c). It was found that the CentroidNet architecture is particularly suitable for counting small and connected objects as individuals by estimating their centroids. This provides an answer to the question: “How can deep learning be used to solve a challenging image-processing task using images produced by low-cost commodity UAVs?”

The trinity of technologies has provided an interesting framework to provide focus for the research on mitigating several inherent limitations of UAVs and hyperspectral cameras by using deep learning. However, the results of this research are not only relevant within this framework of thought. Spectral-band selection can be applied in any hyperspectral application to add efficiency for small platforms. Similarly, it has been shown that CentroidNet is applicable in a broader range of applications like cell-nuclei counting and bacterial-colony counting. CentroidNet shows increased or on-par performance compared to state-of-the-art instance segmentation methods (Redmon and Farhadi, 2018; He et al., 2017). Furthermore, the research in this dissertation contributes to a more fundamental insight in the relation between computer vision and deep learning which is discussed in the next section.

6.2 Computer vision and deep learning

The framework of this research is the trinity of the technologies mentioned in the main title of this dissertation and discussed in the previous section. The subtitle relates to the insight into the relation between computer vision and deep learning gained during this research. During experimentation
within this application framework, the particular usefulness of combining these two technologies became clear. This section elaborates more on this conclusion.

Nowadays many deep learning architectures exist and the number is still growing. Most of these have been designed to solve specific technical shortcomings of predecessor CNN models. For example, vanishing gradients are addressed by ResNet (Szegedy et al., 2017), spatial resolution problems are mitigated by U-Net (Ronneberger et al., 2015) and computational complexity is decreased by Xception (Chollet, 2017). These architectures are general, complex and are applicable in many areas. This makes them very widespread and extremely successful regardless of the application. The question arises if, for some applications, simpler, less computationally complex models could suffice? This could potentially reduce the run time, amount of data needed and even the ecological footprint of deep learning for specific applications.

General architectures can learn to model highly complex functions and generalization methods like weight decay are used to forget details during training, which in turn reduces the complexity of the model and make it perform better on specific applications. If the model of a task is partly known and can be represented with a custom CNN architecture, then this CNN is more suited for the task it was designed for. This method of incorporating prior knowledge results in a simpler model which is less prone to over-fitting. In this dissertation a CNN for crosstalk correction and upscaling was designed for images produced by a specific hyperspectral camera. Prior knowledge about the structure and size of the MCFA, or mosaic, of the hyperspectral camera is used to set the hyperparameters of the underlying CNN model. This includes the convolution filter sizes, amounts and strides. In traditional computer vision a solution would be engineered for a specific application and most parameters would be set manually. In the proposed approach a solution is designed in terms of specific convolutions and its parameters are trained end-to-end using deep learning. This successful combination of computer vision and deep learning to reduce model complexity for a certain task has been discussed in this dissertation in Chapter 3.

Deep learning has endeavored to remove the need for manual feature design and aims to learn solutions to practical problems in an end-to-end fashion without the need for traditional computer vision (LeCun et al.,
In this dissertation the usefulness of combining computer vision and deep learning has been shown by means of CentroidNet. The input image is preprocessed using a CNN so it can be easily processed by traditional computer vision algorithms to segment object instances. Where in the traditional tandem, computer vision is used to preprocess image information (Bougharriou et al., 2017; Suleiman and Sze, 2014), in the CentroidNet algorithm, traditional computer vision is used to postprocess information. This could be generalized to situations where in fully-engineered computer vision solutions, certain elements can be replaced with CNNs. With this method, difficult parts of the vision pipeline can be trained directly from data to improve the overall system performance. Difficult parts in this context can be elements which are hard to parameterize and which can be represented by CNNs. With this hybrid approach, in the computer vision parts, prior knowledge of the application is retained. This combination of computer vision and deep learning was successfully demonstrated in this dissertation in Chapter 4 and Chapter 5.

At the core of CentroidNet is the notion of trained 2D vectors. For each pixel in the image the CNN predicts two sets of vectors. One vector points to the nearest centroid and the other vector points to the nearest border of the object with the nearest centroid. Each vector can be considered a vote, and by aggregating votes the location of centroid and the delineation of the object is determined. This can be viewed as a variant of majority voting similar to a Hough transform or an ensemble of detectors. This approach has shown to be successful in other situations (Mukhopadhyay and Chaudhuri, 2015; Viola and Jones, 2001) and also contributes to the working of CentroidNet. In the ablation studies of CentroidNetV2 it was found that the loss function that was designed to better reflect the nature of the outputs of CentroidNet consistently achieved better results in predicting object instances. This shows that, when designing hybrid CNN models, like CentroidNet, in which the nature of the output of the algorithm changes, other parts of the training process should be redesigned accordingly. This was discussed in Chapter 5.

In the work that has been discussed in this dissertation sometimes a relatively small amount of images was used or very deep neural network architectures with many parameters have been used. It is important that the number of samples and the number of parameters in the deep
learning network is balanced. Each architecture that has been used in this dissertation performs a pixel-to-pixel mapping of the input image to the output. In some cases a traditional sliding window is used (like in Chapter 2) and in other cases a fully convolutional neural network is used to provide this mapping (Chapter 3, 4 and 5). This means that a single image logically consists of a large amount of samples from the perspective of the deep learning architecture. The number of samples from one image then depends on the size of the footprint of the model and the size of the input image. This is probably the reason why in the demosaicking experiments discussed in Chapter 3 no overtraining was observed. In those experiment both the footprint was small and the images were large.

When taking into account the break-through research and the multitude of new applications of deep learning for many classic vision tasks, it seems that the field of computer vision is mostly superseded by deep learning. However, the experiments in this dissertation have shown that designing custom deep learning algorithms for specific computer vision tasks and by combing both technologies, clearly provides an interesting direction for future research into deep learning.

6.3 Future work

Research into other challenges posed by the trinity of deep learning, hyperspectral imaging and UAVs can be performed in the future. In one direction, research could primarily focus on elements where all three technologies need to be combined. Alternatively, the definition of these three areas could be further broadened. For example, instead of deep learning, a more broader range of machine learning and artificial intelligence concepts could be included. Additionally, the hyperspectral imaging notion can be expanded to encompass multiple cameras to provide an even richer view of the electromagnetic spectrum to the deep learning approaches. For example, hyperspectral color cameras, short wave infrared cameras and thermal cameras could be combined. Along this line of thought the image sources can be expanded in multiple dimensions including 3D imaging and video. In future research UAVs could be synonym for “small platforms”, both in weight and computing power. This could include the small deployment platforms used in edge
computing like the Jetson series, Coral or Movidius \(^1\). The main research question can be restated to encompass this broader context: "How can artificial intelligence be utilized to mitigate the limitations imposed by small platforms that collect and process images from sources with high spatial, spectral and temporal dimensionality?"

In the next part of this section future research directions of the specific research topics that were discussed in this dissertation are proposed. In the concluding part of this section, possible future research into the combination of computer vision and deep learning is discussed.

From the 28-dimensional hyperspectral cube, three channels were identified for distinguishing between *Alternaria* and Ozone damage on potato-plant leaves in laboratory conditions. Future research could focus on testing this approach on potato plants in agricultural fields using UAVs. Additionally, using a similar methodology, research could focus on identifying other crop diseases like *Phytophthora*. Deep learning could then also be used to exploit key morphological features of the brownish lesions specific for certain crop diseases.

Filter mosaics are applied in many ways by sensor manufacturers. This includes regular Red Green Blue (RGB) Bayer filters and the hyperspectral MCFA sensors like the one used in this research, but also per-line-filter approaches are available. Recent camera sensors even use a mosaic of polarization filters combined with RGB filters. Future research could focus on crosstalk correction and upscaling within this family of mosaic filter patterns by using the approaches discussed in Chapter 3. Additionally, research could focus on using deep learning for upscaling beyond the native resolution of the imaging sensor using Hyperspectral Single Image Super Resolution (HSISR).

CentroidNet is a hybrid CNN for instance segmentation that is particularly suitable for counting objects in images. Future research could focus on replacing other parts of the algorithm with a CNN. For example, the voting space might be postprocessed with an additional deep learning module to provide clearer centroids which are easier to locate with traditional computer vision methods. This would reduce the amount of hyperparameters while still retaining the core design of centroid and border majority voting. Future research could investigate how, in other

\(^1\)nvidia.com, coral.ai, movidius.com
traditional computer vision methods, parts can be replaced with deep learning to form new hybrid algorithms.

This dissertation started with the classic tandem of feature extraction to feed the machine learning models. Subsequently the idea of designing custom CNN architectures for specific applications was discussed. The final part of this work focused more on preprocessing images with CNNs and then using deterministic computer vision algorithms to do the final processing. As mentioned earlier this can be generalized to replacing parts of computer vision programs by CNNs which consequently allows these parts to be trained from data. This line of thought can be extended into a future where (parts of) traditional software algorithms can be replaced by their trainable counterparts. This does not necessarily have to be limited to redefining software parts in terms of convolutions. But when a software program is differentiable and there is enough training data available it can likely be trained. This would instigate a paradigm shift in how scientific software is developed. Future research could focus on how to develop software that is differentiable in a general sense so it can be trained with gradient descent methods. This is already an active field of research called Differentiable Programming ($\partial$P). In a $\partial$P system, graphs are directly represented by the source code of the software program and can sometimes be compiled and optimized directly. An introduction into $\partial$P is given by Innes et al. (2019). In recent work Riba et al. (2019) provide a differentiable computer vision library. These examples indicate that, in the future, the defining differences between the fields of computer vision, deep learning and, ultimately, software and algorithm development will probably fade when an increasing amount of software is developed to be trainable with gradient descent methods.
Bibliography


Hallermann N. and Morgenthal G., Unmanned aerial vehicles (uav) for the assessment of existing structures. IABSE Symposium Report, volume 101 (International Association for Bridge and Structural Engineering 2013), (pp. 1–8).


Lowe D.G., Object recognition from local scale-invariant features. *International Conference on Computer Vision*, volume 99 (1999), (pp. 1150–1157).


Mazur M., Six ways drones are revolutionizing agriculture. *MIT Technology Review*.


Rebetez J., Satizábal H.F., Mota M., Noll D., Büchi L., Wendling M., Cannelle B., Perez-UrIBE A. and Burgos S., Augmenting a convolutional neural network with local histograms - a case study in crop classification from high-resolution UAV imagery. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2016).


Statt N., The ai boom is happening all over the world, and it’s accelerating quickly. *The Verge*.

Suleiman A. and Sze V., Energy-efficient hog-based object detection at 1080hd 60 fps with multi-scale support. *2014 IEEE Workshop on Signal Processing Systems (SiPS)* (IEEE 2014), (pp. 1–6).


Seventh International Conference on Information Science and Technology (ICIST) (2017), (pp. 464–469).


