A Study on Attention Model and Computational Optics for Pain Face Detection and Fruits Segmentation

by

Jun Yu

Student ID Number:1218005

A dissertation submitted to the
Engineering Course, Department of Engineering,
Graduate School of Engineering,
Kochi University of Technology,
Kochi, Japan

For the degree of
Doctor of Engineering

Assessment Committee:
Supervisor: Prof. Toru Kurihara, School of Information, Kochi University of Technology
Co-Supervisor: Prof. Shinichi Yoshida, School of Information, Kochi University of Technology
Co-Supervisor: Prof. Shu Zhan, School of Information, Hefei University of Technology
August, 2021

Abstract

A Study on Attention Model and Computational Optics for Pain Face Detection and Fruits Segmentation
Jun Yu
Doctoral Program of Engineering Course, Department of Engineering,
Graduate School of Engineering

Pain is one of the primary feelings that encourage people to seek medical care and attention. Hence, automatic estimation of the sense of pain has lots of applications in medical treatment. However, the feeling of pain is an individual and distinctive experience which differs from each person. Recently, there have been several attempts to model the attention ability inside the neural network to improve the performance. Inspired by the previous work in the attention model, we proposed the spatial attention model and inserted it into our convolutional neural network. Our neural network can find the most correlated region on the human face for pain facial expression detection and analysis aiding by our spatial attention model. Experimental results show that our locally spatial attention learning can provide the fine-grained variation on the face region for pain intensity assessment. Our current study extends the prior work in this research area and provides a new method for future studies on painful expression analysis.

Accompanied by the increasing labor cost and the growth of the aging population in Japan, the automatically harvesting system's development is one of the popular research topics in machine vision, computer vision, and robots. The hyperspectral camera can capture not only the visible wavelength but also the near-infrared wavelength. It can

provide many more details about living plants and fruits. Motivated by the previous study of the channel attention model and Long short-term memory(LSTM), we proposed our novel neural-network-based algorithm for green pepper segmentation. Unlike the other research work, our method is a pixel-wise framework for green pepper segmentation. In particular, we select some pixels from the green pepper and some pixel from the foliage. We treat each pixel as a long vector to put into our proposed framework. There are two parts of our deep neural network. The first part is our proposed channel attention module, and the second part is the LSTM. By utilizing the memory function of the LSTM, our proposed structure can use the critical wavelength information to distinguish green pepper. The whole network can train in an end-to-end manner. The experimental results on our hyperspectral dataset show competitive performance for our proposed approach. Our proposed channel attention model considers the intrinsic feature of the hyperspectral data. Besides, our proposed methods require a smaller dataset, lower hardware requirements, and faster, compared with the deep convolutional neural network. However, our approach is vector-based machine learning. The hyperspectral camera can provide not only spectral details but also provide the spatial details of the captured scene.

Computational Optics provides a way to co-design optics devices, camera sensors, and spectral illumination. Recently, the proposal of end-to-end optimization of optics, sensor and camera pipeline is attracting widespread interest. In our research, we proposed to utilized an optical filter to enhance a red-green-blue(RGB) camera system to detect green pepper or immature yuzu citrus. We use the power of the deep neural network to help us to find the best optical transmission curve of the optical filter for a specific task. Specifically, we reported the relationship between the depth-wise convolutional

kernel and the transmission curve of the optical filter. The weights of the first layer and weights of the second layer in our neural network represent the optical filter and color filter array (CFA) inside the camera. Hence, we can represent the physical device by the depth-wise convolutional layer and convolutional layer. To our best knowledge, it is the first proposal to represent the optical devices by the deep neural network layer. Consequently, we can use the deep neural network to find the optimal transmission curve for the optical filter in a data-driven way. Both optical filter and color filter array must satisfy specific physical requirements that constrain both the spectral transmission curve and camera spectral response. The experimental results show our proposed framework can achieve better detection results than an RGB camera without an optical filter. Besides, our proposed camera system is cheaper and easy to use compared with the multispectral and hyperspectral cameras.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

In recent years, we have witnessed the rapid development of computer vision and machine learning research due to the success of deep learning. With the increasing amount of data and computational power, we can more easily train the deep neural network than before. Since the deep neural network has been successfully adopted for image classification, object detection, etc. There are still many new problems and applications worth applying the power of deep neural networks. In this study, we aim to leverage the power of the deep neural network to affective computing and computational optics. Specifically, we combined the attention module and deep neural network to the pain face estimation for aiding the diagnosis of dementia. Our proposed attention-based method can provide a hint to help to analyze the facial pain expressions. In addition, we utilized the deep neural network to end-to-end optimize optical filter design and fruits segmentation for supporting agriculture applications. Following our proposed system, we can quickly adapt to other tasks and reduce the cost of the whole system.

The prevalence of dementia in older people becomes a severe concern worldwide. Significantly, the Japanese elderly confront the need for long-term health care and the risk of high dementia prevalence [2]. The damage lead by the disease dementia is broad in an extensive range, consisting of judgment, language, learning, and social activity [3]. Dementia leads to severe and particular obstacles to pain assessment [4]. Amanda et al., in their study [5], report the facial expression is sensitive and specific non-verbal

evidence of pain. The analysis of the facial pain expression is one of the concise tools for identifying pain within patients with moderate to severe levels of dementia who can no longer self-report pain [6]. Previous research has widely investigated the response to pain in the human face using the Facial Action Coding System(FACS). The FACS is an empirical, fine-grained, and anatomical facial coding system. However, it is labor-intensive and time-consuming and requires much effort in personnel training for skilled experts. Hence, it is not widely available in clinical treatment and increases the economic burden on the whole society. Consequently, there is a growing demand for automatic pain estimation by using facial expression analysis. As reported by previous research, not all the face areas are essential to the pain face expression, and facial expression usually represents by a video sequence. To address these features, we propose the attention model with the deep learning technique for automatic pain estimation

Agriculture is the pillar industry and primary economic income in the Kochi Prefecture. However, Kochi's agricultural sector has been facing many issues, such as an aging population, population migration to the metropolitan, and an increase in the deserted farm field. Our research aims to utilize advanced deep learning technology with computational optics to develop precision agriculture in the Kochi prefecture. Improving the quality and production of the economic crop while aiding the management and marketing strategy is one of the critical aims of precision agriculture. Precision agriculture can provide useful information in the early stage to enable better decisions to make on the management system. In recent years, computer vision and artificial intelligence technology have developed to meet the growing demand for fast and accurate grain crop production [7] [8]. As reviewed by a previous study [9], Machine Learning techniques have been widely used for the early and precise detection of biotic stress in the crop, specifically for the detection of weeds, plant diseases, and insect pests. Estimation of the leaf-to-Fruit ratio is a critical indicator of the yield number and berry composition for fruits [10].

1.2 Attention model

A critical characteristic of the human vision system is the attention mechanism. The human visual system does not tackle to process a whole scene at once. Instead, the system deals with a sequence of partial glimpses of the image and selectively focuses on the important part to better understand visual structure. Combining the information of different partial glimpses of the image can build up a scene representation and guide advanced eye movements and decision-making. Much research in recent years has focused on utilizing attention mechanisms in the neural network structure to improve performance.

Recently, state-of-the-art computer vision methods have relied on intense deep neural networks, which are computationally expensive. The reason behind it is that the most deep neural network process the whole image at once, and the amount of the parameter scales linearly with the number of the layer and image pixels. The principle of attention mechanisms is that they are an imitation of processing in the human visual system. Mnih et al. [11] proposed a novel recurrent visual attention model, which extracts visual information from an image and adaptively processes it. Vaswani et al. [12] proposed using the attention mechanism to replace the traditional RNN and Seq2Seq model [13] in the natural language processing research area. Recently, a form of the attention-based method, self-attention is introduced, which is similar to Non-Local Neural Networks [14]. The attention-based method is a flexible building block and can be easily integrated with the deep neural network.

1.3 Computation optics

In the research field of medical, scientific, and industrial optical and imaging applications, it is feasible to co-design the illumination, optical element, sensor, and imaging processing algorithm. Unlike the traditional hardware(e.g., optical elements) and algorithm(e.g., computer vision and machine learning algorithm) design, , which treat them separately, computational

optics can jointly optimize both of them to achieve better results for a specific task. Hyttinen et al. [15] introduced to use of partially negative filters to optical implement partially negative filters for contrast-enhanced oral and dental imaging. Wang et al. [16] developed multiplexed(coded) illumination to classify similar visuall objects, such as real and synthetic fruit and vegetables. Boominathan et al. [17] demonstrated to design phase-mask-based thin lensless camera system. The key idea of their work is to utilize an optimized mask and computation method to replace the camera lens, which makes the camera more compact than ever before.

There has been an increasing amount of literature on end-to-end jointly optimize optical elements and corresponding algorithms by using deep neural networks in recent years. The concept of deep optics has been demonstrated to contribute significant benefits for various applications in spectral signal reconstruction [18], monocular depth estimation [19] [20], high dynamic range imaging [21], and computational microscopy [22]. The main challenge on deep optics is listed as follows.

- How to represent the optical element as the parameters of one layer in the deep neural network?
- How to design a suitable physical constraint for the fabrication of the optical device?
- The loss function is a critical part of any machine learning and deep learning algorithm. How to choose the proper loss function to optimize the optical element and algorithm jointly?

To address the above issues, we introduce our proposed method for green pepper segmentation in Chapter 4 and Chapter 5.

1.4 Organization of the thesis

The thesis is organized into four parts. The first part is introduction chapter, which provides the background information about the study and summarizes the contribution of the original papers. The second part consists of the Chapter 2 and the Chapter 3. The third part is composed of the Chapter 4 and the Chapter 5. Lastly, we make an conclusion about the whole research in Chapter 6.

In the second part, we reported our proposed attention-based method for pain face estimation and green pepper segmentation. Due to the fact, not all the face areas can contribute to the pain face expression. We proposed the spatial attention module for extracting features from the pain face. We demonstrated our proposed spatial attention method could achieve better results than without the spatial attention method.Hyperspectral imaging can capture more detailed information than RGB images. In this study, we proposed a novel attention-based LSTM model that can effectively leverage the hyperspctral pixels to distinguish green pepper and leave. Specifically, our proposed method uses the band attention mechanism to detect the essential band information for classification. Experimental results in our hyperspectral image dataset suggest our approach can outperform other methods

In the third part, we introduced our proposed method to jointly end-to-end optimize the transmittance curve of the optical filter and the parameters of the image segmentation network. To our best knowledge, it is the first time to use the depthwise convolutional layer to represent the transmittance curve of the optical filter. As a result, we successfully treat the transmittance curve of the optical filter as one layer of the whole deep neural network structure. To aid our proposed optical filter fabrication, we presented the physical-based constraint on the optical filter layer. Under the physical-based constraint, the transmittance curve becomes smooth and non-negative. Inspired by the previous research on the ratio of R, G, B channels, we proposed a color-ratio map enhanced method for optical filter design.

Our proposed method was validated on our hyperspectral image dataset and demonstrated to achieve better results than the purely CNN-based method(no optical filter setting).

In the last part, we summarize the whole research work and discuss the future work. For pain face estimation, we expect to utilize the 3D landmark to estimate the pain face's intensity. In the topic of optical filter design, we are interested in exploiting the color space to aid the optical filter design in our future work.

2 Frame by Frame Pain Estimation Using Locally Spatial Attention Learning

Estimating pain intensity for a patient is a challenging area in clinic treatment and medical diagnosis. The painful facial expression only relates to some areas of the face. Inspired by this fact, we introduce end-to-end locally spatial attention learning for pain estimation. By focusing on an important region in the face with 1×1 locally convolutional layer, the local features related to pain intensity can be captured. Furthermore, the facial expression is the dynamic deformation of the face in the time domain. To model the information, the long short-term memory network(LSTM) is incorporated into our architecture. The feature extracted by the CNN with the locally spatial attention learning is fed to the LSTM network. The results show that our locally spatial attention learning can provide the fine-grained variation on the face region for pain intensity assessment.

2.1 Introduction

Pain is an unpleasant feeling which is related to tissue damage and unhealthy condition. Accurate pain intensity estimation is a central problem in mental health and clinical treatment. Traditionally, pain intensity is evaluated by the observation of expert and self-reported data, such as Observer Pain Intensity (OPI), Visual Analog Scale (VAS). However, for elderly people with dementia who lack the ability to express pain intensity, evaluating pain intensity becomes a basic issue in some medical diagnosis applications. In addition, manual pain estimation is time-consuming, inaccurate without professional training and not available for real-time pain assessment. In such situations, accurate pain intensity

evaluation plays an important role in medical treatment and health care. Hence, there is a large demand to build automatic assessment system for pain intensity estimation.

To solve the great demand, a large majority of research has focused on automatic pain intensity assessment. In the initial stage, detecting pain in video by facial action units has been proposed by Lucey.et al [23]. Later, some methods have been proposed to evaluate pain intensity using multimodal data, such as thermal and depth data from camera [24], biomedical signals from the electrocardiogram signals (ECG) and the electromyogram signals (EMG) [25]. Recently, deep convolutional neural networks have achieved great successful results in face recognition, face detection and so on. Therefore, deep neural networks are attracting widespread interest in the fields of facial expression recognition, especially pain intensity estimation. Recurrent Convolutional Neural Network used for object detection was utilized by Zhou et al. [26] for pain intensity estimation. Another method was developed by fine-tuning deep face verification net with regularized regression loss [27]. Pau et al. [28] proposed a method by combining deep convolutional neural networks with long short-term memory networks for pain intensity estimation. Their study suggested extracting features from both the spatial space and the temporal space can obtain good performance for frame-level pain intensity estimation. Tavakolian et al. [29] developed a method by using binary coding of discriminative statistical feature representation from the convolutional neural network. Hamming distance is applied in the new loss function and benefits the training of the whole framework.

Attention mechanism is one of key properties of human being's visual system which can selectively concentrate on the important areas in an image or a scene for better understanding. Inspired by that, there are several attempts to utilize attention mechanism to improve the performance in Image captioning [30] and other applications. More recently, a concise attention module was proposed by Hu et al. [31] to build the relationship between different channels inside the neural network. The global average pooling was used for estimating

the channel-wise attention. Later, Woo et al. [32] proposed new attention model called Convolutional Block Attention Module(CBAM). In the CBAM module, the attention mechanism was applied not only in the channel space, but also in the spatial space. Extensive experimental results have shown that CBAM module can achieve the best performance in both the image classification task and the object detection task.

Until now, a few research in the field of pain intensity estimation has attempted to utilize the attention mechanism in their research work. The purpose of this study is to propose and examine end-to-end locally spatial attention learning architecture for pain intensity assessment. The overview of the pipeline of our approach is illustrated in Fig. 2.1. The approach we have applied in this work aims to exploit "where" is important in spatial space for pain intensity estimation. Besides, our architecture also exploits the relationship between different frames in the video sequence. The proposed attention-based architecture is validated in the widely used benchmark database [33]. The chapter is organized as follows: In the method part, we describe the details of our approach. In the experiments part, we investigate and analyze our proposed method in the database. Finally, the conclusion is given for our research.

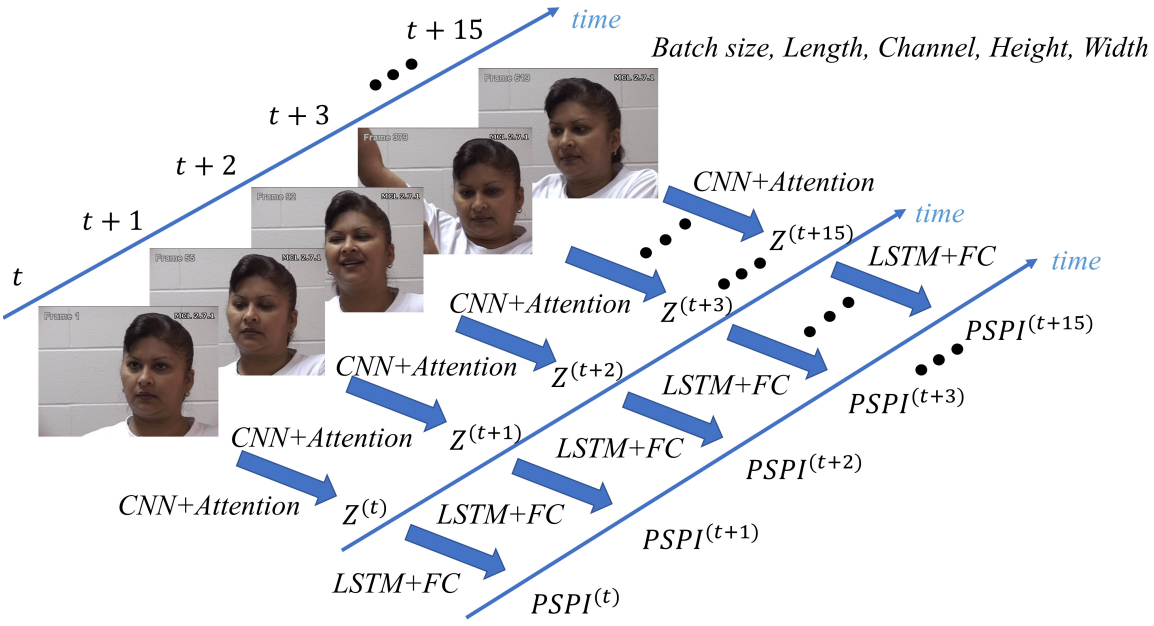


Figure 2.1: The illustration of the whole pipeline of our architecture. The input of the architecture is a five-dimensional tensor, including batch size, length of the sequence, channel, height, width.

2.2 Proposed Method

In a comprehensive literature review of Pain, Craig et al. [34] demonstrated that facial expression could be adequate evidence for identifying pain. The purpose of our study is to estimate the pain intensity directly from the patient's face in a recorded video or in the real-time surveillance system. The motivation of our method is that each region of the face is not equally contributed to the painful expression. In order to capture the local detailed variation of face, we propose the locally spatial attention learning architecture for pain assessment. Our structure is based on the VGG network [35] and previous face recognition work [36]. The input tensor is rescaled by the locally spatial attention model, which can enhance the ability of our network for extracting features from images. Some previous behavioral and emotion study suggests the dynamic information of facial expression is useful and efficient for emotional assessment[37]. In our architecture, the LSTM network is adopted for capturing the dynamic information in the temporal domain. By combining

the spatial variation and the temporal variation in the video sequence of patient's face, we are able to estimate the pain intensity robustly. Our architecture consists of CNN with spatial attention model and LSTM. Each frame in the video sequence is fed into the whole architecture. The CNN block extracts features from single frame, then puts the feature vector into the LSTM block to estimate the pain intensity. The details of our architecture is shown in Fig. 2.2. In the next section, we elaborate on the details of our approach.

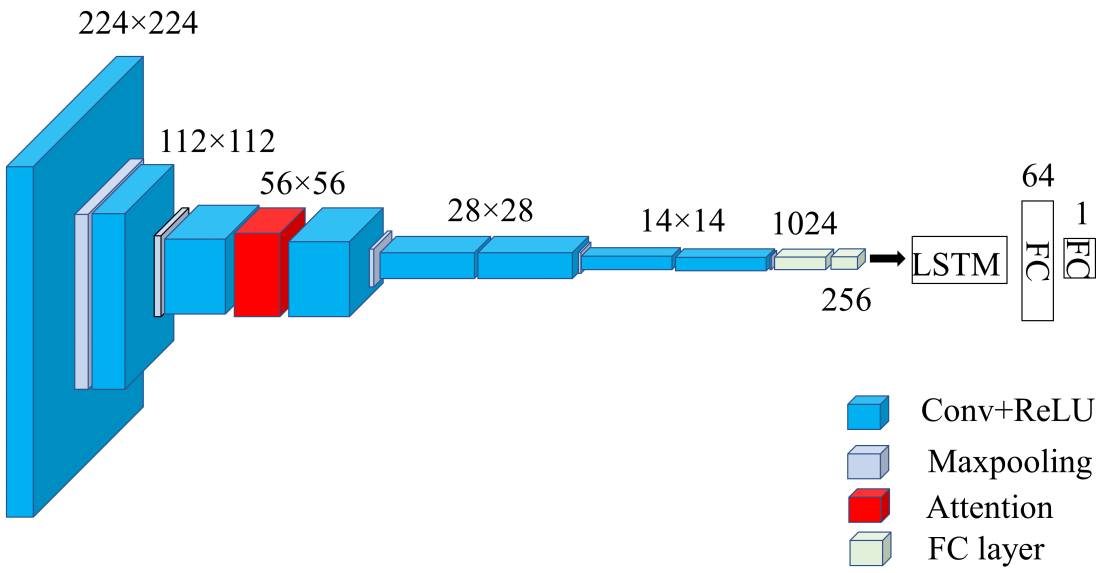


Figure 2.2: In our study, we utilize CNN network and LSTM network as the backbone of our architecture. We incorporate the locally spatial attention learning model into the convolutional network as illustrated in this figure.

2.2.1 Locally spatial attention learning

Figure 2.3 shows the overview of our locally spatial attention learning model. For extracting the static features from the patient face, we utilize the convolutional neural network which is based on VGG11 network(configuration A)[35] for pain intensity estimation. The motivation of our study is that each region of the face is not equally contributed to the painful expression. In an attempt to design the spatial attention model, our intention is to provide a way of detecting the important region for pain intensity estimation. The proposed

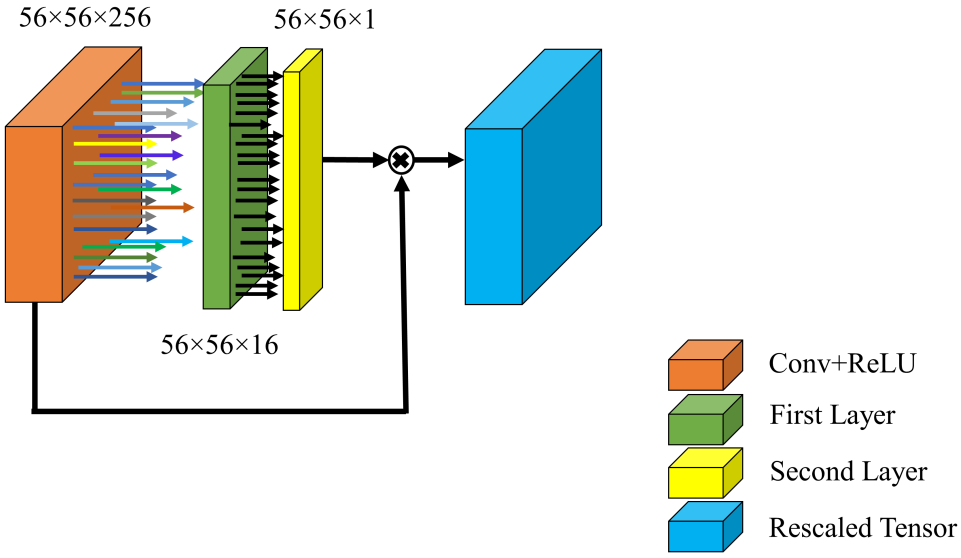


Figure 2.3: The overview of the locally spatial attention learning model. As illustrated in Fig. 2.2, the spatial attention model is inserted inside the third block of convolutional neural network. The first layer is 1×1 locally convolutional layer. The second layer is the conventional 1×1 convolutional layer for dimension reduction. The input of the attention model is the orange block which is the output of first layer in the third block of convolutional network. The spatial attention map is used for rescaling the input tensor.

model which consists of two layers is inserted inside the third block in the convolutional neural network to capture the detailed information from the previous layers.

The aim of the locally spatial attention learning is to capture more detailed information from the face region. A major problem with the previous attention model based on the conventional convolutional kernel is that the generated attention map is translation invariant so that the local details of the image are hard to capture. The geometry attribute of the face image is symmetrical and structural. Inspired by previous face recognition study[36] and other research field based on the facial expression[38], we propose the locally spatial attention learning model which is incorporated into the convolutional neural network. Given the output tensor T from the previous building block in the convolutional neural network, the shape of the tensor is $C \times H \times W$, which C is the channel number of the tensor while

H and W represent height and width respectively. The spatial attention model takes the input tensor T and generates a 2D spatial attention map A_s, with size $H \times W$. To generate the spatial attention map, the locally convolutional layer is adopted in the first layer of the spatial attention model. For each location in the spatial dimension of the input tensor, the first layer of our locally spatial attention learning model uses different convolutional kernel for extracting discriminative appearance feature. The P_{ij} denotes the different weights for the input tensor T of different location T_{ij}. Each T_{ij} has its own receptive field of the face image. In hence, the more the spatial attention model is behind, the larger the receptive field of the attention model becomes. The kernel size of the locally convolutional layer is 1×1, so shape of the output tensor of the first layer is $R \times H \times W$, $R = 16$. The $\tanh$ function is used as activate function in the first layer. In the second layer of spatial attention model, we apply the conventional 1×1 convolutional kernel to generate the spatial attention map, which describes the informative parts of the face region. The sigmoid function is applied on the top of the spatial attention model to let the attention weight lie from zero to one. The shape of the attention map A_s is $1 \times H \times W$. In short, the spatial attention model is calculated as:

$$T_{res} = T \otimes A_s \tag{2.1}$$

where, the operation $\otimes$ denotes the element-wise multiplication. The output of the attention map rescales the input tensor T. In our implementation, the attention map A_s is broadcasted in the channel dimension of the input tensor T. Then, the rescaled tensor T_{res} is fed to the latter convolutional layer in the convolutional neural network. We utilized dropout strategy to avoid the overfitting problems. The dropout ratio of the fully connected layer was set to 0.3. The arrangement of the two layers inside the attention model is a key problem for pain intensity estimation. We compare different spatial attention models in the experiments section, and the results demonstrate that locally convolutional layer in the first order is better than other structures.

2.2.2 Temporal learning

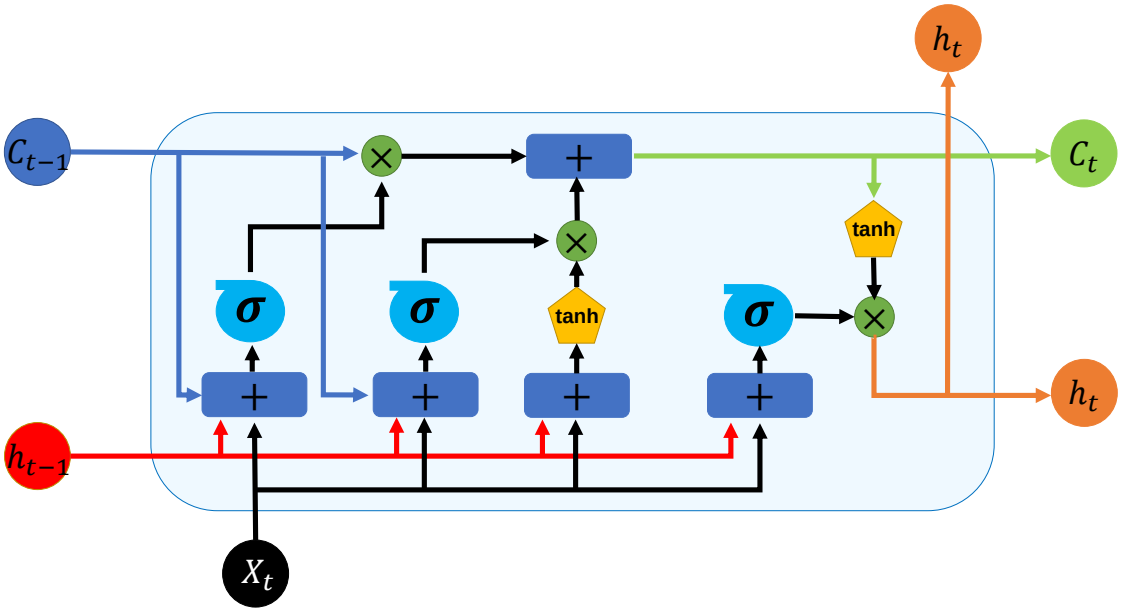


Figure 2.4: The illustration of the LSTM structure. C_{t-1} denotes the memory from the previous block. h_{t-1} represent the output from previous block. X_t denotes the input vector. σ denotes the sigmoid function. C_t and h_t represent the memory from the current block and output of the current block, respectively.

The structure we have used in this study aims not only to extract features from the single frame, but also to build the dynamic temporal relationship among different frames. For this purpose, the LSTM [39] network for learning the long-term relationship among sequence data is adopted in our temporal model. The output feature vector $Z^{(t)}$ of convolutional neural network is fed into the LSTM, so the input node of the LSTM is 256-dimensional. $Z^{(t)}$ denotes the feature vector of tth frame in the video sequence. In our temporal model, we only use one LSTM layer. The hidden node of our LSTM is 128-dimensional. The dropout ratio of our temporal model is 0.3. The LSTM network has three gates to control the information from the previous sequence data and the existing sequence data. The video sequence of the patient is divided into small groups to train our architecture. For instance, the first sequence $s_1 = \{f_1, f_2, \ldots, f_{16}\}$, is 16 consecutive frames from one video , the

next sequence is $s_2 = \{f_2, f_3, \ldots, f_{17}\}$, until the last sequence of this video. Here, f_i denotes the i th frame in the video sequence. The frame label of our training database is Prkachin and Solomon's Pain Intensity metric(PSPI)[40]. The label of each sequence is the PSPI label of the last frame in this sequence. Therefore, the PSPI label is predicted by considering all the 16 frames. Finally, two fully connected layers are used for predicting the PSPI value based on the output of the LSTM network. The dimension of the first layer is 64 and the dimension of the second layer is only 1 for estimating pain intensity. Considering that estimating the pain intensity is the regression task, we chose the mean squared error loss for training our architecture.

2.3 Experiment

In this section, we will discuss the details of our experiments and results for our proposed architecture.

2.3.1 Database and preprocessing details

We trained and validated our spatial attention architecture in the UNBC-McMaster shoulder pain database[33], which consists of 25 subjects with 200 videos. All the participants in this database have got shoulder pain. In the recording stage, they did a list of active and passive range-of-motion tests with their limbs under the professional guidelines. The database provides three types of labels for calculating pain intensity, including VAS, OPI and PSPI. The PSPI intensity can be calculated as:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43 \tag{2.2}$$

As reported in the previous research[41], the PSPI label in the UNBC-McMaster shoulder pain database is not always reliable. In their finding, both VAS and OPI labels for some subjects are not zeros which means the subject feels pain. However, the PSPI label for that

subject is zero. It is known from the literatures[42], that some action units related to the pain expression is not calculated in the original PSPI equation. To improve the accuracy of our proposed spatial attention model, we conduct the data cleaning process for reliable PSPI labels in the database. Therefore, we excluded one subject which has no obvious pain (101-mg101 subjects, including 9 video sequences) and some video sequences without reliable PSPI labels (bg096t2afaff, ib109t2aeaff). We illustrated the sample sequence of the excluded video in Fig. 2.5. Totally, 24 subjects with 189 video sequences were used in

(a) Example of bg096t2afaff

(b) Example of ib109t2aeaff

(c) Example of mg101t1aiaff

Figure 2.5: Example video sequence removed from our experiment. (a) The example shows patient feels pain. However, Lip parting(AU25), Lip stretching(AU20)), raised eyebrows(Au1/2) is not considered in PSPI equation. (b) The VAS(VAS=5) and OPR(OPR=1) of sequence ib109t2aeaff are not zero. Otherwise, the PSPI is annotated as 0. (c) There is no significant pain for 101-mg101 subject.

our experiments. We followed the previous research[27] to preprocess the PSPI label by transforming the range of value from $0 \sim 15$ to $0 \sim 5$. Data preprocessing is a major part in training deep neural networks. As demonstrated in Fig. 2.6, the OpenFace2.0 toolkit[43] was utilized in our experiments for face alignment and cropping.

2.3.2 Implementation and analysis

The convolutional neural network in our architecture was trained from scratch and the whole structure was trained in the end-to-end manner. As mentioned before, we only used 24 subjects with 189 video sequences to train our network. Therefore, we evaluated our network in 24 subjects leave-one-subject-out cross validation. The learning rate of our network was set to 0.0001. Our network was trained in 20 epochs. Furthermore, we utilized Adam[44] optimizer with weight decay 0.001 to train our architecture. The whole architecture was implemented by PyTorch 1.10 [45] framework with batch size 32 in 4 GPUs. The order of the locally convolutional layer and the conventional convolutional

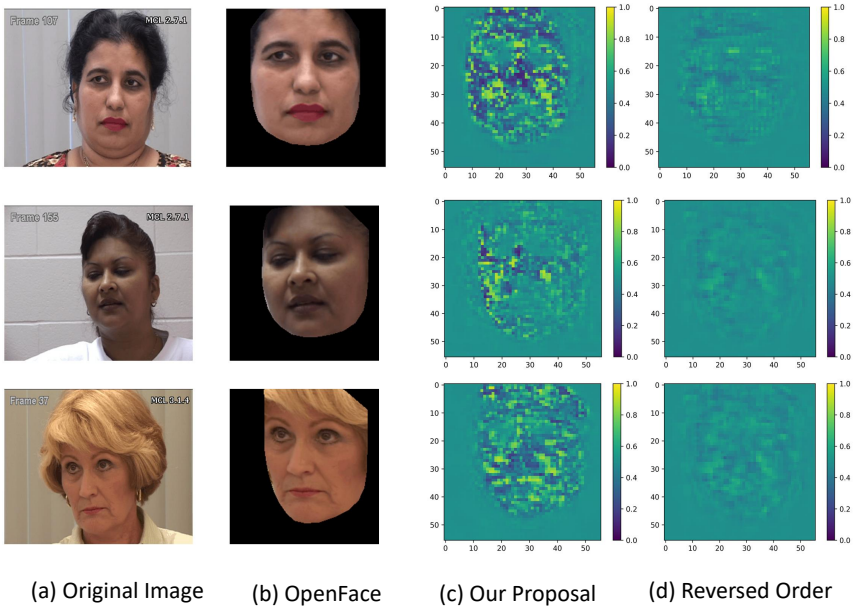


Figure 2.6: Original image and processed image by OpenFace 2.0 toolkit. All the images in the database are resized to 224×224. Example of two different attention maps is illustrated. (c) Attention map is derived from our proposed attention model. (d) The reversed order is comparison model which exchanges the order of 1×1 locally convolutional layer and 1×1 common convolutional layer.

layer in the spatial attention model is an essential issue in our study. Here, we compare two different spatial attention models. As illustrated in Fig. 2.6, the left attention map is derived from the locally spatial attention learning model used in our architecture, while the

Table 2.1: Comparison of different methods.

Methods	MAE↑	MSE↓	PCC↓
Zhou et al.[26]	N/A	1.54	0.65
Wang et al.[27]	0.456	0.804	0.651
Rodriguez et al.[28]	0.5	0.74	0.78
Tavakolian et al.[29]	N/A	0.69	0.81

right attention map is from the different spatial attention model which the first layer is the convolutional layer and the second layer is the locally convolutional layer. As can be seen from Fig. 2.6, our proposed locally spatial attention learning model indicates that the cheek of the face and the region between eyebrows are important for pain intensity estimation. The right attention map shows nearly same importance in the whole face region which indicates this architecture cannot detect significant region for pain intensity estimation. Comparison of two different attention maps shows our proposed model can capture the important region of face more effectively than another model with different order. We also compare the performance of our method with the general architecture and the previous research. Here, the general architecture only contains CNN network without locally spatial attention learning and LSTM network. The mean absolute error (MAE), mean squared error (MSE) and Pearson Correlation Coefficient (PCC) are reported in Table 2.1. As listed in Table 2.1, it is obvious that the locally spatial attention model can achieve an improvement on both MAE and MSE with comparison of general architecture. Comparison between our method and previous research shows the performance of our architecture is not perfect. It should, however, be noted that we use smaller training database for our neural network. Accurate and reliable PSPI label is important for training deep neural networks. Estimating the Pain intensity should be accurately related to the painful expression and feeling which are crucial issues for some medical diagnosis.

2.4 Conclusion

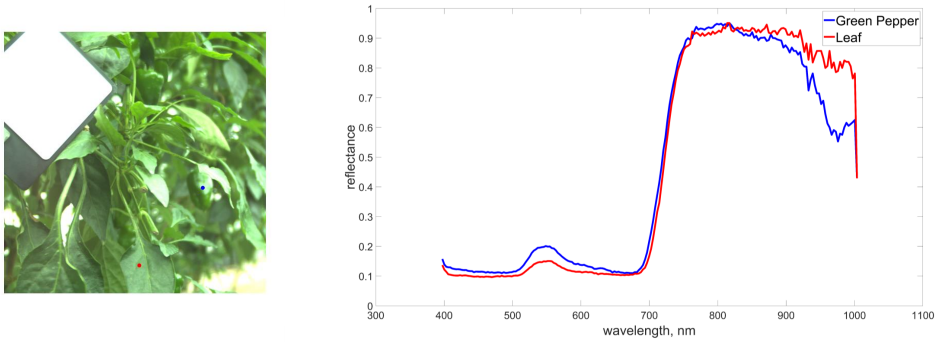
Automatic pain intensity estimation is a key technique in some medical applications. In this research, we propose locally spatial attention learning method to find important region on the face and to enhance the performance of the whole architecture. The results indicate that our proposed method can capture the important area of face for pain intensity estimation. Our current study expands the prior work in this research area and provides a new method for future study on painful expression analysis. We conducted our experiments in the shoulder pain database. At present, the results show the performance of our architecture is better than the general structure without locally spatial attention learning and is not outstanding compared with the state-of-the-art methods. We will improve our spatial attention architecture for better results by effective network engineering in the future.

3 Green Pepper Segmentation by Attention LSTM

3.1 Introduction

Green Pepper is one of the main economic crops in Kochi prefecture, Japan. In 2019, we began the IOP(Internet of Plant) project, aiming to improve agriculture's efficiency and production. Due to the similar color between the green pepper and foliage, our human vision system is difficult to detect green pepper by judging color. As a result, it is harder to develop automatic picking up techniques and more channeling to estimate yield for the marketing decision. An automatic harvesting system for vegetables and fruits attracts widespread interest in research fields, such as computer vision, robots [46], and the Internet of Things. Several methods for distinguishing a green pepper from its leaves in the research fields of image processing and computer vision. Recently, some research has focused on hue, saturation, and the value of the color space to classify a green pepper and its leaves [47]. An alternative approach was developed using a three-dimensional (3D) sensor to detect and segmentation of fruits and vegetables in 3D space [48]. Despite this, there remains a need for an algorithm to identify crops for which leaves and the remaining backgrounds are similar colors. The final goal of this research is to obtain good segmentation results for automatic picking or prediction of green pepper growth. For such a purpose, object detection is not suitable since the detection results cannot be used for cutting the stem of fruit or estimation of fruit size, although it can be used for counting. One approach to improve the system is to incorporate hyperspectral imaging, which collects and transfers various information at different wavelengths; hence, it can provide rich information for classifying vegetables and leaves, even if the colors are similar. Hyperspectral image

Figure 3.1: Where is the green pepper? In the real world, we hardly see the monochromatic object. Instead, the surface of the object reflects a wide range of wavelengths of light. Due to our eyes can only perceive three primary colors, Red, Green, Blue, it leads to the different object surface reflectance but matching the same or near color in our vision system. It is the Metamerism. The photograph was captured in the greenhouse.



(a) RGB image (b) Reflectance of sample points

Figure 3.2: The above image illustrate the different reflectance of the green pepper and green leave. (a) illustrates the sRGB image and selected point location. (b) demonstrate the reflectance of each selected point on the surface of green pepper and leaf.

segmentation is most commonly used in remote sensing [49] [50]. Its research purpose is to achieve land cover classification, including vegetation mapping. In the agricultural field, hyperspectral imaging has been used to detect green citrus using pixel-wise linear discriminant analysis followed by spatial image processing [51]. Its purpose is to improve segmentation results using all the information in high-dimensional space. Therefore, unlike RGB, it does not need to degenerate to three dimensions from a higher spectral dimension.

Image segmentation is a major topic in image processing and computer vision, building a solid foundation for understanding images and solving other computer vision tasks. Traditional image segmentation is based on threshold methods [52], edge detection methods [53], and clustering methods. Much more research in recent years has used deep learning to improve the performance of image segmentation. A breakthrough was achieved by fully convolutional neural networks (FCN), which convert a fully connected layer to a convolutional

layer [54]. Despite the successful results of the FCN model for dense pixel-wise segmentation, it still has some problems regarding losing the spatial resolution of feature maps and some boundary information of objects. To solve these problems, the SegNet [55] structure was proposed, which adopts encoder and decoder architecture to improve performance. The encoder part of SegNet is the same as that of the first 13 convolutional layers of the VGG16 network [35], and the corresponding decoder part uses pooling indices to upsample the feature map. Because of the pooling indices and fully convolutional layer, SegNet is smaller and quicker than other architectures in the training and evaluation stages. However, most image segmentation studies have focused only on RGB space. Few studies have addressed the continuous wavelength in the real world. In the hyperspectral image classification and segmentation area, Gao et al. [51] proposed to use a convolutional neural network with multiple feature learning for hyperspectral image classification. Mou et al. [56] firstly proposed to utilize the deep recurrent neural network for hyperspectral image classification. Their idea is a vector-based method, which treats the one pixel of the hyperspectral image as one vector. By applying the recurrent neural network, the hyperspectral image can be treated as a sequence-based data structure. Besides, they also proposed the new activate function, which is called the parametric rectified $\tanh$ function.

In this chapter, we present a novel approach to use a hyperspectral image and the power of a deep neural network to improves the segmentation results of green peppers. Although a hyperspectral image carries a large amount of information, a hyperspectral camera is expensive, and hyperspectral data processing requires a large amount of memory; hence, there are some difficulties in introducing hyperspectral imaging to many farmers. We propose a vector-based method for green pepper segmentation. Our approach combined channel attention and the LSTM module to segment the green pepper. A conceptual diagram is shown in Fig. 3.3.

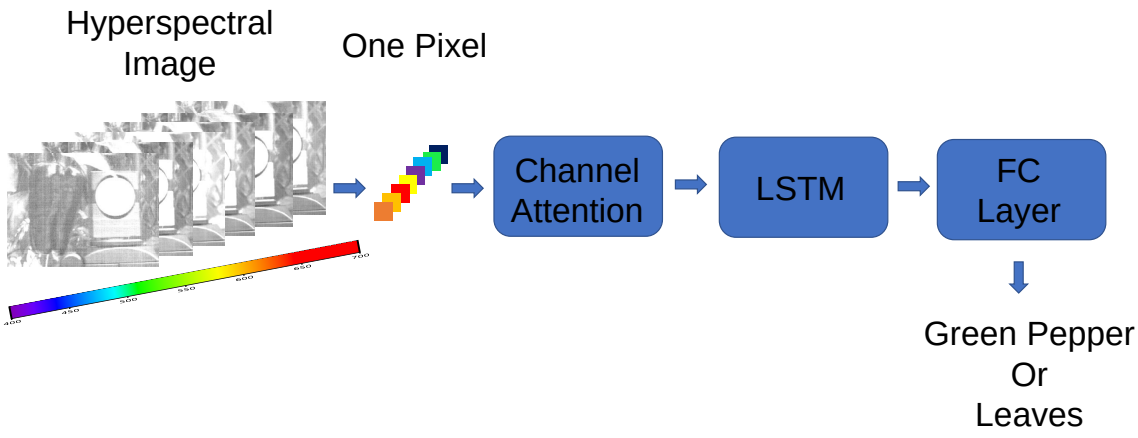


Figure 3.3: The above image illustrates the whole structure of our proposed method. We extra one pixel from the hyperspectral image as one vector. There are two central parts of our proposed method. The first part is the channel attention module. The second part is the LSTM with two fully connected layers.

3.2 Proposed Method

In this section, we introduce the details of our proposed method. Inspired by the Squeeze-and Excitation block in the SeNet [31] and deep recurrent neural network for hyperspectral image classification [56], we presented our channel attention module with LSTM. In order to deal with the issue of exploiting channel dependencies, we first consider reducing the dimension in the output feature space. Unlike the Squeeze-and -Excitation block, we don't need to consider squeezing global spatial information because our approach is vector-based. In figure 3.4, we illustrate the details about our channel attention module. The aim of the channel attention module is to capture channel-wise dependencies fully. To fulfil this objective, the proposed method must satisfy two patterns: first, it must be flexible (in particular, it must be capable of learning a nonlinear interaction between channels), and second, it must retain a non-mutually-exclusive relationship. To meet these criteria, we

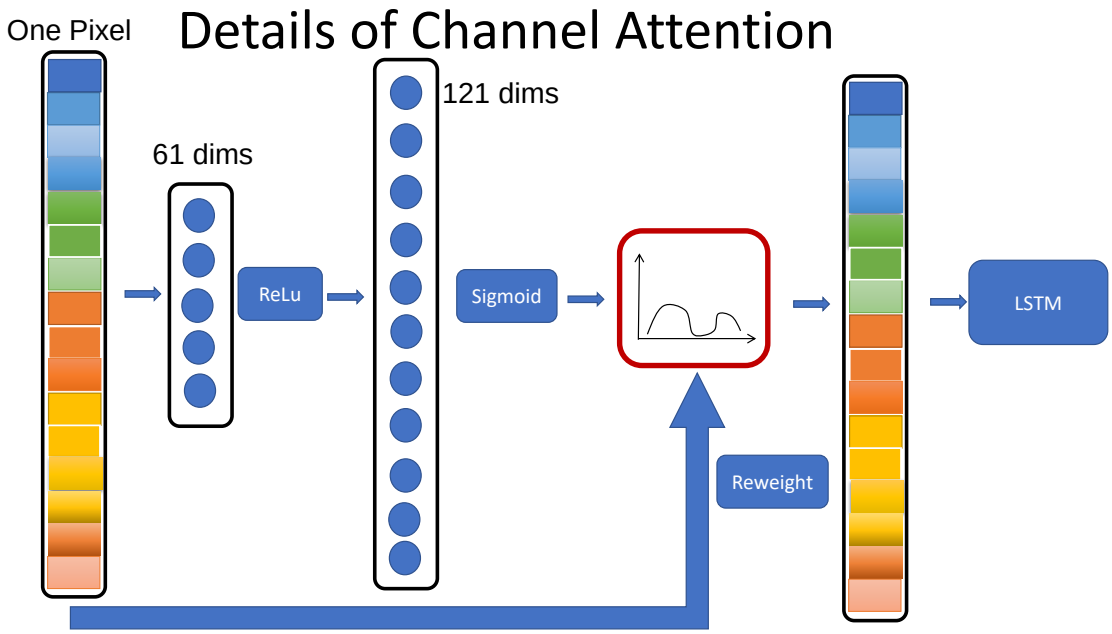


Figure 3.4: The above image shows the details about the channel attention module. Our channel attention module consists of two fully connected layers. The first layer is 61 dimensions, and the second layer is 121 dimensions. After the sigmoid function, we get the weight for each channel of the input vector. We use this weight to reweight the input vector.

opt to employ a simple gating mechanism with a sigmoid activation. It is straightforward to apply the channel attention block to the LSTM structure. The flexibility of the channel attention block means that it can be directly applied to transformations of the hyperspectral input vector. For the proposed channel attention block to be viable in practice, it must provide a useful trade-off between model complexity and performance, which is essential for scalability. We set up the dimension of the first layer to be 61. The final output of the channel attention block is obtained by rescaling the transformation output with the activation. The output of the channel attention block act as channel weights to adapt to the input-specific hyperspectral vector. In this regard, channel attention block intrinsically introduces dynamics conditioned on the input, helping to boost feature discriminability. The layer normalization [57] was utilized in our architecture to speed-up the whole training process.

Most of the existing classification techniques are based on spectral-spatial frameworks. They do not take advantage of the fact that the spectral information in the hyperspectral images is sequential in nature. A recurrent neural network (RNN) is an extension of traditional neural networks and is used to address the sequence learning problem. However, it turns out to be that training RNN models to model the long-term sequence data is difficult. To address this issue, Hochreiter and Schmidhuber proposed LSTM to replace the recurrent hidden node with a memory cell, which defeats the shortcomings of the previously built RNNs [39].

The LSTM structure consists of four essential parts: input gate i_t, output gate o_t, forget gate f_t , and candidate cell value c_t. The equation of the forward pass of an LSTM is followed by:

$$\begin{aligned} f_t &= \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ o_t &= \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c\left(W_c x_t + U_c h_{t-1} + b_c\right) \\ h_t &= o_t \circ \sigma_h\left(c_t\right) \end{aligned} \tag{3.1}$$

3.3 Experiments

Since there are no available hyperspectral pepper datasets on the internet, hyperspectral images of green peppers were taken by our own hyperspectral camera(NH-2 by EBA JAPAN CO., LTD.). Fig 3.5 demonstrates the data acquisition flow for our dataset. We collected our hyperspectral dataset in the Kochi Agriculture center four times from June to December 2019. Let $L(x, y, \lambda)$ be the radiance data index at spatial coordinates (x, y) of the scene at wavelength λ. We also define the spectral reflectance $R(x, y, \lambda)$ and the global

Figure 3.5: The above image demonstrates the data acquisition flow in Kochi Agriculture center for our research

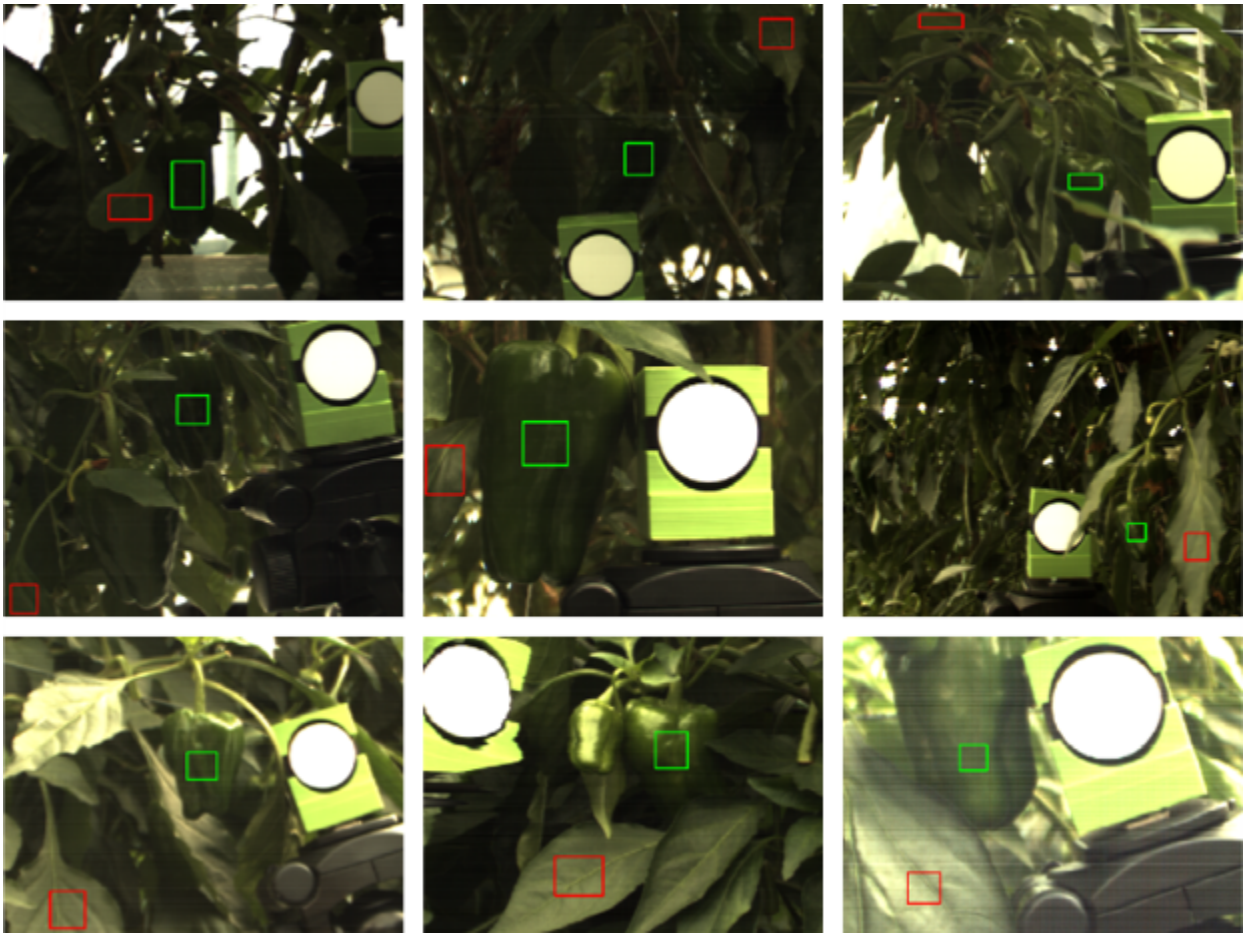


Figure 3.6: The above image shows how to get the pixel-wise data from our dataset. The image is in the raw-RGB color space. The green rectangle represents the pixel for green pepper. The red rectangle represents the pixel for green leave.

illumination $E(\lambda)$. The spectral reflectance of the image is then calculated as:

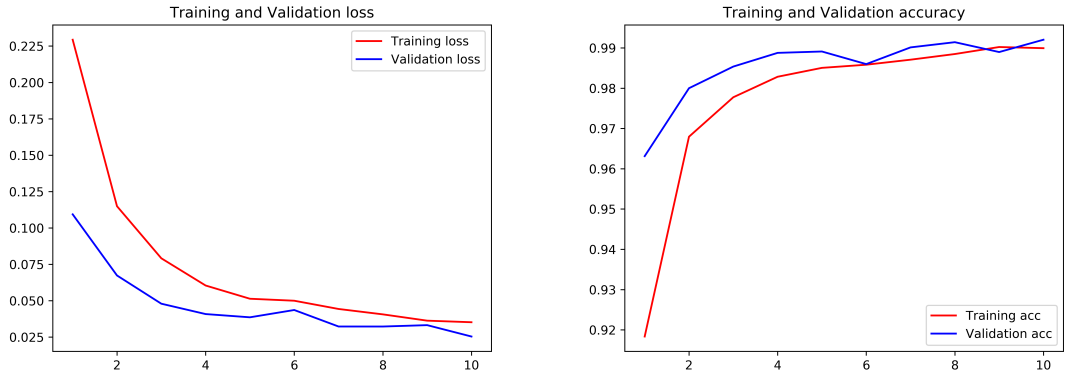
$$R(x, y, \lambda) = \frac{L(x, y, \lambda)}{E(\lambda)}. \tag{3.2}$$

We developed pixel-wise sampling software by using Python Package, such as OpenCV, to get the vector-based dataset. We illustrate the location of the bounding box for collecting pixel data in each sample image in Fig 3.6. The green rectangle represents the pixel for green pepper. The red rectangle represents the pixel for green leave. As reported in the previous research [58], there is no overlapping between each bounding box in our

Table 3.1: Comparison of our proposed method and ACPR pixelwise result.

Methods	accuracy	loss
Our proposed.	0.987	0.0387
ACPR. [1]	0.796	0.45

dataset to avoid an information leak. Because the camera has a spectral resolution of 10 nm potentially, the whole spectral range of 400–1000 nm was divided into 121 wavebands. We selected 120445 samples, including 50,376 positive samples(green peppers) and 42,681 negative samples(green leaves). In our training stage, we adopted adam [44] with a learning rate of 0.001 for optimization. We used ten epochs for training, and the batch size was set to 50. We implemented our model in Keras [59] with TensorFlow backend [60] and trained them by using an NVIDIA™ GeForce 1080Ti GPU. In figure 3.8, we show the segmentation result of the green pepper. As you can see, our proposed method can achieve good green pepper segmentation results.



(a) (b)

Figure 3.7: (a)Training and Validation loss, (b)Training and Validation accuracy.

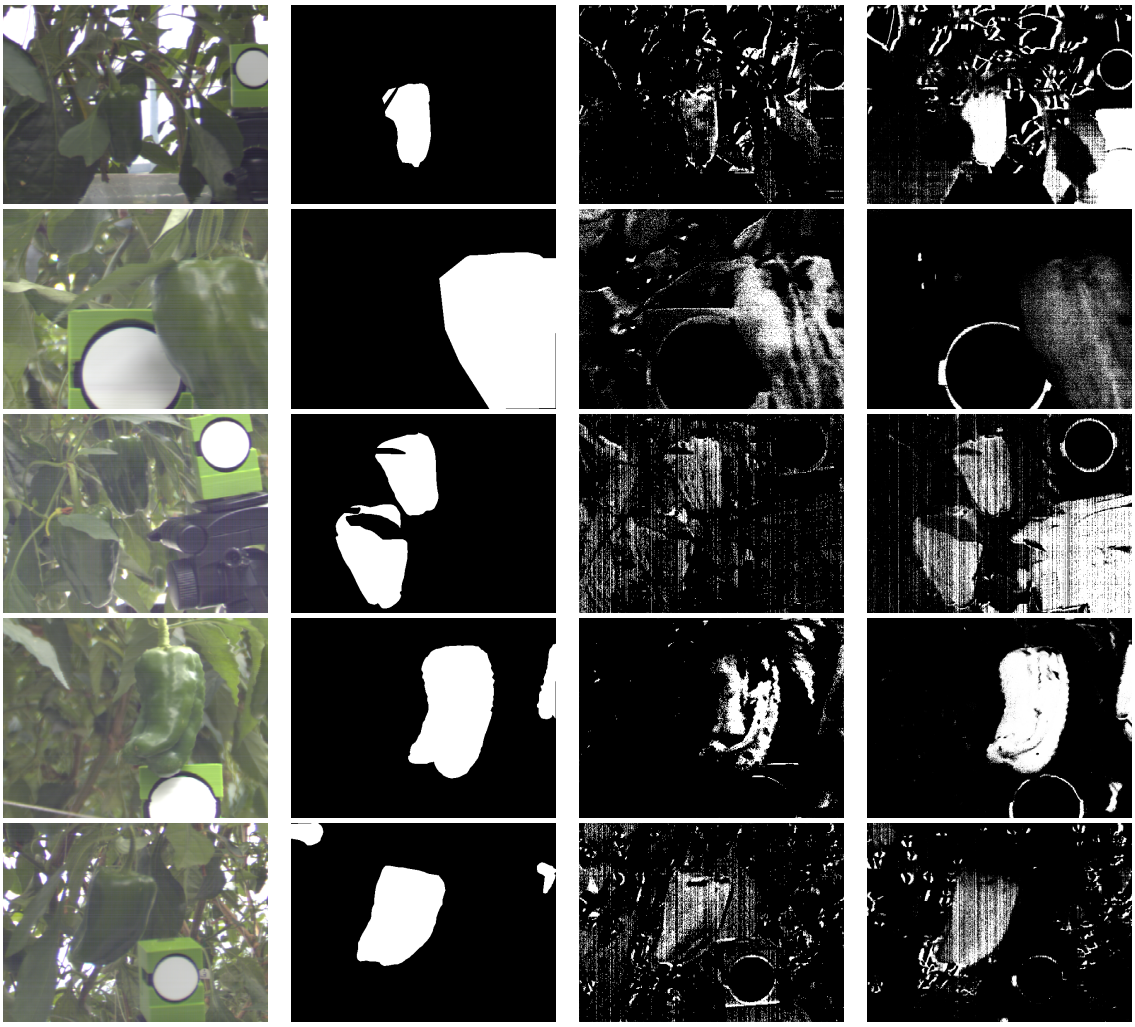


Figure 3.8: **Comprehensive segmentation results of our proposed method with our ACPR paper result.** The above image demonstrates the green pepper segmentation results. The first column is the sRGB image. The second column is the ground truth. The third column is the segmentation results from [1]. The fourth column is the segmentation results by using our proposed methods.

3.4 Conclusion

We have presented a vector-based model that achieves excellent green pepper segmentation problem performance, which can be widely applied in agricultural fields. This research aims not to object detection for green pepper counting but segmentation for automatic picking or growth prediction of green pepper. Our method consists of two parts. The first

part is the channel attention block, which generates the weights for each input channel. The second part is the LSTM. We treat the hyperspectral segmentation as the vector-based machine learning problems. In the future, we will extend our work to the spatial-spectral domain, which we believe can provide more information than the vector-based methods.

4 A Spectral-Aware RGB Camera Framework for Effective Green Pepper Segmentation

Detecting and identifying objects of similar color is a challenging task in computer vision. Green peppers in a natural environment can be found using the abundant information provided by a hyperspectral camera in the spectral domain, but the hyperspectral camera is an expensive device. Therefore, we propose a novel framework called Optical Filter Net, which enables the design of an optical filter that improves the performance of green pepper segmentation by a specific red-green-blue (RGB) camera system. When installed with the optical filter, the system can efficiently utilize the spectral information in the visible wavelength to distinguish green pepper and foliage without requiring an expensive hyperspectral camera. A main finding is the similarity between the transmission curve of the optical filter and the depth-wise convolution kernel without bias. Accordingly, we can treat the transmission curve of the optical filter as one layer of a deep neural network. The whole structure can be trained in an end-to-end manner. To comply with the physical requirement of the optical filter, we further constrain the training process to achieve a non-negative and smooth transmission curve. In an experimental evaluation on our dataset, our proposed spectral-aware RGB camera system outperformed the RGB camera system without an optical filter.

4.1 Introduction

Green pepper is one of the major vegetables found in Kochi Prefecture, which contributes to approximately 10% of the overall green pepper production in Japan. Hence, the production

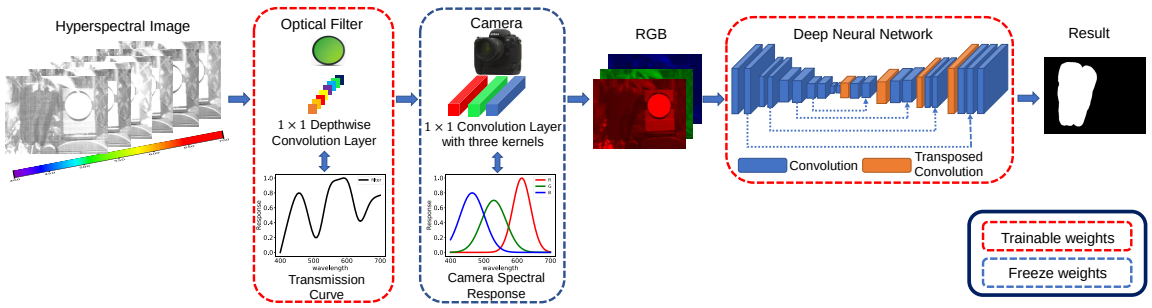


Figure 4.1: Pipeline of our work. Based on our hyperspectral dataset, an end-to-end network structure designs the transmission curve of the optical filter. Our structure consists of two parts: a depth-wise convolution layer and a traditional convolution layer (representing the spectral transmission curve of the optical filter and the camera spectral response, respectively), followed by a U-Net shaped structure for green pepper segmentation. During the training stage, the weight of the camera spectral response is fixed. At the bottom right, the blue and red dashed-line rectangles indicate the trainable framework parameters and the frozen weights representing the camera spectral functions, respectively. During the application stage, the optical filter is attached in front of the camera lens. Consequently, the optical filter changes the spectral distribution of the incident light by its transmission curve.

efficiency of green pepper must be improved using recent technology. Green pepper segmentation is useful to estimate the size and shape of green pepper, which has potential application in precision agriculture applications, e.g., automatic harvesting, green pepper growth estimation, harvest prediction, and marketing aid. However, to the human eye and in red-green-blue (RGB) images, green pepper is the same color as the foliage. Hyperspectral cameras can provide much more detailed clues than a consumer RGB camera, because a hyperspectral image records the interaction between the material surface and incident light at different wavelengths. Therefore, it captures the intrinsic physical and chemical properties of the imaged material. Hyperspectral imaging is a useful tool in agricultural applications [61], food industries [62], and scientific research [63], but hyperspectral cameras are expensive and not widely available to most farmers.

Under a particular illumination condition, two objects with different spectral reflectances may appear as the same color. This phenomenon is called metamerism [64]. Metamerism arises because the surface reflectance and scene illumination have more degrees of freedom

than trichromatic vision responses. Consequently, color provides insufficient evidence for identifying and detecting objects. To increase the accuracy of surface classification, Blasinski et al. [65] proposed to design the spectral power distributions of illumination. A critical component of a consumer RGB camera is the Bayer color filter array [66]. The camera spectral response (CSR) describes the sensitivity of the camera sensor with a Bayer color filter array to incident light of different wavelengths. The CSR, which relates the image intensity to the scene radiance, plays an essential role in computer vision tasks such as hyperspectral reconstruction [67], color constancy [68], and integrated signal processing inside digital cameras [69]. However, the main purpose of most consumer cameras is generating visually pleasing images, which does not require precise measurements of the scene radiance. Previous research [70] has shown that the CSR is not necessarily the best choice for specific computer vision tasks such as hyperspectral estimation. Moreover, the CSR varies notably among the different manufacturers and models of cameras [71]. Fu et al. [72] proposed a deep neural network based method to jointly select the optimal CSR and learn the hyperspectral image reconstruction network. Another major current focus is computational optics [73]. Jointly optimizing optical elements with differentiable end-to-end algorithms has generated considerable recent research interest, such as sensor design [74], automatic lens design [75] and HDR imaging [21].

Inspired by these works, we here introduce a spectral-aware RGB camera system that economically improves the accuracy of detecting green peppers in their natural environment. To meet our ultimate goal (optimizing the transmission curve for green pepper segmentation), we add an optical filter to a specific consumer digital camera. The transmission curve of the optical filter is designed by a deep neural network, which is driven by the task of green pepper detection. Figure 4.1 is an outline of our presented study.

The main contributions of our research are summarized below:

• Rather than redesigning or selecting the CSR, we design an optical filter that can be

added to any available camera owned by farmers and workers. The optical filter enables an automatic harvesting system in a cost-efficient way.

• We show that the transmission curve of the optical filter can be considered as a 1×1 depth-wise convolution kernel without bias. Following previous research, we treat the CSR of the RGB camera as fixed-weight 1×1 traditional convolution kernels without bias[76], and thereby represent our proposed physical camera model. The CSR layer is followed by a segmentation part, a namely, a U-Net like neural network with three input channels, for green pepper segmentation. The whole structure simultaneously learns the spectral transmission curve and optimizes the green pepper segmentation neural network in an end-to-end manner.

• Owing to the physical limitations of filter production, the transmission curve of an optical filter is non-negative and smooth. During the training stage, we impose a particular constraint on our deep neural network. We then show that our deep neural network can learn the spectral transmission curve in a non-negative and smooth space.

The remainder of this chapter is organized as follows. Section 4.2 briefly reviews the related work on CSR, deep neural networks, and object segmentation and detection. Section 4.3 introduces the details of our method, and Section 4.4 presents the experimental results and our green pepper dataset. Section 4.5 reports our fabricated optical filter and the whole imaging sensor system. The paper concludes with Section 4.6.

4.2 Related work

Agriculture, horticulture, and food industries continue to rely largely on the manual labor of farmers and workers. However, as labor costs increase and the population ages, automatic harvesting systems have gained attention in the computer vision and robotics field. Kitamura et al.[77] developed a picking and cutting robot system equipped with a parallel stereo vision system and a HALCON image processing application. In a follow-up

study, Eizentals et al.[47] proposed a novel three-dimensional (3D) pose estimation method for green pepper detection by their automatic harvesting robot. Their approach utilizes illumination by a light-emitting diode in hue-saturation-value color space for target recognition, and laser measurement with point segmentation for 3D pose estimation. Alternatively, automatic fruit segmentation has been achieved by feature learning algorithms with conditional random fields on multi-spectral images[78]. Although such algorithms achieve high performance on multi-spectral images, the method is uneconomic and insufficient for large-scale applications in practice. Our proposed method efficiently utilizes the spectral information for low-cost segmentation of green pepper among its foliage. A critical advantage of our strategy is its generalization, as RGB cameras are widely available.

Mimicking the visual attention of the human vision system, salient object detection aims to detect and segment the most perceivable objects from an image. This method is attracting widespread interest in computer vision fields such as object detection and image manipulation, and has been extensively improved by the rapid development of deep neural networks in segmentation tasks. A variety of methods and datasets have been proposed in this research area[79]. Recently, salient object detection has focused on new approaches using hyperspectral images. For example, Liang et al.[80] used the hyperspectral information to distinguish objects with similar appearance but constructed from different materials, which is a challenging problem in salient object detection. The HS-SOD dataset[81] was one of the first large-scale hyperspectral image datasets for evaluating salient object detection algorithms. Hyperspectral imaging achieves high spectral resolution, as it provides the full spectral reflectance information of the object. In a recent study, İmamoğlu et al.[82] proposed feature learning on hyperspectral images by unsupervised segmentation tasks.

Many computer vision algorithms require accurate measurements of the scene radiance in (for example) high dynamic range [83], photometric stereo [84] and shape from shading[85]. CSR links the scene radiance to image intensity. The CSR is most precisely estimated

using a monochromator and a spectrometer, but such measurements are time-consuming and expensive. Much research in recent years has focused on modeling and evaluating the CSR. In one significant study, Han et al.[86] estimated the CSR from the fluorescence signals of a single image. From a different angle, Jiang et al.[71] analyzed the CSR by a statistical method, and reported that principal component analysis achieves more accurate and robust results than other methods. Their results showed that the CSR can be estimated from a single image with known or unknown illumination.

Recently, Nie et al.[76], who analyzed the working process of the CSR, reported that the CSR is similar to a convolution layer, and can be optimized by leveraging robust deep neural networks. Examining their results in detail, it was found that the machine learning algorithm can automatically design the CSR for a specific task. Because CSR performs dimensional reduction in the spectral domain, most consumer cameras cannot provide the full spectral information. Kimachi et al.[87] proposed a three-phase quadrature spectral matching imager based on a correlation image sensor. Their technique offers a new way of utilizing spectral information without a hyperspectral camera. Unlike the methods mentioned above, an alternative approach was developed by selecting limited band information from the hyperspectral image and increasing features by morphological calculation. In the remote sensing area, Kwan et al.[88] proposed to utilize RGB-NIR bands and augment features by EMAP[89] which enlarges feature map considering morphological attribute for land cover classification and achieved better performance than using hyperspectral images.

Detecting objects in an image is one of the fundamental tasks in computer vision. Much research progress in recent years has focused on CNN-based detectors, such as RCNN[90], YOLO[91], and so on. Gan et al.[92] developed an immature green citrus detection approach based on faster RCNN using color and thermal cameras. Kwan et al.[93][94][95] firstly applied the YOLO algorithm on coded exposure cameras and achieved

good detection accuracy with low power consumption and small bandwidth. Object detection is out of focus for our own application but focusing on low power consumption is also important even in agricultural application.

The above studies highlight the need for developing computer-aided technology in agriculture, and emphasize the critical role of the CSR function in computer vision applications. However, few studies have considered the spectral information in a ready-made camera system. The present study explores the distinguishing ability of a transmission curve for similarly colored objects, such as green pepper and foliage. The proposed method is generalizable to the segmentation of other objects composed of materials with similar colors but different spectral reflectances.

4.3 Proposed Method

This section provides the details of our proposed method. We first describe the relationship between the 1×1 depth-wise convolution kernel and the optical filter. We then report the whole structure of our neural networks. Finally, we present the non-negative and smooth constraint of our neural network.

4.3.1 Optical filter simulation

Our proposed method is overviewed in Fig. 4.1. Our framework uniquely simulates the optical filter as a 2D depth-wise convolution kernel without bias. By this approach, we can explore the transmission curve of the optical filter through the deep neural network algorithm. Fig. 4.2 demonstrates the similarity between depth-wise convolution and the transmission curve of an optical filter.

The essential equipment in the imaging system is the optical filter, which selectively transmits the incident light at different wavelengths. Tensor factorization is a popular method for analyzing 3D hyperspectral data [96]. The captured scene is represented as

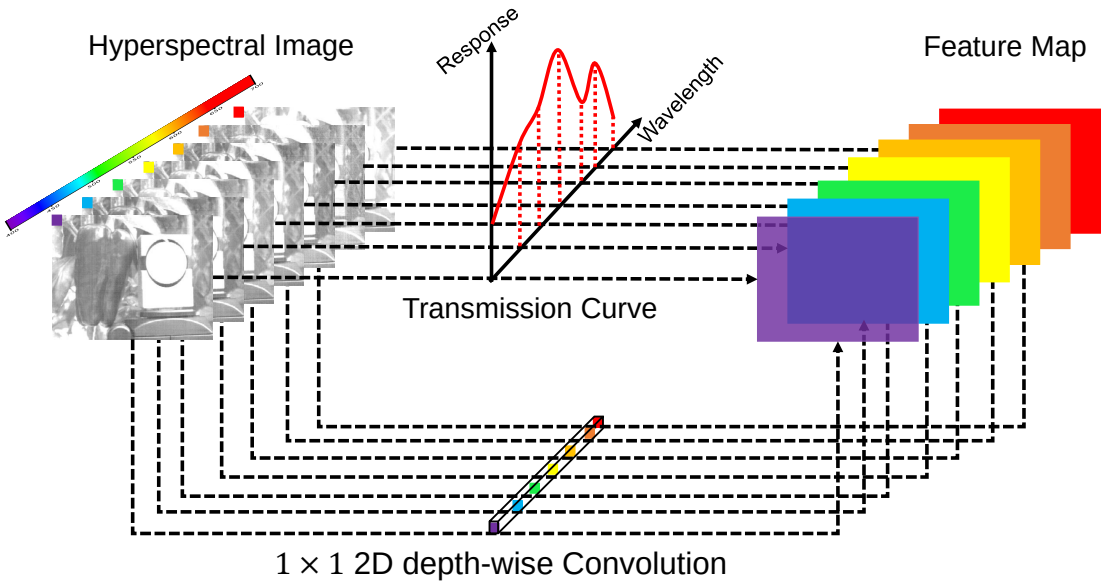


Figure 4.2: Similarity between the 1×1 depth-wise convolution and the spectral transmission curve of an optical filter. The spectral transmission curve of the optical filter can be represented by a non-biased 1×1 depth-wise convolution kernel.

a 3D tensor $L(x, y, \lambda)$, where λ denotes the wavelength, and x, y is the spatial position of the analyzed point. Given the radiance of the captured scene $L(x, y, \lambda)$, the workflow of the optical filter can be described by the following element-wise product:

$$S(x, y, \lambda) = L(x, y, \lambda) \circ T(\lambda), \tag{4.1}$$

where $1 \leq x \leq W$, $1 \leq y \leq H$, W and H represent the width and height, respectively, of the hyperspectral image in the spatial domain. $T(\lambda)$ represents the spectral transmission curve of the optical filter, and $S(x, y, \lambda)$ denotes the output of the optical filter. The elements of the column vector $T(\lambda)$ are the transmittance at different sampled wavelengths $[t(\lambda_1), t(\lambda_2), t(\lambda_3), \cdots, t(\lambda_n)]$. Depth-wise convolution was proposed for the Xception architecture [97], which uses depth-wise separable convolution operations (depth-wise and point-wise convolution) in its Inception module. Here, we chose a depth-wise convolution kernel because it executes independently over each channel of the input tensor. By employing

the depth-wise convolution kernel, we can treat the optical filter's transmission curve as one layer in the deep neural network, and design the most efficient transmission curve in an end-to-end manner. The transmission curve of an optical filter is usually assumed to be spatially uniform. Under this assumption, the spatial size of the depth-wise convolution kernel is 1×1, and the kernel weight is constant at different positions in the spatial domain.

Incident light passes through the optical filter and enters the camera equipped with a Bayer color filter array. Inspired by previous research [76], we simulated the camera's spectral response with a three-kernel 1×1 convolution layer. The three kernels represent the R, G, and B channels, respectively. The intensity of the pixel located at (x, y) in the k-th channel of the image is

$$P_k(x, y) = \sum_{i=1}^{N} C_k(\lambda_i) S(x, y, \lambda_i), \qquad k = R, G, B \tag{4.2}$$

where $C_k(\lambda_i)(i = 1, 2, \cdots, N)$ is the related spectral response function of the camera, and $S(x, y, \lambda_i)$ is the output tensor of the previous optical filter. As shown in Fig. 4.1, our deep neural network is appended with a 1×1 depth-wise convolution layer followed by a 1×1 convolution layer with three kernels. Following the above approach, our architecture extracts the spectral transmission curve of the optical filter from the optimized weight of the 1×1 depth-wise convolution kernel.

4.3.2 Network structure

The hyperspectral image passes through the one-kernel depth-wise convolution layer and the three-kernel convolution layer. After this passage, it is converted to three channel feature maps representing the three color intensities of the raw-RGB image. The feature maps from the two convolution layers are input to our green pepper segmentation module, which is configured as an encoder–decoder U-Net [98] structure with a skip connection design. This module provides more accurate segmentation results than other module designs

on a small training set. Moreover, the U-Net structure simultaneously captures the low-level features and high-level contexts. The skip connection is particularly useful for sharing the low-level details and features across the whole structure. Here we investigate the possibility of designing an optical filter by a deep neural network. Applying a more effective neural network structure in the green pepper segmentation module is left for future study.

In the encoder part of our U-Net structure, the basic module is a 2D convolution layer composed of a batch normalization layer [99] and a rectified linear unit activation function[100]. After spatial down-sampling by the max-pooling layer in the encoder part, the spatial dimension is further reduced to 20×20. The decoder part is nearly structurally symmetric to the encoder part, and upsampling is performed by a transposed convolution layer of stride 2. The transposed convolution layer is followed by basic feature-extraction modules in each block.

4.3.3 Constraint ensuring a non-negative and smooth function

The optical filter must satisfy certain physical requirements that constrain its spectral transmission curve. First, the weight of the spectral transmission curve must be non-negative. Second, as a smooth curve is more feasible than other arbitrary curves during actual implementation, the spectral transmission curve cannot be spiky and randomly fluctuating. Last but not least, the spectral transmission curve must be spatially invariant. To optimize the optical filter design, we trained the whole framework under the following loss-minimization formula:

$$\ell = \ell_{bce} + \eta \|\mathbf{GW}\|_2^2 \qquad s.t. \quad W \geq 0, \tag{4.3}$$

where ℓ_{bce} is the binary cross entropy, the most commonly used loss function in binary classification and binary image segmentation. The coefficients η control the smoothness of the spectral transmission curve, G denotes the first or second derivative matrix of the depth-wise convolution layer, and W denotes the shape of the transmission curve of the

Figure 4.3: The sample color images were rendered from our hyperspectral dataset and the corresponding ground truth of each image

optical filter. We impose a non-negative constraint $W \geq 0$ on the depth-wise convolution kernel in forward and backward propagation modes. To experimentally verify the effectiveness of the constraint in different settings, we varied η and controlled the smoothness of the spectral transmission curve by changing the derivative matrix. Therefore, the optimal weight of the optical filter can be obtained from the learned weight of the depth-wise convolution kernels W.

4.4 Experiment

4.4.1 Dataset and Setup

Thus far, the research community has lacked a public hyperspectral dataset for green pepper segmentation. To construct such a dataset, we collected hyperspectral images at the Kochi Agriculture Research Center, Kochi, Japan. Various types of hyperspectral cameras—wavelength-scan type, line-scan type, and snapshot type[101]— are available for scientific and industrial research. Our dataset was acquired by a line-scan type camera (Model NH-2-KTK, EBA Japan Co., Ltd, Japan). The NH-2 hyperspectral camera can provide 640×480 pixel images with 121 spectral bands ranging from 400 to 1000 nm. For compatibility with our target camera (a common RGB camera), we utilized only 61 spectral

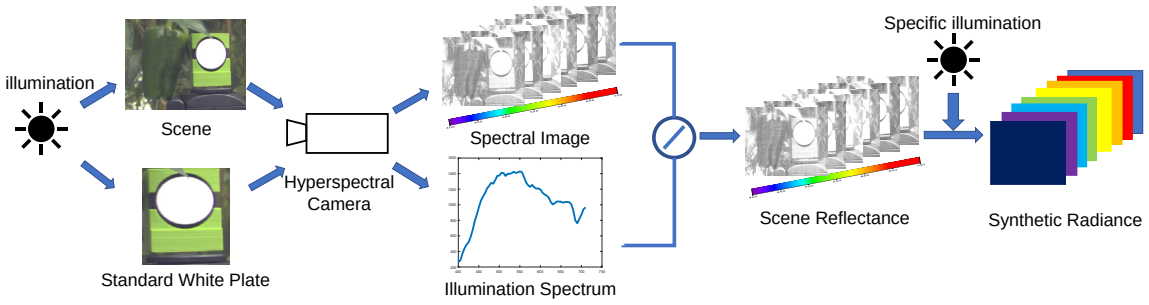


Figure 4.4: Pre-processing of hyperspectral images in our proposed method. After inserting a standard white plate over the captured scene, we take an image using the hyperspectral camera. Second, we divide the spectral image by the illumination spectrum to obtain the reflectance of the scene. Finally, we multiply the scene reflectance by specific illumination spectrum such as 6,500 K for daylight and 4,000 K for early evening.

bands ranging from 400 to 700 nm. The measurement resolution of our camera was 5 nm. When acquiring the images, we fixed the camera on a tripod to avoid motion distortions in the recorded data and adjusted the focus lens, the aperture of the camera. The recording time depended on the weather, and ranged from 30 seconds to 2 minutes. To obtain the reflectance data of the scene, we inserted a standard white plate in the field of view of the camera. Sample images are presented in Fig. 4.3.

Image capture by a hyperspectral camera is time-consuming and is often technically difficult, especially in outdoor scenes. The recorded radiance data are much more variable in outdoor environments than in laboratory settings, as natural illumination is not constant, but are greatly affected by weather conditions and seasons. The viewing geometry also alters the natural illumination. To construct our experimental dataset under different weather conditions and natural illumination conditions, we collected data at four times from June to December, 2019. Let $L(x, y, \lambda)$ be the radiance data index at spatial coordinates (x, y) of the scene at wavelength λ. We also define the spectral reflectance $R(x, y, \lambda)$ and the global illumination $E(\lambda)$. The global illumination $E(\lambda)$ is determined by averaging the spectral

information over the standard white plate:

$$E(\lambda) = \frac{1}{N}\sum_{n=1}^{N} W(n, \lambda), \tag{4.4}$$

where $W(n, \lambda)$ represents the spectral information recorded by measurement of the standard white plate, and N is the total number of pixels on the standard white plate. As shown in Fig. 4.4, the global illumination $E(\lambda)$ of each image is estimated by the measurement of reference plate reflectance. The spectral reflectance of the image is then calculated as

$$R(x, y, \lambda) = \frac{L(x, y, \lambda)}{E(\lambda)}. \tag{4.5}$$

To solve the different illuminations in our dataset,

we simulated the synthetic radiance data by a specific illumination at different color temperatures, e.g., 4,000 K, 6,500 K. The radiance data under a particular illumination were then constructed as follows:

$$L_{synthetic}(x, y, \lambda) = R(x, y, \lambda)E_{specific}(\lambda). \tag{4.6}$$

In experiments, we conduct our proposed method in two datasets. One has a 6500K illumination condition, and the other has different illumination conditions. We rendered the hyperspectral image in standard RGB color space and created the ground truth for green pepper segmentation using the annotation tool LabelMe [102].

4.4.2 Implementation Details

We trained our framework on the above-constructed dataset, which has 104 and 9 hyperspectral images for training and testing, respectively. Before the training step, we enlarged our training dataset by data augmentation (random crop and random horizontal flip

operations). To extract an image patch from the original image, we set the crop size to 320×320. All convolution layers (except the three-kernel convolution layer) were initialized by the He uniform method[103]. The spectral resolution of the CSR of Nikon D1X was 4 nm [104]. Figure 4.6 shows the CSR of Nikon D1X, simulated by Gaussian functions at 5 nm resolution.

The whole network was trained by an Adam optimizer[44], and the hyperparameters were set to their default values. The initial learning rate was 10^{-3}, the weight decay was 0, and beta was 0.9 or 0.999. A dynamic learning rate was assumed with a learning rate drop of 0.1, determined by monitoring the changing test loss. The batch size and total epochs were set to 32 and 60, respectively. All experiments were implemented in the deep learning tool PyTorch 1.14[45] run on our GPU server, which is equipped with an Intel® Xeon® Silver 4110 central processing unit @ 2.10 GHz, 512 GB DDR4 memory, and an NVIDIA™ Tesla V-100 graphics processing unit. The total time of our training process was approximately 10 hours.

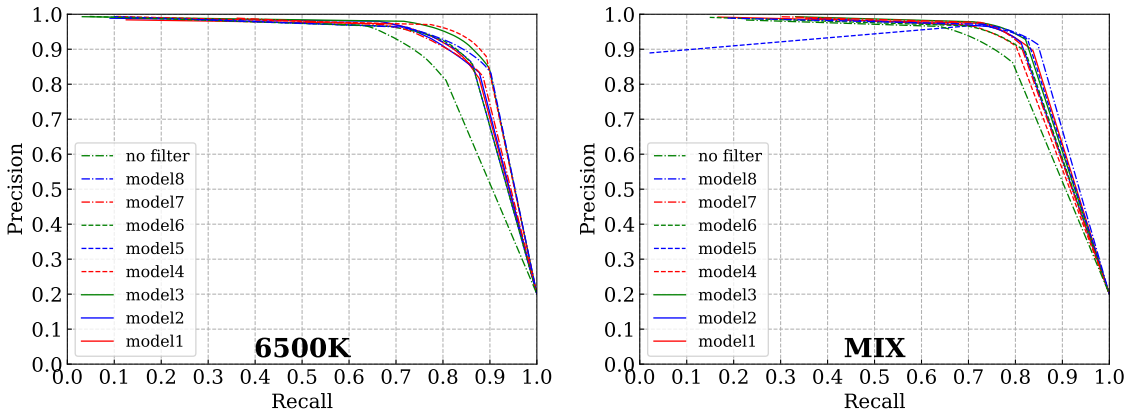


Figure 4.5: PR curve for comparing no-filter method and proposed methods under (a) 6,500 K dataset, (b) MIX dataset.

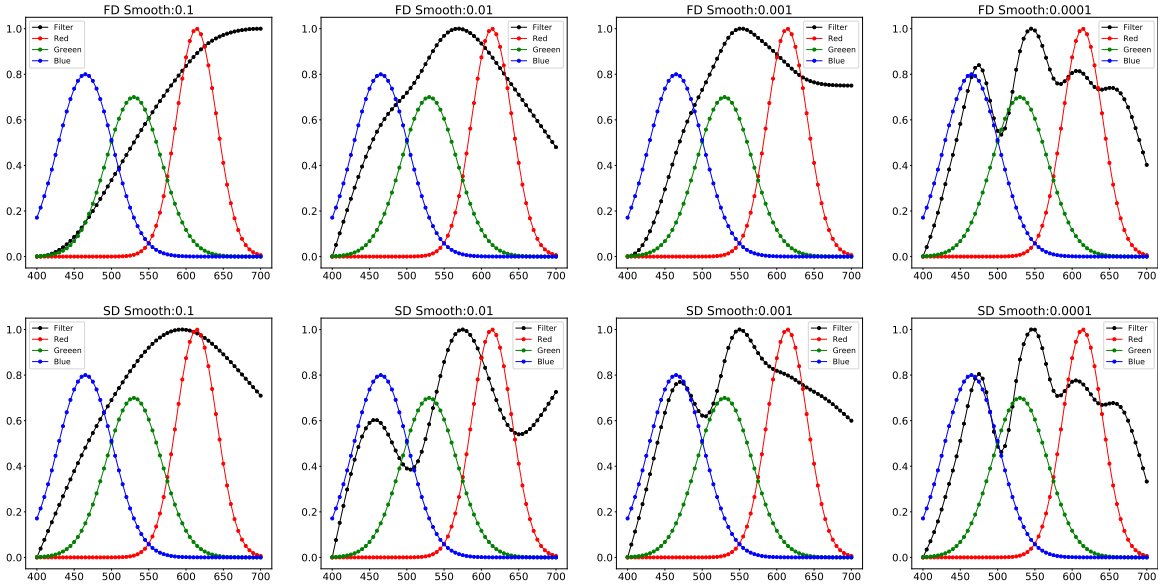


Figure 4.6: Learned spectral transmission curves under different settings in our proposed approach. The horizontal axis of each graph is the wavelength (in nm), and the vertical axis represents the normalized response. The black curves plot the learned spectral transmission responses of the optical filter. The R, G, B curves denote the spectral response functions of the camera to red, green, and blue light, respectively. FD: first derivative, SD: second derivative.

4.4.3 Results

Our aim was to design a data-driven spectral transmission curve of an optical filter for the effective segmentation of green peppers among foliage. This section presents the learned spectral transmission curves of the optical filter and the green pepper segmentation results.

Evaluation metrics

To qualitatively evaluate our proposed method, we adopted three common evaluation metrics: the precision-recall (PR) curve, F-measure, and mean absolute error (MAE). The MAE is computed by the average pixel-wise dissimilarity between the predicted object

Table 4.1: Ablation study of our model with different parameter settings and architectures by using two types of dataset.

Configuration	6500K		MIX	
	MAE↓	F_1↑	MAE↓	F_1↑
model1(FD η=0.1)	0.049	0.879	0.047	0.895
model2(FD η=0.01)	0.051	0.878	0.046	0.898
model3(FD η=0.001)	0.043	0.896	0.047	0.905
model4(FD η=0.0001)	0.040	0.906	0.051	0.888
model5(SD η=0.1)	0.048	0.884	0.048	0.900
model6(SD η=0.01)	0.048	0.885	0.050	0.886
model7(SD η=0.001)	0.051	0.877	0.046	0.899
model8(SD $\eta = 0.0001$)	0.047	0.885	0.045	0.904
no-filter	0.061	0.855	0.057	0.862

segmentation map and the ground truth.

$$MAE = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} |P(h,w) - G(h,w)|, \tag{4.7}$$

where $P(h,w)$ and $G(h,w)$ denote the predicted results and ground truth, respectively. The F-measure is determined from the precision (proportion of relevant instances among all obtained instances) and recall (proportion of obtained relevant instances among all relevant instances) as follows:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}. \tag{4.8}$$

Ablation Study on illumination

At the real application stage, the illumination condition in the captured scene largely varies[105]. To effectively verify our proposed method in the different illumination conditions, we created a synthetic mixed illumination dataset (MIX dataset) with different spectral emission of black body radiators at different temperatures: under early evening (4,000 K), daylight (6,500 K), and unknown illumination from our original dataset. The selected illumination provides adequate approximations to light conditions in real natural environments.

As presented in Table 4.1, under different illumination conditions, the MAE and

Figure 4.7: Comprehensive segmentation results of our proposed method under different parameter settings. For each test image, we show the segmentation results in three different color temperatures in our MIX dataset

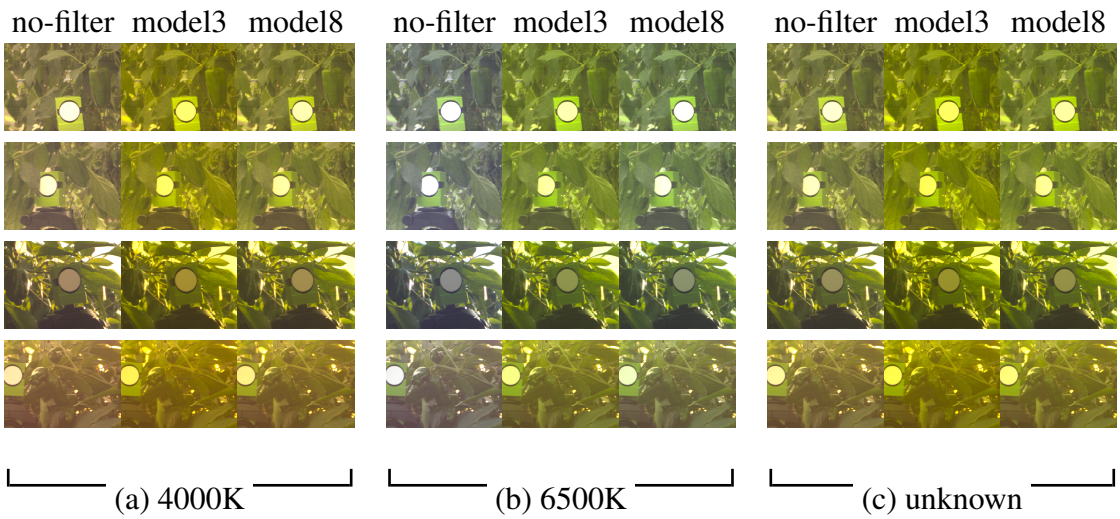


Figure 4.8: Example Rendered sRGB image under different illumination conditions. Each row shows different test images. Best viewed in color.

F-measure were lower and higher, respectively, in our proposed method than in the no-filter method, confirming that our filter improved the model performance. Besides, MIX dataset results show better performance in the most experimental configuration than the only 6,500 K dataset. These improvements in the MIX dataset could be attributed to:

1) The MIX dataset is a relatively larger dataset than the 6,500 K dataset.

2) The MIX dataset contains a variety of illumination conditions. In particular, the non-uniformity of lighting conditions in the scenes, e.g., the upper and lower sides of the peppers or shading, affects the lighting conditions at each part of the scene. Therefore, it is inferred the MIX dataset compensates for such illumination variations and improves the performance of our proposed methods.

Our proposed method can achieve better results in both datasets than those of the no-filter setting, showing the importance of spectral selection. To prove the effectiveness of our proposed method, we also plot the (precision, recall) pairs of the nine models in Fig. 4.5. The PR curve is a traditional machine learning measure for imbalanced data.

Compared with the no-filter setting, all of the proposed methods show better results. Evidently, the learned optical filter improved the green pepper segmentation ability.

Under the non-negative and smooth constraint described in the previous section, we can obtain the spectral transmission curve of our optical filter that satisfies the physical requirement. In our loss function, the parameter η controls the smoothness, and G is the first- or second-derivative matrix of the depth-wise convolution kernel. Figure 4.6 compares the model results under different settings of η and G in the MIX dataset.

Obviously, different η and derivative matrices yield different spectral transmission curves. According to the curves in Fig. 4.6, spectral transmission curves in the blue channel are lower or keep similar to the CSR in the different settings. It turns out to be the wavelength from 400 nm to 450 nm is not crucial for identifying the green pepper and foliage. When the η is 0.1 in both FD and SD, the shape of the spectral transmission curve has proven to satisfy the smooth constraint with less concern for critical wavelength because of large η. Expect for the setting FD $\eta = 0.1$, the maximum response for most spectral transmission curve is around 550 nm. Interestingly, however, both FD and SD for the $\eta = 0.0001$ can yield a similar spectral transmission curve, and there are four similar local maxima. Empirically, it is suggested that with low smoothness constraint, the spectral transmission curve has selected maximum local responses across wavelengths to generate blue, green, and red information. Apparently, according to these graphs, it can be inferred that the most critical wavelength is around 550 nm. And there are also other essential wavelengths around 460 nm, 620 nm, and 660 nm. However, it seems that wavelength near 500 nm makes a little contribution to distinguish green pepper and foliage.

Figure 4.7 compares the pepper segmentation results of test input by our proposed method under different parameter settings. The results of the no-filter system are also presented. In our experiments, the no-filter system describes the system with the naked RGB camera (no optical filter). Comparing each column in Fig. 4.7, we observe that our

proposed method significantly improved the segmentation result of most test images. The result implies that along with the spatial domain, the spectral domain is vitally important for green pepper detection. Unlike the traditional RGB camera, our proposed system can effectively and fully utilize the spectral information at visible-light wavelengths. We inferred that the optical filter captures the prior knowledge of green pepper segmentation in the spectral domain. The present research suggests that combining the spatial and spectral information is an economical approach for detecting and identifying objects with similar colors. To illustrate color imaging by our proposed method, we rendered the simulated RGB images of the no-filter method and the optical filter method under different settings of η. The images are presented in Fig. 4.8. Each row in Fig. 4.8 represents the different test images in our dataset. As illustrated in Fig. 4.8, the optical filter changes the image's color information compared with the no-filter setting. And combining the results of Fig. 4.7 and Fig. 4.8, the green pepper segmentation results are improved by such modification of the test images.

4.5 Realization of the Designed Optical Filter

This section exhibits the proposed spectral-aware RGB camera system with a proof-of-concept prototype by fabricating an optical filter that can be attached as an add-on device to a camera lens. As demonstrated in Fig. 4.9, we asked Chroma Technology Corp. to implement our designed Optical filter for the Lucid Vision Labs TRI050S-CC camera [106]. Our designed optical filter can be easily used for the camera. We installed our optical filter in front of the camera lens. Here, we chose FUJINON HF12XA-5M F1.6/12mm C-mount fixed focal Len for the TRI050S-CC camera.

Fig. 4.10 shows the transmittance curve of our implemented optical filter. The optic manufacturer try their best to closely our design transmittance curve. As illustrated in Fig. 4.11, We took photo with optical filter and without optical filter for the color check

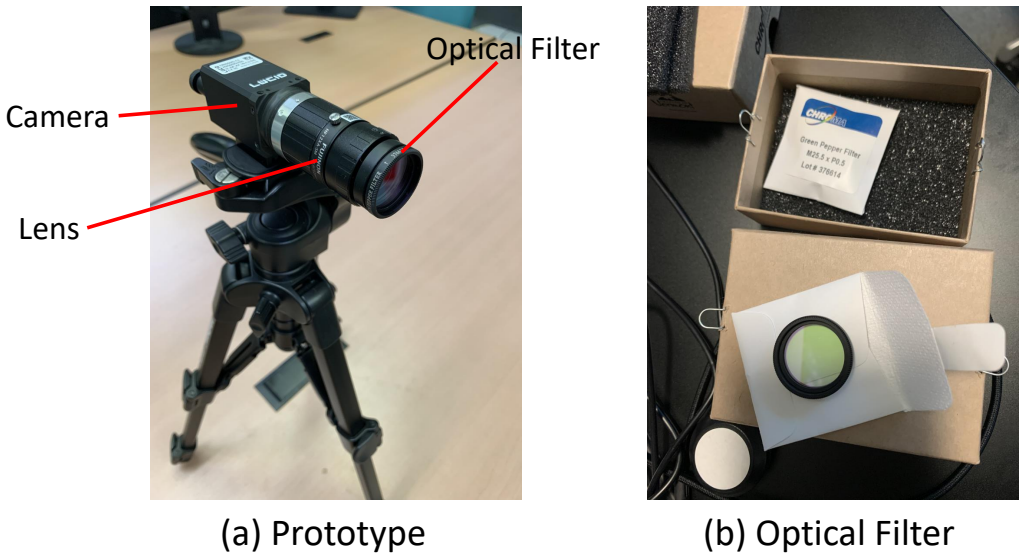


Figure 4.9: (a)Lucid Vision Labs Camera, (b) Manufactured Optical Filter.

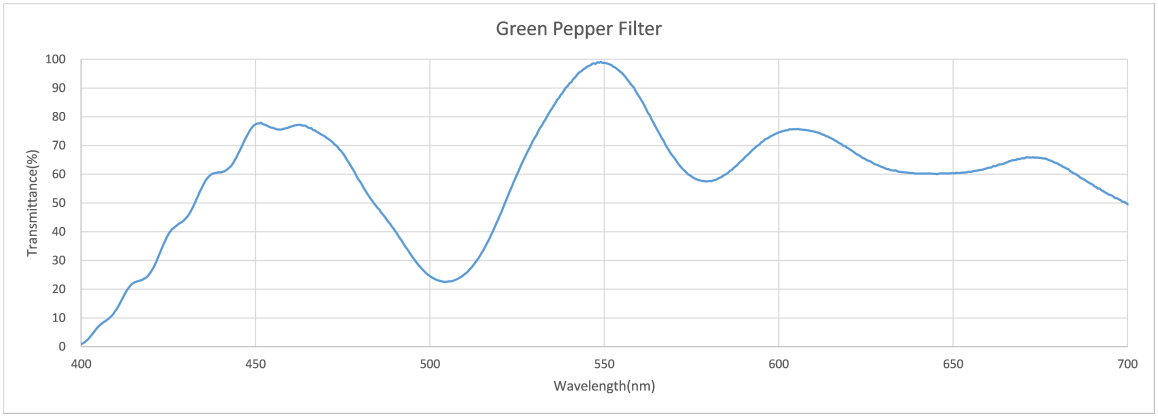


Figure 4.10: The transmittance curve of our implemented Optical Filter

broad.

(a) (b)

Figure 4.11: (a)With Optical Filter, (b)Without Optical Filter.

4.6 Conclusions

We proposed a spectral-aware RGB camera system that enhances the performance of green pepper segmentation by installing an optical filter on the RGB camera. The spectral transmission curve of the optical filter was designed by an end-to-end deep neural network. Our framework explicitly simulates the transmission curve of the optical filter and the CSR of the RGB camera by assigning trainable depth-wise convolution weights and freeze convolution weights, respectively. Our system exploits the full spectral information in RGB camera images for object identification in a straightforward manner, and provides a data-driven approach for optical device design. This study improves our prior work[1] by composing an end-to-end structure with a non-negative and smooth constraint for optical filter design and enlarging dataset under different illumination conditions. Our proposed system is technically feasible and demonstrates the benefits of a spectral-aware system for identifying near-color objects.

The most significant benefit for agricultural applications is that we improved the green pepper segmentation by adding an optical filter to an RGB camera, negating the need for expensive and specialized equipment. To our knowledge, this solution has not

been previously reported. Replacing the expensive hyperspectral camera with a cheaper camera-filter system, we can greatly reduce the cost of automatic harvesting systems for farmers. Although the deep neural network can design the transmission curve of the optical filter, the actual optical filter has not been demonstrated. In a follow-up study, we will evaluate the physical implementation of the optical filter and its performance in green pepper segmentation.

5 Color-Ratio Maps Enhanced Optical Filter Design and its application

5.1 Introduction

Improving the quality and production efficiency of the economic crop while aiding the management and marketing strategy is one of the critical aims of precision agriculture. Precision agriculture can provide useful information in the early stage to enable better decisions to make on the management system. In recent years, computer vision and artificial intelligence technology have developed to meet the growing demand for fast and accurate grain crop production [7] [8]. As reviewed by a previous study [9], machine learning techniques have been widely used for the early and precise detection of biotic stress in the crop, specifically for the detection of weeds, plant diseases, and insect pests.

Green pepper is one of the chief crops in Kochi Prefecture, which contributes to approximately 11% of the total production in Japan. Therefore, there is a significant need for using the latest precision agricultural technology to improve the production efficiency of green pepper. Developing automated green pepper harvest and growth prediction technology is essential for farmers to enhance their carriage efficiency and aid their marketing strategies. However, due to the same color of green pepper and its leaves, there remains a need for developing robust methods to recognize and segment green pepper. Recently, a new sensor system [107] for the detection and localization of green pepper has been proposed by utilizing multiple camera positions and viewing angles. Li et al. [108] proposed a novel pose estimation algorithm for sweet pepper.

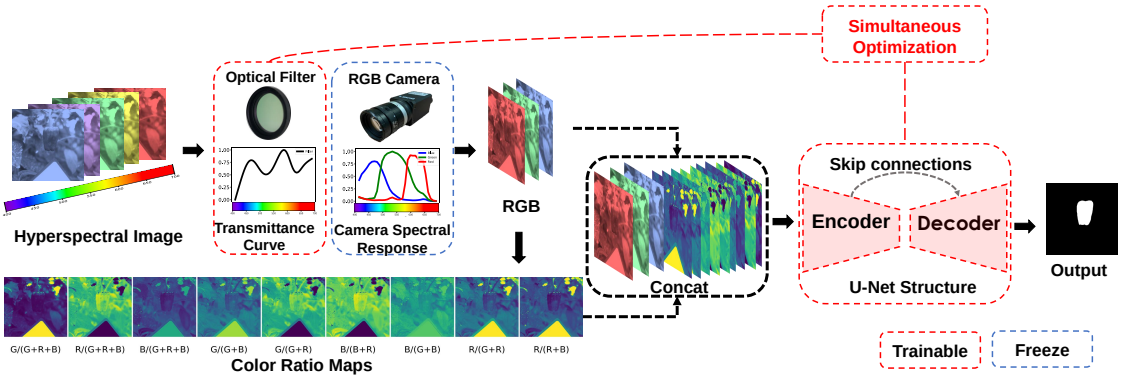


Figure 5.1: Our proposed computational optics framework incorporates both optics and image segmentation algorithm designs. Rather than optimizing these two parts separately and sequentially, the whole framework was treated as one neural network and establish a simultaneous end-to-end optimization framework. Explicitly, the first layer of the network corresponds to the physical optical filter, the second layer of the network is related to RGB camera spectral response, and all subsequent layers represent the segmentation algorithm. Inspired by previous research, instead of generating red-green-blue(RGB) three channels for the segmentation module, we augment the RGB three channels by color-ratio maps to exploit useful spectral information for green pepper segmentation. All the parameters of the framework are optimized based on segmentation loss on our hyperspectral dataset. Once the transmittance curve is optimized, we can fabricate the corresponding optical filter using multilayer thin-film technology. The fabricated optical filter is mounted in front of the camera lens, and the optimized segmentation network is integrated with the whole system.

Instead of independently optimizing the optical device and relevant image segmentation algorithm, we proposed the optical filter designing method for segmentation neural networks whose input is enhanced by color ratio maps. The transmittance curve(TR curve) of the optical filter can be treated as the weight of the neural network, and we can simultaneously optimize both an optical element and green pepper segmentation module by back-propagation. We illustrate the overview of our proposed method in Figure 5.1. Recently, Yu et al. [109] proposed an end-to-end deep learning optimization algorithm to search the optimal TR curve of an optical filter in the smooth and non-negative space. However, their method didn't fully utilize all the color-ratio maps from the R, G, B channels captured by the RGB camera. Historically, agricultural studies investigating color ratios and their linear combinations have shown the effectiveness for fruits and vegetables distinguish [110].

In this study, we enhanced the method by adopting the color-ratio maps as input for segmentation neural network. Color is one of important clue of object surface property. The benefit of the color-ratio maps is that it can help us to retrieve the adequate ratio of three chromatics in chromaticity space to derive the optimal TR curve for a specific CSR. In our segmentation module, a U-Net-like structure network [98] is utilized for extracting the spatial features of the RGB images captured by the optimal TR curve of our designed optical filter. After optimization, the designed optical filter can be implemented by optical technology and is attached in front of the camera lens. The spectral property of the incident light is changed by our designed optical filter.

The main contributions of our study conclude as follows:

- We developed the computational optics framework for co-design an optical filter and segmentation algorithm that can achieve a better image sensor system for green pepper segmentation. The whole framework simultaneously optimize the front-end optical device(optical filter) and the back-end green pepper segmentation algorithm.
- We introduced the color-ratio maps as additional input feature maps to improve the green pepper segmentation results. The experimental results demonstrate the benefits of the improved performance by color-ratio maps.

The rest of this chapter is organized as follows. Section 5.2 presents the research works related to our work. Section 5.3 presents the details of our proposed methods. Section 5.4 describes our green pepper dataset and experimental results. Lastly, Section 5.5 concludes our presented work and our future work.

5.2 Related Work

5.2.1 Color space

Color space is the fundamental research topic in colour image processing and has various computer vision applications. One of the major currents focuses in the advanced driving assistant system is to find appropriate color space for the detection of the traffic light. In their study, various color spaces were applied for their deep learning model, and the experimental results showed the RGB and normalized RGB color spaces [111] achieved the best performance. In an earlier study, Kondo et al. [110] established to utilize the color ratio map to search the most suitable wavelength to distinguish fruits and leaves. In precision agriculture, Zhao et al. [112] proposed to use an adaptive RB chromatic aberration map(ARB) based on OHTA color space [113] and the sum of absolute transformed difference feature in RGB camera to detect immature green citrus. Recently, a novel global image enhancement method, Neural Curve layers [114], was developed by exploiting global image adjustment curves in three different colors spaces, e.g., CIELLab, HSV, RGB.

5.2.2 Application of optical filter

The color filter array(CFA) or multispectral filter array [115] plays an essential role in acquiring the color information or spectral information in the RGB camera and multispectral camera. One of the early and intuitive studies of the optical filter is filter-wheel camera [116]. A series of special optical filters are installed in the rotating filter wheel, where each optical filter can be placed in the optical path of a monochrome camera by rotating the filter wheel. A complete multispectral image is constructed by multiple exposures for different optical filters at a time. Inspired by the CFA in the RGB camera, a multispectral filter array approach was proposed in both academic and industrial areas [117]. Lapray et

al. [118] reported a detailed study of the snapshot multispectral imaging and the analysis of spectral filter array. In the real application, Nakauchi [119] proposed a data-driven selection algorithm of a set of bandpass optical filters for ice detection by using hyperspectral imaging. They implemented their proposed optical filter by installing two bandpass filters with a near-infrared camera. Another important application of spectral optical filter array is skin oxygenation measurement for medical monitor and diagnosis [120]. Recently, Ono proposed an innovative multi-spectral imaging system using a polarization camera that captures nine bands at once [121].

5.2.3 Computational optics

Computational optics, which can be interpreted as jointly optimization optics elements (i.e. Bayer color filter, lenses and optical filters), image processing, and computer vision task, have generated considerable research interest [122] [19]. Chang et al. [20] proposed the end-to-end optimization paradigm by combining a differentiable optical image formation layer and a depth estimation network for jointly optimizing both camera lens and neural network weights. Inspired by the recent deep optics approach, A.Metzler et al. [73] developed an end-to-end method to jointly optimize the point spread function of the custom diffractive optical element(DOE) and the deep neural network for High-dynamic-range imaging. Nie et al. [76] reported the relationship between the 1×1 convolution operation and the camera spectral response(CSR) function. They developed a data-driven method to design a camera spectral filter array for hyperspectral reconstruction. Zou et al. [123] proposed the CSR-Net, which can effectively design the optimal CSR to achieve high classification accuracy with limited image bands. A mathematical approach [124] to improve the color measurement of the camera was developed by designing the spectral sensitivity of an optical filter. Their study demonstrated a numerical computation way of optical filter design based on both the Luther condition and the Vora-Value.

5.3 Proposed Method

In this section, we elaborate on our proposed method. We first introduce the filtered RGB camera module. Then, we report the green pepper segmentation module. Lastly, we describe the loss function and physical constraint.

5.3.1 Filtered RGB camera module

As illustrated in Figure 5.1, our proposed filtered RGB camera module is consists of two major parts: 1) A differentiable optical filter layer, whose trained weight is the transmittance curve of the optical filter, that can take in as input radiance and output a modified spectral radiance; 2) The frozen weights of a convolutional layer with three filters represent the camera spectral response function of the Bayer color filter array.

Optical filter layer

As the same with the photographic filter(e.g., UV filter, ND filter), the designed optical filter is mounted directly on the camera lens. Hence, the spectral information of the incident light at different wavelength is selectively filtered by the TR curve of the optical filter. We can describe the wavelength-wise product as:

$$L(x, y, \lambda) = R(x, y, \lambda) \circ T(\lambda), \tag{5.1}$$

where the $R(x, y, \lambda)$ denotes the radiance data in the captured scenes, $T(\lambda)$ represents the transmittance curve of the optical filter, and the $L(x, y, \lambda)$ represents the output radiance data of the designed optical filter, respectively. The range of x and y is $1 \leq x \leq W$, $1 \leq y \leq H$, W and H represent the width and height, respectively, of the captured image in the spatial domain. According to the equation (5.1), we found the similarity between the depth-wise convolution layer without bias and the TR curve of the optical filter. The

depth-wise convolution layer was proposed in the Xeption network structure [97], which the purpose is to reduce the computation resources. By utilizing the depth-wise convolution layer without bias, the TR curve of the optical filter can be regarded as one layer of the whole neural network structure. One feature of the TR curve is the spatially invariant, i.e., it only works in the spectral dimension and keeps the same transmittance across the spatial dimension in the captured radiance $R(x, y, \lambda)$. Due to the above feature of the TR curve, we chose the 1×1 as the kernel size of the depth-wise convolution kernel. Each weight in the depth-wise convolution kernel only works on the corresponding wavelength, which selectively transmits the input incident light. Also, the filter keeps the same weights across all the spatial domains.

CSR layer

Considered the output radiance data $L(x, y, \lambda)$ at position (x, y), the captured intensity by a fabricated image sensor equipped with CFA is calculated by

$$P_k(x, y) = \int_{\lambda} C_k(\lambda) L(x, y, \lambda) d\lambda, \qquad k = R, G, B \tag{5.2}$$

where λ denotes the wavelength, and C_k is the corresponded CSR function of the CFA, where k denotes the red, green, blue channels . $P_k(x, y)$ is the pixel intensity of the captured scenes. Essentially, we can discretely formulate the above equation by this equation

$$P_k(x, y) = \sum_{i=1}^{N} C_k(\lambda_i) L(x, y, \lambda_i), \qquad k = R, G, B \tag{5.3}$$

where the CSR function is represented by the vector form of $C_k(\lambda_i) = (C(\lambda_1), C(\lambda_2), C(\lambda_3), \cdots, C(\lambda_N))$ at different sampled wavelength, and N represents the total number of the spectral bands. As reported by previous research [76], the CSR function can be represented by three kernels with weights of 1×1 convolutional layer. Consequently, the $P_k(x, y)$ can

be calculated by the feature map generated by the 1×1 convolution layer with three kernels. In our approach, the simulated image pixel intensity is determined by three factors, i.e., the TR curve of the optical filter, the CSR function of specific CFA and the radiance of the captured scenes. To account for specular highlights and dark current in the simulated RGB image, we normalized the simulated sensor image as follows equation(5.4). The actual values for min and max are determined from the training dataset. We add a small number ϵ to avoid the division by zero in the color-ratio maps introduced in the following subsection. In our experiment, through trial and error, we set the $\epsilon = 0.01$.

$$P_k(x, y) = \frac{P_k(x, y) - \min}{\max - \min} + \epsilon \qquad k = R, G, B \tag{5.4}$$

We presume the camera has a linear response function, which commonly clips the simulated image sensor RGB value to emulate sensor saturation by the following equation.

$$f(c) = \begin{cases} 0, & \text{if } c < 0, \\ c, & \text{if } 0 \leq c \leq 1, \\ 1, & \text{if } c > 1. \end{cases} \tag{5.5}$$

5.3.2 Color-ratio maps

Unlike the previous research [109] to utilize only generated RGB images, we augment the simulated RGB image by combing different color-ratio maps. The simulated RGB sensor images are determined by the three main chromatics, red(R), green(G), and blue(B). Inspired by previous research that the color component ratio could help to distinguish fruits and leaves [110], we utilize the color-ratio maps as additional color cues to help the whole framework search the optimal TR curve of the optical filter. To solve the numerous

illumination condition in the green pepper grove, we utilized the normalized RGB color maps [125] in our augment color-ratio map. The normalized RGB color-ratio maps are expressed as d_1, d_2, and d_3, in later. They can be computed by the following equations:

$$\begin{aligned} d_1 &= \frac{G}{R+G+B} \\ d_2 &= \frac{R}{R+G+B} \\ d_3 &= \frac{B}{R+G+B} \end{aligned} \tag{5.6}$$

To efficiently derive the adequate transmittance curve of the optical filter, we applied multiple color-ratio maps. Our proposed multiple color-ratio maps are shown in the equation(5.7), respectively.

$$\begin{cases} d_4 = \frac{G}{(G+B)} \\ d_5 = \frac{G}{(G+R)} \\ d_6 = \frac{B}{(B+R)} \\ d_7 = \frac{B}{(G+B)} \\ d_8 = \frac{R}{(G+R)} \\ d_9 = \frac{R}{(R+B)} \end{cases} \tag{5.7}$$

We extract features from multiple input(RGB+color-ratio maps) to exploit the valuable information of different color ratio. Specifically, we concatenate the simulated RGB sensor image and their color-ratio maps as a tensor. We send it to the segmentation module that takes the RGB sensor image with the color-ratio maps as input to generate the final segmentation result.

5.3.3 Segmentation module

For green pepper segmentation, we attach the segmentation module to our filtered RGB camera module. Note that the principal goal of our research is not to propose the state-of-the-art neural network structure for green pepper segmentation, but to explore the relative benefit of color-ratio maps enhancement and co-design optical filter with segmentation module. In particular, we adopt the U-Net-like structure [98] in this work because it is commonly used for pixel-wise estimation(e.g., image segmentation, image-to-image translation) and great generalization performance on various tasks.

Table 5.1 summarizes the overall structure of the segmentation module. Followed by the filtered RGB camera module, the segmentation module accepts tensors of size $H \times W \times 12$ and lastly yields the corresponding green pepper segmentation results of size $H \times W \times 1$. In the encoder part, the basic block is a convolution layer followed by a batch normalization layer [99] and rectified linear unit activation function [100]. We can express the building block in the segmentation module formed as follows: $(Conv - BN - ReLU) \times 2$. The spatial size of the feature maps in the encoder part is reduced by the max-pooling layer. In the decoder part, the transposed convolution layer [126] is utilized to increase the spatial size of the feature maps while reducing the number of feature maps. In the end, a 1×1 convolution layer handles the feature maps to generate the final green pepper segmentation map. The skip connection design lets the feature maps in the encoder part directly share with the decoder part to avoid to lose essential spatial information. In our experiment, the only difference for the segmentation module in each model is the number of the input channel. Unlike the color-ratio maps enhancement methods, the model without color-ratio maps only needs three channels of input.

Table 5.1: The "U-Net-like" based segmentation module.

U-Net-like encoder			U-Net-like decoder		
Layer	**Details**	**Size**	**Layer**	**Details**	**Size**
input	R,G,B feature map+ color-ratio map	256x256 x12	upsampling1	2x2 upsample of block5 concatenate with block4	32x32 x1024
block1	{conv(3x3, pad=1)+Batch Norm ReLU}x2	256x256 x64	block6_1	{conv(3x3, pad=1)+Batch Norm ReLU}x2	32x32 x256
pool1	2x2 max pool; stride 2	128x128 x64	upsampling2	2x2 upsample of block6 concatenate with block3	64x64 x512
block2	{conv(3x3, pad=1)+Batch Norm ReLU}x2	128x128 x128	block7	{conv(3x3, pad=1)+Batch Norm ReLU}x2	64x64 x128
pool2	2x2 max pool; stride 2	64x64 x128	upsampling3	2x2 upsample of block7 concatenate with block2	128x128 x256
block3	{conv(3x3, pad=1)+Batch Norm ReLU}x2	64x64 x256	block8	{conv(3x3, pad=1)+Batch Norm ReLU}x2	128x128 x64
pool3	2x2 max pool; stride 2	32x32 x256	upsampling4	2x2 upsample of block8 concatenate with block1	256x256 x128
block4	{conv(3x3, pad=1)+Batch Norm ReLU}x2	32x32 x512	block9	{conv(3x3, pad=1)+Batch Norm ReLU}x2	256x256 x64
pool4	2x2 max pool; stride 2	16x16 x512	outconv	1x1x1	256x256 x1
block5	{conv(3x3, pad=1)+Batch Norm ReLU}x2	16x16 x512			

5.3.4 Loss function and physical constraint

As illustrated in Fig 5.1, we simultaneously optimize the TR curve and the segmentation module via the end-to-end system. The total loss function can be described as

$$\mathcal{L}_{total} = \mathcal{L}_{bce} + \eta \mathcal{L}_{smooth} \tag{5.8}$$

where $\mathcal{L}_{bce}$ denotes the binary cross-entropy loss for green pepper segmentation. It is defined as:

$$\mathcal{L}_{bce} = -\sum_{(x,y)}^{(H,W)} [G(x,y) \log P(x,y) + (1 - G(x,y)) \log(1 - P(x,y))] \tag{5.9}$$

where (x, y) is the pixel coordinates and (H, W) is image size: height and width. $G(x, y)$ and $P(x, y)$ denote the pixel values of the ground truth and the predicted segmentation probability map, respectively.

To aid the physical requirements of the TR curve in the optical filter fabrication, we introduce specific physical constraints for optical filter layer. As the filtered incident light should be positive, all the weights in the TR curve are non-negative. Besides, from the manufacturing perspective, the TR curve of the designed optical filter should avoid arbitrary and sudden variation between adjoining wavelengths. Hence, we proposed the physical constraint which can satisfy non-negative and smooth features as follows:

$$\mathcal{L}_{smooth} = \|\mathbf{GW}\|_2^2 \qquad s.t. \quad W \geq 0 \tag{5.10}$$

where the G denotes the second derivative matrix for optical filter layer, the W represents the weights of the 1×1 depth-wise convolution layer. The parameter η controls the smoothness of the TR curve for the optical filter. Due to the non-negative property of the optical filter weights, we enforced the non-negative $W \geq 0$ to the depth-wise convolution kernel of optical filter layer in the backward training procedure. In our experimental setting, we verified the different smoothness parameter η, e.g. $\eta = 0.1$, $\eta = 0.01$, $\eta = 0.001$. By explicitly modeling the TR curve of the optical filter with the specific physical constraint, our proposed optical filter layer can represent the property of a physical device in the real world. The actual optical filter will be fabricated to have the same transmittance curve as the learned weights in a further study.

5.4 Experimental Results and Analysis

To clearly explain our proposed method and determine the suitable parameters in the design space, we conduct several experiments and report in this section. In this section, we report the details of our experimental results and identify the essential parts that contribute to the overall system.

5.4.1 Hyperspectral dataset

Up to now, there is no public green pepper dataset in the research community and the Internet. Consequently, to construct a green pepper dataset for our research, we collected hyperspectral images at Next Generation Green House of the Kochi University of Technology and Kochi Agriculture Center, Kochi Prefecture, Japan. We selected a portable push-broom hyperspectral camera(Specim IQ, Specim Ltd., Finland) [127] as our data acquisition device. Hyperspectral images of green pepper were collected four times during May 2021 under sunny and cloudy weather conditions. The Specim IQ camera was set to Default Recording Mode(without any processing). In our workflow for image recording progress, we fixed the camera on the tripod, adjusted the camera position and white reference plate position. After that, we manually changed the camera focus and the integration time according to the captured scene. To accurately measure the illumination conditions in the captured scene, we put a standard white reference plate next to our target green pepper in the camera field of view. Sample images are illustrated in Figure 5.2.

Our hyperspectral camera can record 512×512 pixels image, with 204 spectral bands ranging from 400 nm to 1000 nm. The recording time of our hyperspectral camera for one image is from 40 seconds to 2 minutes in the different captured scenes. Compared with the laboratory illumination setting, natural illumination is always inconstant. On the one hand, various factors can affect the spectral power distribution of the illumination, e.g., climate, solar elevation [128]. On the other hand, mutual reflections between different surfaces, occlusions also lead to illumination variation in the natural environment [129]. In the end, we rendered the hyperspectral image to the sRGB image and made the ground truth for green pepper segmentation using the annotation tool LabelMe [102].

(a) The next generation Green House in KUT **(b)** Sample sRGB image and Ground Truth

Figure 5.2: The photograph of the Next Generation Green House in Kochi University of Technology(KUT) and Sample sRGB image and Ground Truth.

5.4.2 Experimental settings

As mentioned above, we collected green pepper hyperspectral dataset in our university Green House. Totally, we have 133 hyperspectral images. We randomly selected 101 images as our training set, 16 images as our validation set and 16 images as our test set. In our experiment, we apply the random crop 256×256 image patch from the original hyperspectral image and random horizontal flip to augment the dataset size of our training dataset. As a result, we obtained 7116 training patches for our experiment. Due to the spectral response of our camera is in the visible wavelength, we only used the hyperspectral image from 400 nm to 700 nm. Our experiment was conduct on an NVIDIA Tesla V100 GPU with the deep learning framework PyTorch [45]. The batch size is set to 32. The Adam optimizer [44] with beginning learning rate of 0.001 and $\beta = 0.5, \beta = 0.999$ was used in our experiment. We dynamic changed the learning rate by monitoring the performance on the validation set. The total epoch was set to 50, and the best model in the validation set is evaluated on the test dataset. The CSR of Lucid Triton 5.0 MP Model [106] was used in our experiment.

5.4.3 Experimental results

In this section, we compare the performance of different settings of our proposed method with the optical filter(without the color-ratio map) and no optical filter. After that, we illustrate the TR curve of the different settings of our proposed method. Lastly, we evaluate the effectiveness of the color-ratio maps.

Evaluation results

We refer to our proposed method as OF-CRM, Yu et al. [109] as OF, and no optical filter setting as NF, respectively. To evaluate and compare different setting approaches, we compute the mean intersection over union(mIoU) and F1 measure as following equations.

$$mIoU = \frac{1}{N_{class}} \sum_{i} \frac{p_{ii}}{t_i + \sum_j p_{ij} - p_{ii}} \tag{5.11}$$

$$\mathcal{F}_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.12}$$

where the p_{ij} be the number of pixels of class i predicted to belong to class j, there are totally N_{class} different classes, and let $t_i = \sum_j p_{ij}$ be the total number of pixels of class i. In our experiment, the number of classes is set to 2 (pepper or non-pepper). We evaluated different settings of max value in the normalization step, which is used to simulate camera saturation, and smoothness value η, which constraint transmittance curve smoothness. Table 5.2 shows the results of different maximum values and smoothness η settings. Empirically, we find the proposed model with $\eta = 0.001$ and $max = 4.470$ achieves the best performance in all the settings.

Remarkably, we notice that both the color-ratio maps and smoothness have influenced the shape of the designed TR curve. Looking at Figure 5.3, it is apparent that the optimal TR curve of $\eta = 0.1$, $\eta = 0.01$ turns to be the multiple bandpass optical filter. Intuitively, there is a clear trend of increasing η to generate more clear bandpass wavelength with the

Table 5.2: Quantitative comparison of different models. Our model outperform the optical filter design without color-ratio maps and no filter settings in the test dataset. The minimum value in Eq.(5.4) is same in all setting(min=0.008).

Models	smoothness	max	mIoU	F1
OF-CRM	$\eta = 0.001$	1.725	0.877	0.864
		2.615	0.878	0.874
		4.470	**0.899**	**0.891**
	$\eta = 0.01$	1.725	0.884	0.875
		2.615	0.866	0.853
		4.470	0.887	0.869
	$\eta = 0.1$	1.725	0.877	0.862
		2.615	0.874	0.862
		4.470	0.877	0.864
OF[109]	$\eta = 0.001$	1.725	0.875	0.858
		2.615	0.870	0.855
		4.470	0.869	0.846
	$\eta = 0.01$	1.725	0.850	0.823
		2.615	0.865	0.849
		4.470	0.877	0.862
	$\eta = 0.1$	1.725	0.864	0.841
		2.615	0.868	0.845
		4.470	0.852	0.822
NF	N/A	1.725	0.867	0.853
		2.615	0.857	0.832
		4.470	0.823	0.815

color-ratio map. Closer inspection of Figure 5.3, the transmittance in wavelength around 510 nm and 650 nm is almost zero in all settings, which is not helpful for green pepper segmentation.

In general, most plants look green in our human eyes due to Chlorophyll, which is vital for photosynthesis. There are two types of Chlorophyll in the land plants, Chlorophyll a and b. As reported in the previous study [130], they have different absorption spectrums. For example, the absorption peak of chlorophyll b is just below 650 nm. Interestingly, we can observe that the TR curves of all models are suppressed at wavelengths just below 650 nm. It can thus be suggested that the content of ChlorophyII a and b is different in

green pepper and leaves. The present study raises the possibility that our optical filter has found this Chlorophyll ratio differences in the same green color. This difference appears in the red channel, and it has proven to play an essential role in distinguishing green pepper and leaves, as we review in the following subsection. However, until now, we haven't found related studies to support this hypothesis, which supports the Chlorophyll ratio difference between green pepper and leaves. A further study with more focus on the ratio of Chlorophyll a and b is therefore suggested.

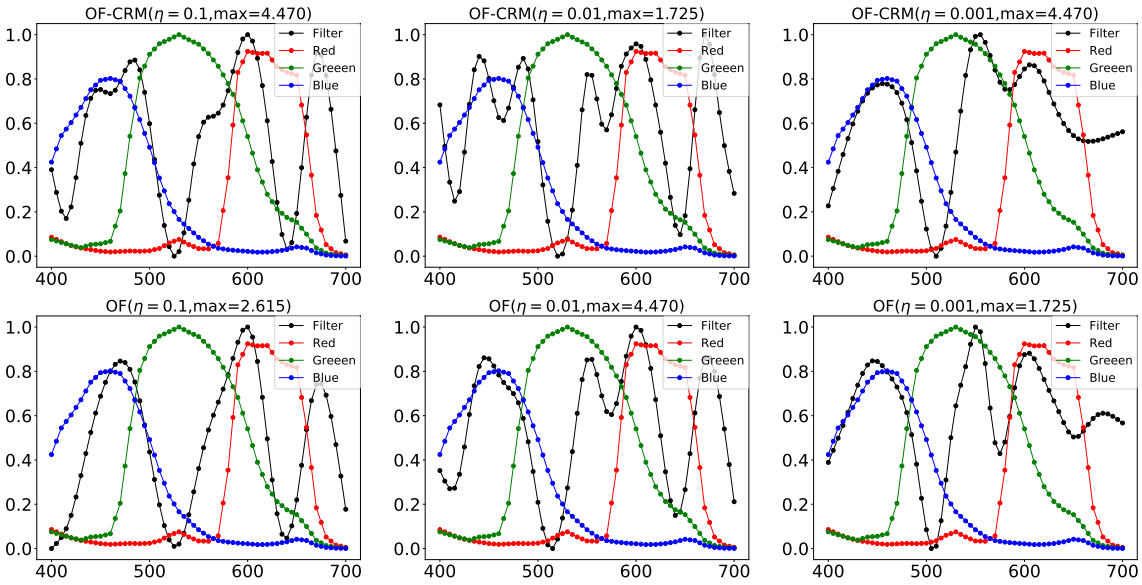
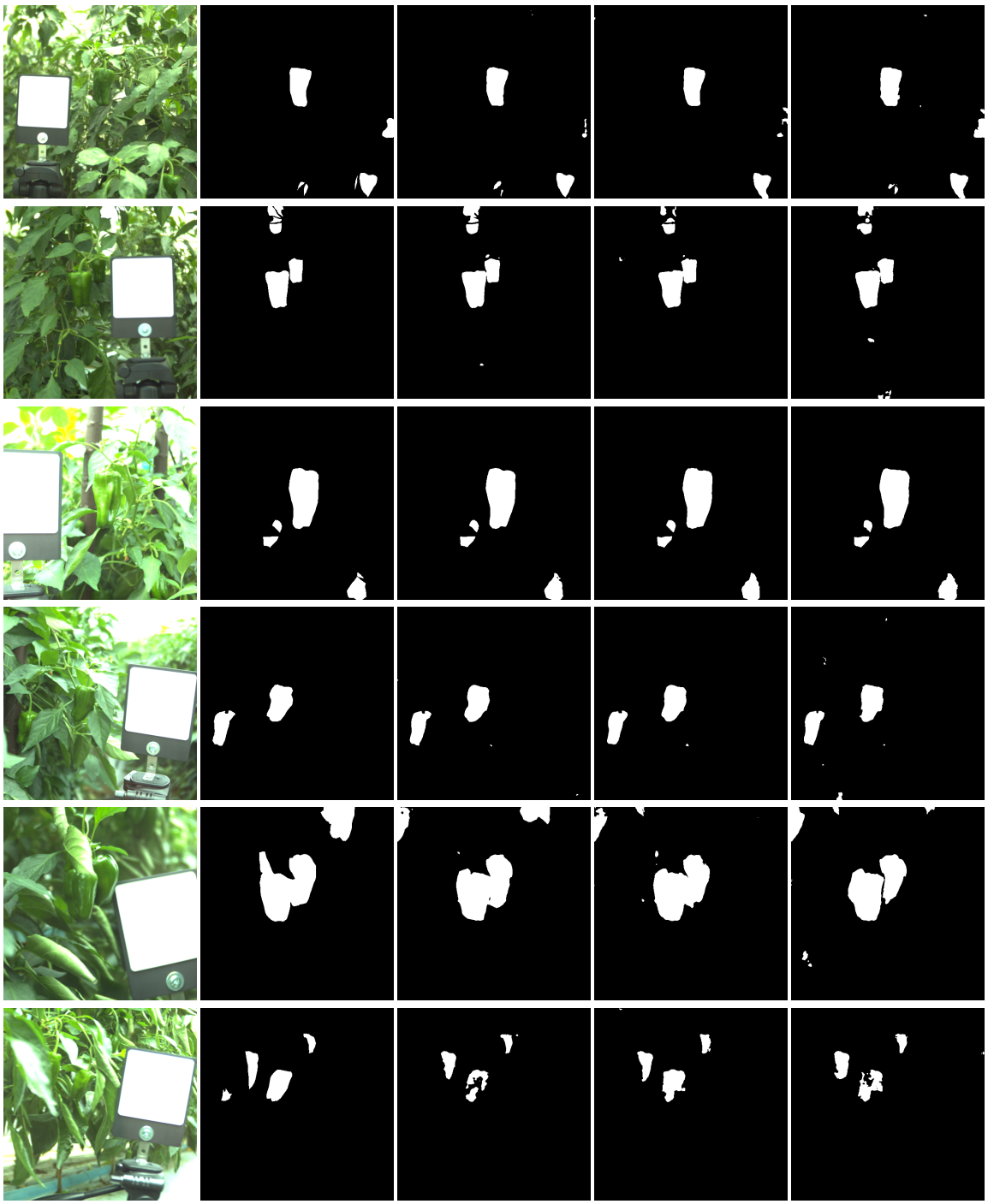


Figure 5.3: The TR curves of each model is illustrated in the above image. The top row shows the proposed model(with CRM), bottom row shows the optical filter design without CRM. In each η setting, we only demonstrate the best model among different max values.

Effectiveness of color-ratio maps

To empirically analyze how our proposed color-ratio maps work, we demonstrate the color-ratio maps on the test data. As shown in Figure 5.5, it is apparent that in the color-ratio map of $d5$ and $d8$, the green pepper is more distinguished. In the color-ratio map $d5$, the green pepper is highlighted. On the contrary, in $d8$, the green pepper looks dark than other



(a) RGB **(b)** label **(c)** OF-CRM **(d)** OF **(e)** NF

Figure 5.4: Segmentation results of each model in the test dataset. We only illustrate the best performance of each settings. **(c)** shows the best model in OF-CRM with smoothness $\eta = 0.001$ and max value 4.470. **(d)** illustrates the best model in OF with smoothness $\eta = 0.001$ and max value 1.725. **(e)** shows the best model in NF setting with max value 1.725.

parts of that color-ratio map. However, there are some of leaves also look dark, which is similar to the green pepper.

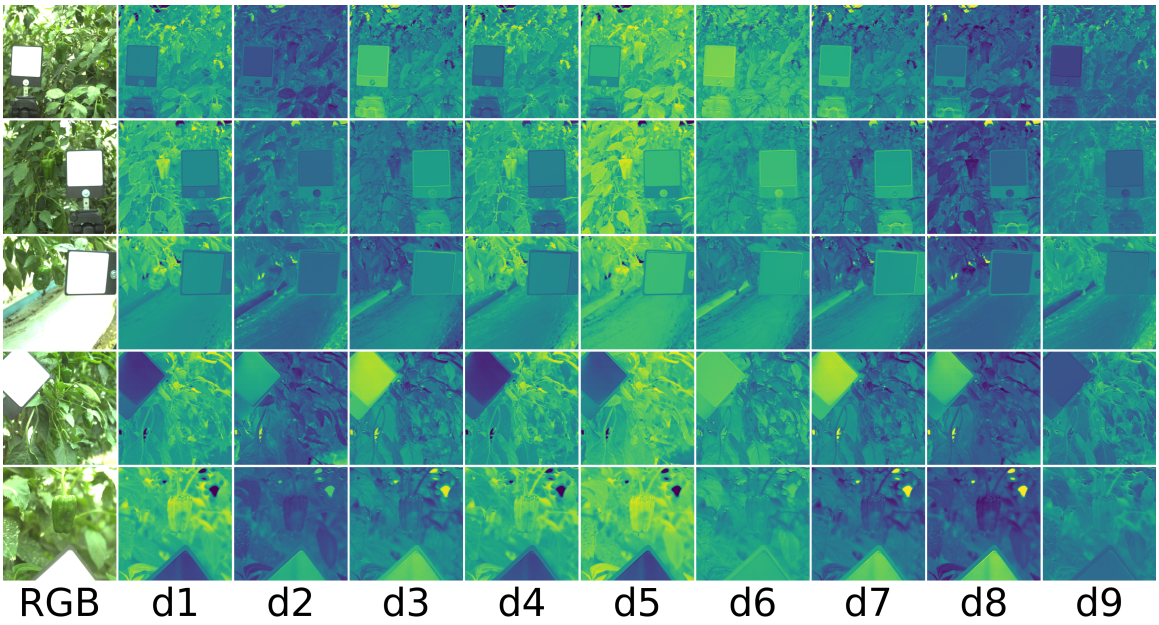
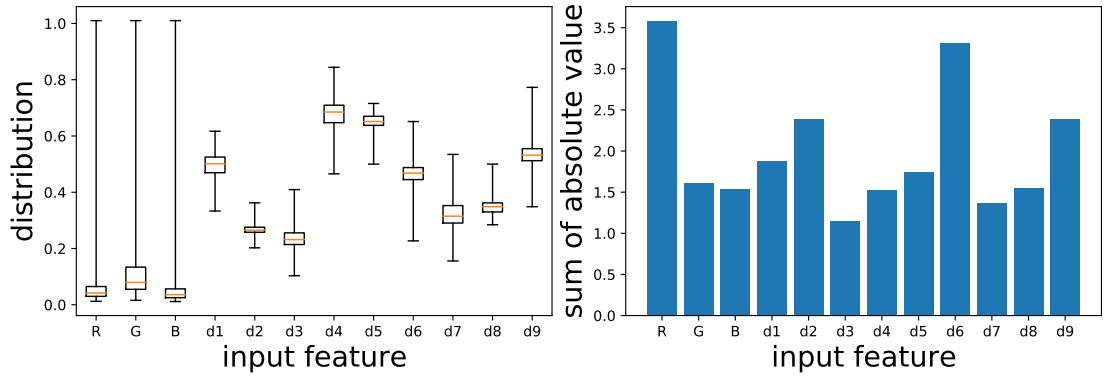


Figure 5.5: The above figures illustrate color-ratio maps of each test data. All color-ratio maps are shown in the same range [0,1].



(a) (b)

Figure 5.6: **(a)** The boxplot of the distribution of the input tensor of test dataset for OF-CRM($\eta = 0.001$, 4.470). **(b)** Sum of absolute value of all kernels for the input features of segmentation module in OF-CRM($\eta = 0.001$, 4.470). R channel and $d2 = R/(R+G+B)$, $d6 = B/(B+R)$, $d9 = R/(B+R)$ are more important than the other features and channels.

To analyze the importance of each input feature, especially color-ratio maps, we adopted the idea of the sum of absolute values of kernels used in filter pruning [131].

In figure 5.6(a), the distribution of each input feature is demonstrated by the boxplot. As can be seen from the figure, the distribution of the green channel and its corresponding color-ratio maps have slightly larger values than the other channels. However, since the differences are small, we introduced the sum of the absolute values of the kernels into the analysis to interpret the importance of each input feature. Figure 5.6(b) illustrates the sum of the absolute values of all kernels for input features, including R, G, B, and all color ratio maps. It seems that the red channel, $d2$, $d6$ and $d9$ are essential inputs for the segmentation module. It turns out that the ratio of the red channel to other channels can provide more meaningful information than other color-ratio maps. A critical hypothesis that emerged from the figure is the red channel is vital for green pepper segmentation. These results provide further support for the hypothesis that the red channel is vital for green pepper segmentation as mentioned in the previous part. We also illustrate the sum of the absolute value of each kernel for each input feature in Figure 5.7. As we know, each kernel in the convolutional layer pays attention to different input features. The fact that some kernels show large weights for CRM features indicates that the CRM features play an important role in distinguishing green peppers from leaves.

Figure 5.7: The sum of the absolute values of each input channel for each kernel. The horizontal axis represents the different input features of the segmentation module, R, G, B, $d1$, $d2$, $d3$, $d4$, $d5$, $d6$, $d7$, $d8$, $d9$ from left to right. The vertical axis of all subfigures are shown in the same range [0, 0.5]. The graph above shows that the color-ratio maps play an essential role, with some kernels showing larger in the color-ratio maps than in the R, G, B feature map.

5.5 Conclusions

In this chapter, we present an end-to-end optimization approach for the simultaneous design of optical filters and green pepper segmentation neural networks. We aim to leverage an end-to-end deep learning framework to find the optimal TR curve for green pepper segmentation. To accomplish this purpose, we model the critical components inside our

end-to-end framework, including the TR curve of the optical filter, CSR of the RGB camera, and our proposed color-ratio maps. Throughout our experiments, we demonstrate the proposed method achieved the best performance in mIoU and F1 measure.

As opposed to any deep learning based methods that operate directly on hyperspectral image or RGB image, our proposed approach has the ability to optimize the both TR curve of an optical element and weight of segmentation module simultaneously. Particularly, the design of TR curve of an optical element is enhanced by the color-ratio maps, which is useful for exploiting the spectral information. This study has been one of the first attempts to thoroughly examine the enhancement of color-ratio maps for optical filter optimization. Our future study will fabricate the optical filter according to the designed weights and evaluate its performance in a real application scenario.

6 Conclusion

This thesis has reported several works targeting research related to the attention-based approach and computational optics. The thesis presented to leverage the attention-based method to pain facial expression estimation for older people with moderate and severe dementia. In addition, we also propose a computational optics technique to jointly optimize the transmission curve of the optical filter and segmentation algorithm. In chapter 1, we introduced the background information about our research. Specifically, we introduced the importance of pain facial expression estimation for the elderly and the concept of computational optics and deep optics.

In chapter 2, we demonstrated the details of our proposed attention-based method for pain facial expression estimation. Because not all the facial areas contribute to the pain facial expression, we developed the spatial attention model to exploit the critical area in the human face for pain expression estimation. However, there is still a limitation of our proposed method. The facial expression happened not only in the spatial domain but also temporal domain. In the follow-up study, we will explore combining both spatial attention and temporal attention method. Until now, most facial pain expression database has provided 2D facial landmark to measure the geometry change of human face. However, the 2D landmark regards the face as a 2D object, which is frontal and planar. Some research reported the large head-poses would occur when patients feel pain. It will lead to the 2D landmark localization failed due to the self-occlusion. We will utilize the recently advanced 3D landmark localization method for pain facial expression estimation in future work.

Chapte 3 proposed the attention-based LSTM network structure for green pepper

segmentation by using hyperspectral imaging. Due to the accumulation effect of the cone cell in our eyes and camera spectral response function, the difference in spectral reflectance of the object surface will lead to the same color in our eyes and RGB cameras, such as green pepper and green leave. Hyperspectral imaging can provide much more details of the reflectance of the object surface than RGB camera. We treat each pixel of the hyperspectral image as sequence data. The proposed channel attention module can automatic select the important hyperspectral band for distinguishing green pepper and leave. Then, the LSTM network can effectively analyze and determine the input pixel categories via network training. Considering that hyperspectral imaging is a data cube, it has abundant spectral information and spatial information. Our future study will utilize spatial and spectral information to distinguish similar color fruits and vegetables.

Chapter 4 and Chapter 5 proposed the computational optics method for the vegetable segment, which can aid the automatic harvest, fruit and leave ratio estimation, and marketing strategy. Our proposed method successfully achieves the three critical problems mentioned in Chapter 1. Firstly, we utilized the depth-wise convolutional layer without bias and activation function to represent the optical filter in the whole network. Secondly, we proposed the physical-based constraint, which lets the weight of the optical filter layer be non-negative and smooth. Lastly, we leverage the binary classification loss function to optimize the whole structure. In addition, we successfully fabricate the designed optical filter. Furthermore, we propose to use the color-ratio map to enhance the optical filter design. It seems the color-ratio map can help us to exploit the different ratios of Chlorophyll a and b in green pepper and leave. In the future study, we will explore the new color space to enhance the optical filter design.

7 Academic Journal Publications

1. Jun Yu, Toru Kurihara, Shu Zhan, "Optical Filter Net: A Spectral-Aware RGB Camera Framework for Effective Green Pepper Segmentation," IEEE Access, Vol.9, pp.90142–90152, 2021.

2. Jun Yu, Toru Kurihara, Shu Zhan, "Color-Ratio Maps Enhanced Optical Filter Design and its application in green pepper segmentation," MDPI Sensors(under review)

3. Qiuyu Li, Jun Yu, Toru Kurihara, Haiyan Zhang and Shu Zhan, "Deep Convolutional Neural Network with Optical Flow for Facial Micro-expression Recognition," Journal of Circuits, Systems and Computers, Vol.29, No.1, 2050006, 2020. (Q4)

8 Academic Conference Publications

1. Jun Yu, Toru Kurihara, Zhan Shu, "Frame by Frame Pain Estimation Using Locally Spatial Attention Learning," Pattern Recognition and Image Analysis, pp.229-238, IbPRIA ,Madrid, Jul.1-4, 2019.

2. Jun Yu, Xinzhi Liu, Pan Wang,and Toru Kurihara, "Design of an optical filter to improve green pepper segmentation using a deep neural network," Pattern Recognition ACPR2019, LNCS, Vol.12047 pp.653–666, Auckland, New Zealand, Nov.26–29, 2019.

3. Jun Yu, Toru Kurihara, "New Deep Learning Architecture for Pain Intensity Estimation," International Workshop on Human-Engaged Computing, Kochi, Jan.11, 2019. (2019/01/11)

4. Toru Kurihara, Jun Yu, "Optical flow estimation using a correlation image sensor based on FlowNet-based neural network," Proc. of 15th International Conference on Computer Vision, Theory and Applications(VISAPP), Vol.4, pp.847–852, Malta, Feb.27-29, 2020.

5. Yijiu Yang, Jun Yu and Toru Kurihara, "Immature Yuzu Citrus Detection Based on DSSD Network with Image Tiling approach," Proc. of SICE Annual Conference, Tokyo, Japan, Sep.08-10, 2021.

6. Dong Ji, Jun Yu,Toru Kurihara, Liangfeng Xu, and Shu Zhan, "Automatic Prostate Segmentation on MR Images with Deeply Supervised Network," IEEE CoDIT, pp.309–314, Thessaloniki, Greece, Apr.10-13, 2018.

7. Qiuyu Li, Jun Yu,Toru Kurihara,and Shu Zhan, "Micro-expression Analysis by Fusing Deep Convolutional Neural Network and Optical Flow," IEEE CoDIT, pp.265-270, Thessaloniki, Greece, Apr.10-13, 2018.

9 Acknowledgment

Firstly, I would like to thank my academic supervisors at the Kochi University of Technology and Hefei University of Technology, Prof. Kurihara and Prof. Zhan, for helping me to improve my academic interests and provide educational guidance. Next, I would also like to thank the Kochi University of Technology and the Japan Ministry of Education, Culture, Sports, Science, and Technology for their generous financial support. Of course, I must thank the International Relations Center at the Kochi University of Technology, whose assistance in solving cultural and language problems has made a good living in a foreign country possible. Besides, The authors thank the anonymous reviewers for their helpful and constructive suggestions and comments. We also thank the cooperation of Kochi Agriculture Research Center and Go Ohira and Makoto Takehara for their hard work in data acquisition. And finally, I would like to thank my father and mother for instilling in my academic curiosity and for constantly pushing me to improve myself.

REFERENCES

[1] J. Yu, X. Liu, P. Wang, and T. Kurihara, "Design of an optical filter to improve green pepper segmentation using a deep neural network," in *Proc. of 5th Asian Conference on Pattern Recognition*. Auckland, New Zealand: Springer, 2019, pp. 653–666.

[2] H. Okamura, S. Ishii, T. Ishii, and A. Eboshida, "Prevalence of dementia in japan: a systematic review," *Dementia and geriatric cognitive disorders*, vol. 36, no. 1-2, pp. 111–118, 2013.

[3] M. Sado, A. Ninomiya, R. Shikimoto, B. Ikeda, T. Baba, K. Yoshimura, and M. Mimura, "The estimated cost of dementia in japan, the most aged society in the world," *PloS one*, vol. 13, no. 11, pp. 1–13, 2018.

[4] S. M. Zwakhalen, J. P. Hamers, H. H. Abu-Saad, and M. P. Berger, "Pain in elderly people with severe dementia: a systematic review of behavioural pain assessment tools," *BMC geriatrics*, vol. 6, no. 1, pp. 1–15, 2006.

[5] A. C. d. C. Williams, "Facial expression of pain, empathy, evolution, and social learning," *Behavioral and brain sciences*, vol. 25, no. 4, pp. 475–480, 2002.

[6] S. Lautenbacher and M. Kunz, "Facial pain expression in dementia: a review of the experimental and clinical evidence," *Current Alzheimer Research*, vol. 14, no. 5, pp. 501–505, 2017.

[7] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and electronics in agriculture*, vol. 153, pp. 69–81, 2018.

[8] M. Benavides, M. Cantón-Garbín, J. A. Sánchez-Molina, and F. Rodríguez, "Automatic tomato and peduncle location system based on computer vision for use in robotized harvesting," *Applied Sciences*, vol. 10, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5887

[9] J. Behmann, A.-K. Mahlein, T. Rumpf, C. Römer, and L. Plümer, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection," *Precision Agriculture*, vol. 16, no. 3, pp. 239–260, 2015.

[10] G. Gutiérrez-Gamboa, I. Díaz-Galvéz, N. Verdugo-Vásquez, and Y. Moreno-Simunovic, "Leaf-to-fruit ratios in vitis vinifera l. cv. "sauvignon blanc", "carmenère", "cabernet sauvignon", and "syrah" growing in maule valley (chile): Influence on yield and fruit composition," *Agriculture*, vol. 9, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2077-0472/9/8/176

[11] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, Utah, USA: IEEE, 2018, pp. 7794–7803.

[15] J. Hyttinen, P. Fält, H. Jäsberg, A. Kullaa, and M. Hauta-Kasari, "Optical implementation of partially negative filters using a spectrally tunable light source,

and its application to contrast enhanced oral and dental imaging," *Optics express*, vol. 27, no. 23, pp. 34 022–34 037, 2019.

[16] T. Wang and D. G. Dansereau, "Multiplexed illumination for classifying visually similar objects," *Applied Optics*, vol. 60, no. 10, pp. B23–B31, 2021.

[17] V. Boominathan, J. K. Adams, J. T. Robinson, and A. Veeraraghavan, "Phlatcam: Designed phase-mask based thin lensless camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1618–1629, 2020.

[18] B. Sun, J. Yan, X. Zhou, and Y. Zheng, "Tuning ir-cut filter for illumination-aware spectral reconstruction from rgb," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. virtual: IEEE, 2021, pp. 84–93.

[19] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d—learning phase masks for passive single view depth estimation," in *Proc. of IEEE International Conference on Computational Photography (ICCP)*. Tokyo, Japan: IEEE, 2019, pp. 1–12.

[20] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proc. of the IEEE/CVF International Conference on Computer Vision*. Seoul, Korea: IEEE, 2019, pp. 10 193–10 202.

[21] Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, "Learning rank-1 diffractive optics for single-shot high dynamic range imaging," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. virtual: IEEE, 2020, pp. 1386–1396.

[22] M. Kellman, E. Bostan, M. Chen, and L. Waller, "Data-driven design for fourier ptychographic microscopy," in *2019 IEEE International Conference on Computational Photography (ICCP)*. Tokyo, Japan: IEEE, May 2019, pp. 1–8.

[23] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 664–674, 2011.

[24] R. Irani, K. Nasrollahi, M. O. Simon, C. A. Corneanu, S. Escalera, C. Bahnsen, D. H. Lundtoft, T. B. Moeslund, T. L. Pedersen, M.-L. Klitgaard *et al.*, "Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Santiago, Chile, 2015, pp. 88–95.

[25] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *Proc. of 22nd International Conference on Pattern Recognition (ICPR)*. Stockholm, Sweden: IEEE, 2014, pp. 4582–4587.

[26] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Las Vegas, Nevada, USA: IEEE, 2016, pp. 84–92.

[27] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille, "Regularizing face verification nets for pain intensity regression," in *Proc. of IEEE International Conference on Image Processing (ICIP)*. Beijing, China: IEEE, 2017, pp. 1087–1091.

[28] P. Rodriguez, G. Cucurull, J. Gonzalez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE transactions on cybernetics*, no. 99, pp. 1–11, 2017.

[29] M. Tavakolian and A. Hadid, "Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition," in *Proc. of 25th IEEE International Conference on Image Processing (ICIP)*. Athens, Greece: IEEE, 2018, pp. 1952–1956.

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of International conference on machine learning*, 2015, pp. 2048–2057.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[32] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. of the European Conference on Computer Vision (ECCV)*. Munich, Germany: Springer, 2018, pp. 3–19.

[33] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Proc. of IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 57–64.

[34] K. D. Craig, K. M. Prkachin, and R. E. Grunau, "The facial expression of pain," in *Handbook of pain assessment*. The Guilford Press, 2001, pp. 153–169.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. of the IEEE conference on computer vision and pattern recognition*. Columbus, OH, USA: IEEE, 2014, pp. 1701–1708.

[37] E. G. Krumhuber, A. Kappas, and A. S. Manstead, "Effects of dynamic aspects of facial expressions: A review," *Emotion Review*, vol. 5, no. 1, pp. 41–46, 2013.

[38] W. Pei, H. Dibeklioğlu, T. Baltrušaitis, and D. M. J. Tax, "Attended end-to-end architecture for age estimation from facial expression videos," *IEEE Transactions on Image Processing*, vol. 29, pp. 1972–1984, 2020.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[41] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 286–299, 2017.

[42] M. Kunz and S. Lautenbacher, "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain," *European Journal of Pain*, vol. 18, no. 6, pp. 813–823, 2014.

[43] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. of the 13th IEEE International Conference on Automatic Face Gesture Recognition*. Xian, China: IEEE, 2018, pp. 59–66.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[46] C. McCool, I. Sa, F. Dayoub, C. Lehnert, T. Perez, and B. Upcroft, "Visual detection of occluded crop: For automated harvesting," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*. Stockholm, Sweden: IEEE, 2016, pp. 2506–2512.

[47] P. Eizentals and K. Oka, "3d pose estimation of green pepper fruit for automated harvesting," *Computers and Electronics in Agriculture*, vol. 128, pp. 127–140, 2016.

[48] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez, "Autonomous sweet pepper harvesting for protected cropping systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 872–879, 2017.

[49] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[50] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[51] Q. Gao, S. Lim, and X. Jia, "Hyperspectral image classification using convolutional neural networks and multiple feature learning," *Remote Sensing*, vol. 10, no. 2, 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/2/299

[52] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[53] R. Muthukrishnan and M. Radha, "Edge detection techniques for image segmentation," *International Journal of Computer Science & Information Technology*, vol. 3, no. 6, pp. 259–267, 2011.

[54] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 2015, pp. 3431–3440.

[55] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[56] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[57] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[58] J. Nalepa, M. Myller, and M. Kawulok, "Validating hyperspectral image segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1264–1268, 2019.

[59] N. Ketkar, "Introduction to keras," in *Deep learning with Python*. Springer, 2017, pp. 97–111.

[60] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

[61] H. Okamoto and W. S. Lee, "Green citrus detection using hyperspectral imaging," *Computers and electronics in agriculture*, vol. 66, no. 2, pp. 201–208, 2009.

[62] P. M. Mehl, Y.-R. Chen, M. S. Kim, and D. E. Chan, "Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations," *Journal of food engineering*, vol. 61, no. 1, pp. 67–81, 2004.

[63] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern trends in hyperspectral image analysis: A review," *IEEE Access*, vol. 6, pp. 14 118–14 129, 2018.

[64] C. T. Committee *et al.*, "Colorimetry," *CIE pub*, vol. 15, 2018.

[65] H. Blasinski, J. Farrell, and B. Wandell, "Designing illuminant spectral power distributions for surface classification," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2164–2173.

[66] B. E. Bayer, "Color imaging array," *United States Patent 3,971,065*, 1976.

[67] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *Proc. of European Conference on Computer Vision(ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 19–34.

[68] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational color constancy: Survey and experiments," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2475–2489, 2011.

[69] H. C. Karaimer and M. S. Brown, "A software platform for manipulating the camera imaging pipeline," in *Proc. of European Conference on Computer Vision(ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 429–444.

[70] B. Arad and O. Ben-Shahar, "Filter selection for hyperspectral estimation," in *Proc. of the IEEE International Conference on Computer Vision*, 2017, pp. 3153–3161.

[71] J. Jiang, D. Liu, J. Gu, and S. Süsstrunk, "What is the space of spectral sensitivity functions for digital color cameras?" in *Proc. of IEEE Workshop on Applications of Computer Vision (WACV)*. Tampa, FL, USA: IEEE, 2013, pp. 168–179.

[72] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Joint camera spectral sensitivity selection and hyperspectral image recovery," in *Proc. of the European Conference on Computer Vision (ECCV)*. Munich, Germany: Springer, 2018, pp. 788–804.

[73] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1375–1385.

[74] J. N. Martel, L. K. Mueller, S. J. Carey, P. Dudek, and G. Wetzstein, "Neural sensors: Learning pixel exposures for hdr imaging and video compressive sensing with programmable sensors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1642–1653, 2020.

[75] G. Côté, J.-F. Lalonde, and S. Thibault, "Deep learning-enabled framework for automatic lens design starting point generation," *Optics Express*, vol. 29, no. 3, pp. 3841–3854, 2021.

[76] S. Nie, L. Gu, Y. Zheng, A. Lam, N. Ono, and I. Sato, "Deeply learned filter response functions for hyperspectral reconstruction," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, Utah: IEEE, June 2018, pp. 4767–4776.

[77] S. Kitamura and K. Oka, "Recognition and cutting system of sweet pepper for picking robot in greenhouse horticulture," in *Proc. of IEEE International Conference Mechatronics and Automation*, vol. 4. Niagara Falls, ON, Canada: IEEE, 2005, pp. 1807–1812.

[78] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo, Japan: IEEE, 2013, pp. 5314–5320.

[79] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[80] J. Liang, J. Zhou, L. Tong, X. Bai, and B. Wang, "Material based salient object detection from hyperspectral images," *Pattern Recognition*, vol. 76, pp. 476–490, 2018.

[81] N. Imamoglu, Y. Oishi, X. Zhang, G. Ding, Y. Fang, T. Kouyama, and R. Nakamura, "Hyperspectral image dataset for benchmarking on salient object detection," in *Proc. of Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–3.

[82] N. İmamoğlu, G. Ding, Y. Fang, A. Kanezaki, T. Kouyama, and R. Nakamura, "Salient object detection on hyperspectral images using features learned from unsupervised segmentation task," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE, 2019, pp. 2192–2196.

[83] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *Proc. of 4th international Conference on Computer Vision*. Berlin, Germany: IEEE, 1993, pp. 173–182.

[84] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *International Journal of computer vision*, vol. 72, no. 3, pp. 239–257, 2007.

[85] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: a survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[86] S. Han, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato, "Camera spectral sensitivity estimation from a single image under unknown illumination by using fluorescence," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 805–812.

[87] A. Kimachi, S. Nishi, S. Ando, and M. Doi, "Three-phase quadrature spectral matching imager using correlation image sensor and wavelength-swept monochromatic illumination," *Optical Engineering*, vol. 50, no. 12, pp. 1 – 18, 2011.

[88] C. Kwan, B. Ayhan, B. Budavari, Y. Lu, D. Perez, J. Li, S. Bernabe, and A. Plaza, "Deep learning for land cover classification using only a few bands," *Remote Sensing*, vol. 12, no. 12, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/12/2000

[89] M. Dalla Mura, J. Atli Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010.

[90] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[91] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[92] H. Gan, W. S. Lee, V. Alchanatis, R. Ehsani, and J. K. Schueller, "Immature green citrus fruit detection using color and thermal images," *Computers and Electronics in Agriculture*, vol. 152, pp. 117–125, 2018.

[93] C. Kwan, B. Chou, J. Yang, A. Rangamani, T. Tran, J. Zhang, and R. Etienne-Cummings, "Deep learning-based target tracking and classification for low quality videos using coded aperture cameras," *Sensors*, vol. 19, no. 17, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3702

[94] C. Kwan, D. Gribben, B. Chou, B. Budavari, J. Larkin, A. Rangamani, T. Tran, J. Zhang, and R. Etienne-Cummings, "Real-time and deep learning based vehicle detection and classification using pixel-wise code exposure measurements," *Electronics*, vol. 9, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/6/1014

[95] C. Kwan, D. Gribben, A. Rangamani, T. Tran, J. Zhang, and R. Etienne-Cummings, "Detection and confirmation of multiple human targets using pixel-wise code aperture measurements," *Journal of Imaging*, vol. 6, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/40

[96] U. B. Gewali, S. T. Monteiro, and E. Saber, "Machine learning based hyperspectral image analysis: a survey," *arXiv preprint arXiv:1802.08701*, 2018.

[97] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[98] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[99] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[100] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[101] N. A. Hagen and M. W. Kudenov, "Review of snapshot spectral imaging technologies," *Optical Engineering*, vol. 52, no. 9, p. 090901, 2013.

[102] K. Wada, "labelme: Image Polygonal Annotation with Python," Available online: https://github.com/wkentaro/labelme, (Accessed: 2021-06-20).

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[104] R. Kawakami, H. Zhao, R. T. Tan, and K. Ikeuchi, "Camera spectral sensitivity and white balance estimation from sky images," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 187–204, 2013.

[105] D. H. Foster and K. Amano, "Hyperspectral imaging in color vision research: tutorial," *Journal of the Optical Society of America A*, vol. 36, no. 4, pp. 606–627, 2019.

[106] "Triton 5.0 mp model (imx264)," Available online: https://thinklucid.com/product/triton-5-mp-imx264/, (Accessed: 2021-08-25).

[107] J. Hemming, J. Ruizendaal, J. W. Hofstee, and E. J. Van Henten, "Fruit detectability analysis for different camera positions in sweet-pepper," *Sensors*, vol. 14, no. 4, pp. 6032–6044, 2014.

[108] H. Li, Q. Zhu, M. Huang, Y. Guo, and J. Qin, "Pose estimation of sweet pepper through symmetry axis detection," *Sensors*, vol. 18, no. 9, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/9/3083

[109] J. Yu, T. Kurihara, and S. Zhan, "Optical filter net: A spectral-aware rgb camera framework for effective green pepper segmentation," *IEEE Access*, vol. 9, pp. 90 142–90 152, 2021.

[110] K. Naoshi and E. Shunzo, "Studies on fruit recognizing visual sensor (i) calculation of the most suitable wavelength bands and discriminating experiment(in japanese)," *Journal of the Japanese Society of Agricultural Machinery*, vol. 49, no. 5, pp. 563–570, 1987.

[111] M. Omachi and S. Omachi, "Traffic light detection with color and edge information," in *Proc.of 2nd IEEE International Conference on Computer Science and Information Technology*. IEEE, 2009, pp. 284–287.

[112] C. Zhao, W. S. Lee, and D. He, "Immature green citrus detection based on colour feature and sum of absolute transformed difference (satd) using colour images in the citrus grove," *Computers and Electronics in Agriculture*, vol. 124, pp. 243–253, 2016.

[113] Y.-I. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," *Computer graphics and image processing*, vol. 13, no. 3, pp. 222–241, 1980.

[114] S. Moran, S. McDonagh, and G. Slabaugh, "Curl: Neural curve layers for global image enhancement," in *Proc.of 25th International Conference on Pattern Recognition (ICPR)*. Virtual-Milano: IEEE, 2021, pp. 9796–9803.

[115] Y. Monno, S. Kikuchi, M. Tanaka, and M. Okutomi, "A practical one-shot multispectral imaging system using a single image sensor," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3048–3059, 2015.

[116] Z. Frentress, L. Young, and H. Edwards, "Field photometer with nine-element filter wheel," *Applied Optics*, vol. 3, no. 2, pp. 303–308, 1964.

[117] J.-B. Thomas, P.-J. Lapray, P. Gouton, and C. Clerc, "Spectral characterization of a prototype sfa camera for joint visible and nir acquisition," *Sensors*, vol. 16, no. 7, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/7/993

[118] P.-J. Lapray, X. Wang, J.-B. Thomas, and P. Gouton, "Multispectral filter arrays: Recent advances and practical implementation," *Sensors*, vol. 14, no. 11, pp. 21 626–21 659, 2014.

[119] S. Nakauchi, K. Nishino, and T. Yamashita, "Selection of optimal combinations of band-pass filters for ice detection by hyperspectral imaging," *Optics express*, vol. 20, no. 2, pp. 986–1000, 2012.

[120] J. R. Bauer, A. A. Bruins, J. Y. Hardeberg, and R. M. Verdaasdonk, "A spectral filter array camera for clinical monitoring and diagnosis: Proof of concept for skin oxygenation imaging," *Journal of Imaging*, vol. 5, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2313-433X/5/8/66

[121] S. Ono, "Snapshot multispectral imaging using a pixel-wise polarization color image sensor," *Optics Express*, vol. 28, no. 23, pp. 34 536–34 573, 2020.

[122] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3081–3089.

[123] Y. Zou, Y. Fu, Y. Zheng, and W. Li, "Csr-net: Camera spectral response network for dimensionality reduction and classification in hyperspectral imagery," *Remote Sensing*, vol. 12, no. 20, p. 3294, 2020.

[124] Y. Zhu and G. D. Finlayson, "A mathematical investigation into the design of prefilters that make cameras more colorimetric," *Sensors*, vol. 20, no. 23, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/23/6882

[125] W. Wintringham, "Color television and colorimetry," *Proceedings of the IRE*, vol. 39, no. 10, pp. 1135–1172, 1951.

[126] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[127] J. Behmann, K. Acebron, D. Emin, S. Bennertz, S. Matsubara, S. Thomas, D. Bohnenkamp, M. T. Kuska, J. Jussila, H. Salo *et al.*, "Specim iq: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection," *Sensors*, vol. 18, no. 2, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/2/441

[128] J. Hernández-Andrés, J. Romero, J. L. Nieves, and R. L. Lee, "Color and spectral analysis of daylight in southern europe," *Journal of the Optical Society of America A*, vol. 18, no. 6, pp. 1325–1335, 2001.

[129] S. M. Nascimento, K. Amano, and D. H. Foster, "Spatial distributions of local illumination color in natural scenes," *Vision research*, vol. 120, pp. 39–44, 2016.

[130] W. T. Shoaf and B. W. Lium, "Improved extraction of chlorophyll a and b from algae using dimethyl sulfoxide," *Limnology and oceanography*, vol. 21, no. 6, pp. 926–928, 1976.

[131] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=rJqFGTslg