https://doi.org/10.2352/issn.2694-118X.2020.LIM-14

Deep Optimal Filter Responses for Multi-Spectral Imaging

Tarek Stiebel and Dorit Merhof
RWTH Aachen University

Abstract

Spectral recovery from measured camera signals based on deep learning lead to significant advancements of the potential reconstruction quality. However, most deep learning based approaches only consider RGB cameras and are targeting object classification in particular or remote sensing in general as their final application.

Within this work, we analyze the influence of a joint filter optimization and spectral recovery for multi-spectral image acquisition with the underlying goal of capturing high-fidelity color images. An evaluation on the influence of the total camera channel count on the reproduction quality is provided. Finally, a possible normalization of spectral data is discussed.

Introduction

Capturing high-fidelity color images typically relies on an accurate measurement of the incident spectral stimuli. The only possible alternative are dedicated RGB devices having a spectral response that is within a linear transform of the human color matching functions [1, 2]. However, such an approach is limited to exactly one human observer and does not allow for a person specific calibration of the reproduction. Capturing images at a high spectral resolution while retaining a reasonable spatial as well as temporal resolution forms an active field of research. Directly associated is the question on the amount of camera channels that are actually needed to obtain a sufficient spectral image in terms of its color accuracy.

The measurement of spectral images using multi-spectral imaging as well as an adequate signal processing to perform the spectral recovery is a long-standing problem. Classical examples comprise an approach based on the Wiener inverse [3] or techniques based on basis functions [4]. Recently, a variety of deep learning based methods was proposed and was found to significantly outperform non-deep learning based approaches. However, the focus of the deep learning community was almost exclusively on the recovery of spectral images from RGB signals. A concise overview can for example be found in [5]. The underlying motivation was almost never the aquisition of high-fidelity color-images.

In contrast to assuming imaging devices as fixed entities, it is also possible to consider a combined filter optimization for achieving an optimal spectral recovery. Recent non-deep learning based examples of such approaches are [6, 7]. Within this work, we focus on the potential color-accuracy of multi-spectral imaging systems in combination with deep learning and respectively optimized camera response functions. To the best of our knowledge, there exist three comparable published approaches: Fu et al. [8] perform a joint filter and spectral recovery optimization using deep learning. However, they restrict themselves to RGB imaging. Analogously, Nie et al. [9] succesfully learned a superior RGB Bayer-style 2x2 filter array and constructed a bi-spectral camera using their obtained response functions. Finally, Gewali et al. [10] perform a multi-spectral filter optimization for remote sensing with the spectral data reaching significantly beyond the visible wavelength range.

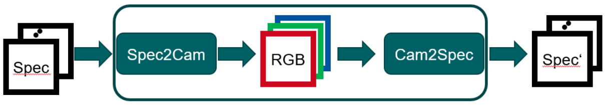


Figure 1: Autoencoder like setup for a joint optimization of a camera response and the spectral reconstruction.

The contribution of this work is as follows: First of all, a joint optimization of both the algorithm providing the spectral reconstruction as well as the camera response functions is conducted based on modern deep learning techniques. This optimization is not restricted to RGB imaging, although the amount of considered channels must be fixed beforehand. An analysis of the influence of the camera channel count on the accuracy of the spectral reconstruction is provided. Distinct spectral image datasets were considered, containing different scenarios ranging from pure spectral object reflectances to scenes captured in the wild under different lighting conditions. Finally, a normalization of spectral data is investigated.

Methods

The process of signal formation is modeled using the discretized process

$$\vec{\varphi} = (\varphi_1, ..., \varphi_n)^T = \mathbf{R}\vec{s}, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{nxq}$ denotes the camera response functions and $\vec{s} \in \mathbb{R}^q$ a spectral stimulus, assuming n camera channels and q spectral sampling points. Eq. 1 can be implemented as a custom layer of a neural network [9]. Such a layer receives a hyper-spectral image as input and computes the corresponding raw camera image, e.g. RGB. We implemented our own version not restricted to RGB imaging devices and therefore capable of simulating multi-spectral image acquisition. It will from now on be referred to as projection network. The projection network precedes an established network architecture performing the spectral recovery from camera signals. Both the projection network and the recovery network can then be jointly trained based on spectral image datasets under the premise, that the input spectral image, I_{in}, should match the output spectral image, I_{out}, using the error metric $L_{spectral}(I_{in}, I_{out})$. The resulting network combination is visualized in Fig. 1 and can be interpreted as an autoencoder. The projection network represents the encoder and learns the camera response functions such that the recovery network, the decoder, may perform the spectral reconstruction in an optimal way.

Constraints

A real-world camera response function is subject to several physical constraints, that were explicitly considered during the optimization process.

First of all, **positivity** of the sensor response is enforced by passing all coefficients of the response function through a rectified linear unit before performing the actual projection from spectral space onto camera signal space.

Next, the **smoothness** of the response functions is considered.

Especially in the context of camera device calibration, a wide variety of approaches has been proposed to enforce smoothness on the spectral response functions. One possibility is to describe the sensor response in terms of some inherently smooth basis, e.g. the Fourier basis as proposed by Finlayson [2]. Within a deep learning framework, Nie et al. [9] recommend using the L2 norm of the response as regularization term. The approach we found to work best is to introduce a regularization term

$$L_{smooth} = \frac{1}{n}\sum_{i=1}^{n} \vec{r_i}^{T}\mathbf{D}\vec{r_i}, \tag{2}$$

with

$$D = \begin{pmatrix} 1 & -1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \ddots & \ddots & 0 \\ 0 & -1 & 2 & -1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & -1 & 2 & -1 & 0 \\ \vdots & \ddots & \ddots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & 0 & -1 & 1 \end{pmatrix} \tag{3}$$

and r_i corresponding to the i'th row of the matrix $\mathbf{R}$, as proposed by Paulus et al. [11]. Equation 2 can be interpreted as the sum of squared differences of adjacent coefficients or as the energy of the second order derivative of the response function.

Finally, the **signal energy** of the response function is considered. The averaged signal energy of the response function needs to be minimized and is therefore introduced as another regularization term,

$$L_{energy} = \frac{1}{qn}\sum_{i=1}^{n}\sum_{k=1}^{q} \vec{r_i}[k]. \tag{4}$$

These constraints lead to the total loss function that was used during training:

$$L = L_{spectral} + \alpha L_{smooth} + \beta L_{energy}, \tag{5}$$

where α and β denote individual weighting terms. Within this work, the values $\alpha = 5$ and $\beta = 0.1$ were used.

Spectral reconstruction

The spectral recovery was performed with a modern convolutional neural network that was previously shown to reach state-of-the-art performance [5]. We chose the architecture proposed by Stiebel et al. [12] since there is an official implementation publicly available[1]. It is a U-Net based architecture that was modified to perform the regression task of spectral signal recovery. The only difference to the original architecture is the amount of input channels, which was modified to match the amount of considered camera channels. For all experiments, the model is trained from scratch using the original settings. In summary, the relevant hyper-parameters are a patch size of 32, a batch size of 10 and an initial learning rate of 0.001 using adam optimization. The mean relative absolute error (MRAE) was used as the spectral loss term $L_{spectral}$. Given two spectral cubes of spatial resolution MxN, the error metric is defined as

$$MRAE(I_{in}, I_{out}) = \frac{1}{MNq}\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{q}\left|\frac{I_{in}[i,j,k] - I_{out}[i,j,k]}{I_{in}[i,j,k]}\right|. \tag{6}$$

All parameters of the recovery network were randomly initialized utilizing a uniform distribution. In contrast, all parameters of the projection network were fixed to the value 1, i.e. the initial camera response function equals the value 1 for all channels. The complete model is implemented in Python using PyTorch and the training was executed on a GTX 1080 TI.

Spectral Image Data

A total of three different spectral datasets are considered, each of which was subdivided into a respective training, validation and testing set. Both the **CAVE** data [13] as well as the **NUS** dataset [14] are used offering image data containing pure spectral object reflectances. The NUS dataset contains a variety of different scenes. We restrict ourselves to the general scenes consisting of 52 outdoor and 36 indoor images. Both datasets were respectively subdivided into three random subsets of approximately equal size. Additionally, an extended version of the **ICVL** [15] dataset as it was used during the NTIRE 2018 reconstruction challenge [5] is considered[2]. In contrast to before, the ICVL data consists of spectral images captured in the wild, containing spectral reflectance functions under unknown illumination. In general, there is no information on the illuminant within the ICVL images available.

If not already the case, all spectral images were resampled within the spectral domain using a spline interpolation such that all images have a common spectral sampling range from 400nm to 700nm in 10nm steps. Lighting is computationally added to the CAVE and NUS images assuming the standard illuminant CIE D65.

Spectral Data Normalization

We found that working with multiple and distinct spectral datasets poses a non-trivial challenge for deep learning. A significant issue of the different spectral datasets is a difference in their signal value ranges. We therefore normalized every spectral dataset, respectively, to a zero mean and unit standard variation

$$\vec{s}_n = \frac{(\vec{s} - e\vec{1})}{\sigma}, \tag{7}$$

where e denotes the average signal value across all channels and images of a single dataset and σ the respective standard deviation. Solely the normalized data is used for training. We will now show that normalizing the data in the proposed way does not have any influence on the learned response functions:

$$\begin{aligned} \vec{\varphi} &= \mathbf{R}\vec{s} \\ &= \mathbf{R}(\sigma\vec{s}_n + e\vec{1}) \\ &= \sigma\mathbf{R}\vec{s}_n + e\mathbf{R}\vec{1} \\ \vec{\varphi} - e\mathbf{R}\vec{1} &= \sigma\mathbf{R}\vec{s}_n \end{aligned} \tag{8}$$

The neural network is effectively learning Eq. 9 when using the normalized spectra as input,

$$\vec{\varphi}_n = \mathbf{R_n}\vec{s}_n. \tag{9}$$

The analogy is directly apparent when comparing Eq. 8 and Eq. 9. In contrast to learning the original response function, a normalized version is learned, $\mathbf{R_n} = \sigma\mathbf{R}$, that only differs up to scale. However, this is not even a concern, since we are usually interested in the relative response function anyway (and the

[1] https://github.com/tastiSaher/SpectralReconstruction

[2] available at http://icvl.cs.bgu.ac.il/ntire-2018/

Dataset	Metrics	2	3	4	5	6	7	8
	RMSE	17.3	11.3	9.48	9.32	9.40	9.15	8.86
ICVL	MRAE	2.27	1.28	1.13	1.08	1.06	0.999	0.964
	GFC	0.99957	0.99986	0.99989	0.99989	0.99989	0.99991	0.99991
	ΔE_{00}	3.8	1.3	1.1	0.92	1.1	1.2	1.2
CAVE	RMSE	1.66	0.827	0.704	0.632	0.709	0.679	0.7
	MRAE	20.5	15.0	13.6	12.9	13.5	13.7	13.6
	GFC	0.98434	0.9944	0.99556	0.99567	0.99536	0.99533	0.99523
	ΔE_{00}	4.1	2.0	1.5	0.89	0.78	0.62	0.77
NUS	MRAE	11.1	8.14	6.15	5.05	4.49	4.24	4.33
	RMSE	2.18	1.72	1.14	0.919	0.789	0.748	0.721
	GFC	0.99279	0.996	0.99778	0.99856	0.99881	0.99891	0.99894

Table 1: Average reconstruction quality over the channel count.

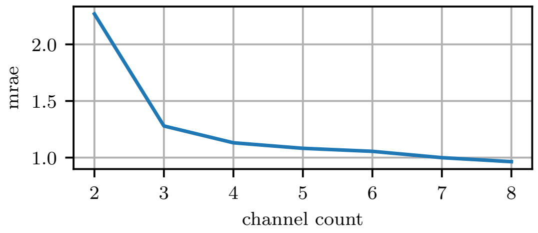


Figure 2: The spectral reconstruction error over the camera channel count for the ICVL data.

scaling factor is even known). $\mathbf{R_n}$ retains all major properties of $\mathbf{R}$. Of greater interest is the introduced constant offset, $\vec{c} = e\mathbf{R}\vec{1}$, of the camera signal values, which might be interpreted as a weighted ideal white point. In the final application, the decoder network will be used as a stand-alone for only performing the task of spectral reconstruction from camera images. Any potential input signal, $\vec{\varphi}$, must be modified beforehand according to

$$\vec{\varphi}_n = \vec{\varphi} - \vec{c}. \tag{10}$$

Results & Discussion

The joint projection and recovery network was respectively trained on each dataset with an increasing amount of camera channels ranging from two to nine. For the evaluation, all images of the testing sets were processed in the same way. A simulated multi-spectral image was created using the learned response functions. Subsequently, the spectral recovery was performed. The quality of the obtained spectral reconstruction was evaluated using multiple error metrics. First of all, the root mean squared error (RMSE), the mean relative absolute error (MRAE) and the goodness-of-fit coefficient (GFC) [16] are provided to evaluate the reconstruction within the spectral domain. For both the CAVE and NUS datasets, the color accuracy was evaluated based on the CIEDE2000 color difference [17], ΔE_{00}, with respect to CIE D65 lighting.

The average reconstruction errors over the respective test sets are summarized in Tab. 1. Fig. 2 visualizes the general quality of the spectral reconstruction vs. the channel count for the ICVL data. The intuitive result is directly observable: the reconstruction error decreases the more channels are available. When comparing the 3 channel system trained on the ICVL data to the results achieved at the NTIRE challenge, a significant performance increase can be observed due to the optimized response functions, whereas the challenge employed human color matching functions [5]. The high reconstruction errors on the CAVE dataset can be attributed to its comparably low training data size. For comparison, the ICVL training data size is approximately 180 times of the size when using the CAVE dataset instead. Still the trend holds, that with an initially increasing channel count the reconstruction error decreases. However, the higher the channel count becomes, the more training data can be expected to be required due to an increased complexity of the reconstruction. This is why for channel counts higher than six the reconstruction quality becomes worse for the CAVE data. The complexity increases while the underlying dataset size was probably insufficient for the considered workflow from the start. The achieved results on the NUS dataset are consistent and within expectation. An exemplary image of an achieved reconstruction when using 3 camera channels and the NUS data is visualized in Fig. 3. Noteworthy are the color checker charts which appeared to be the most challenging while the underlying scenes were well reconstructed.

In general it can be summarized, that the major and reliable gain in reconstruction quality can be found within the first channels. For example, adding an additional channel to a RGB system has a significantly higher impact than increasing the channel count from 7 to 8. It can be stated that the potential improvement is comparably negligible once five channels have been reached.

Finally, the learned camera response functions are analyzed. Fig. 4 shows the learned relative camera response functions for an increasing channel count when trained on the ICVL data. The learned response functions when instead trained on the CAVE or NUS data are comparable. At first glance, the learned functions show a remarkable resemblance to known filter design. They basically appear as Gaussian-like functions that are equally spaced over the wavelength range. A major difference to human design is a lack of uni-modality, although most of the learned response functions can be interpreted as uni-modal Gaussians. A close examination reveals that some channels contain multiple bandpasses at once. For common multi-spectral camera design, the channels usually become the more narrowband the more camera channels there are. This does not appear to be the case for the learned response functions. Broadband channels appear to be preferred, even for higher channel counts. A closer examination will be necessary in order to decide if this is simply an effect due to the chosen constraints on the camera response functions or if the broadband characteristic is actually superior, which will be investigated in future work.

Conclusion

Within this work, we successfully optimized both the spectral sensitivity as well as the spectral recovery from measured camera signals in a joint fashion for multi-spectral imaging devices. This could be achieved by employing modern deep learning techniques in conjunction with physically based regularization terms. A normalization for distinct hyper-spectral datasets was proposed, allowing for a general application of deep learning based approaches without the necessity for any modifications. It was found that a channel count of 5 is already sufficient for pro-

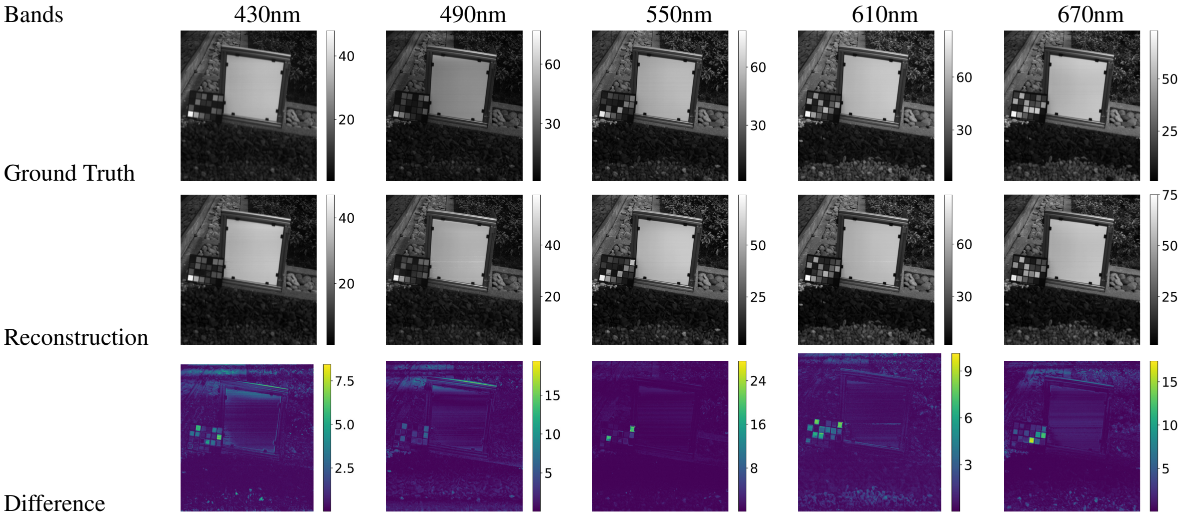
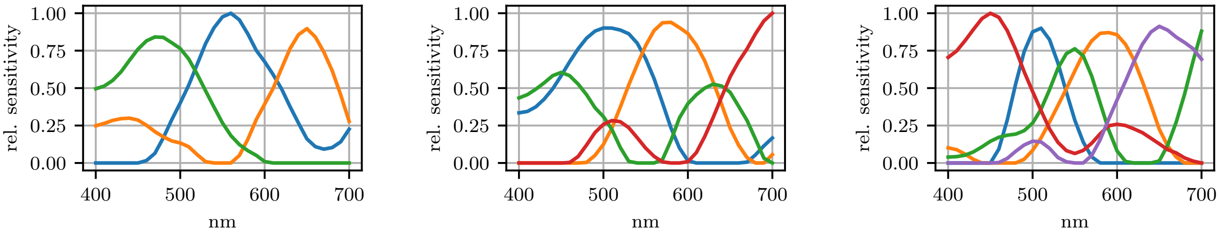


Figure 3: Visualization of reconstructed spectral bands for an NUS test image under simulated CIE D65 illumination.



(a) 3 channel

(b) 4 channel

(c) 5 channel

Figure 4: Exemplary learned relative spectral sensitivity functions on the ICVL dataset.

viding highly accurate spectral images within the visible wavelength range, when the spectral recovery is performed using a state-of-the-art neural network. A further increase of channels does only slightly improve the accuracy of the spectral measurement. However, it should be noted that depending on the application, the used spectral resolution of the ground-truth spectral images themselves comprising 31 channel ranging from 400nm to 700nm can be seen as too low. Extremely narrowband spikes originating from e.g. fluorescent light sources can hardly be expressed in such a way. The underlying spectral image data should therefore be considered as rather well behaved and smooth.

References

[1] R. Luther, “Aus dem Gebiet der Farbreizmetrik,” *Zeitschrift für technische Physik*, vol. 8, pp. 540–558, 1927.

[2] G. Finlayson, Y. Zhu, and H. Gong, “Using a simple colour pre-filter to make cameras more colorimetric,” *Color and Imaging Conference*, vol. 2018, pp. 182–186, 11 2018.

[3] B. Hill, “Optimization of total multispectral imaging systems: best spectral match versus least observer metamerism,” in *9th Congress of the International Colour Association*, vol. 4421, International Society for Optics and Photonics. SPIE, 2002, pp. 481 – 486.

[4] W. Praefke and T. Keusen, “Optimized basis functions for coding reflectance spectra minimizing the visual color difference,” in *Third Color Imaging Conference*. Society for Imaging Science and Technology, 1995, pp. 37–40.

[5] B. Arad, O. Ben-Shahar, R. Timofte, L. Van Gool, L. Zhang, M.-H. Yang *et al.*, “Ntire 2018 challenge on spectral reconstruction from rgb images,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[6] B. Arad and O. Ben-Shahar, “Filter selection for hyperspectral estimation,” in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] H.-L. Shen, J.-F. Yao, C. Li, X. Du, S.-J. Shao, and J. H. Xin, “Channel selection for multispectral color imaging using binary differential evolution,” *Appl. Opt.*, vol. 53, no. 4, pp. 634–642, Feb 2014.

[8] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, “Joint camera spectral sensitivity selection and hyperspectral image recovery,” in *The European Conference on Computer Vision (ECCV)*, September 2018.

[9] S. Nie, L. Gu, Y. Zheng, A. Lam, N. Ono, and I. Sato, “Deeply learned filter response functions for hyperspectral reconstruction,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 4767–4776.

[10] U. Gewali, S. Monteiro, and E. Saber, “Spectral super-resolution with optimized bands,” *Remote Sensing*, vol. 11, p. 1648, 07 2019.

[11] D. Paulus, J. Hornegger, and L. Csink, “Linear approximation of sensitivity curve calibration,” in *8. Workshop Farbbildverarbeitung*, 2002, pp. 3–10.

[12] T. Stiebel, S. Koppers, P. Seltsam, and D. Merhof, “Reconstructing spectral images from rgb-images using a convolutional neural network,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[13] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, “Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum,” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 19, pp. 2241–53, 03 2010.

[14] R. M. H. Nguyen, D. K. Prasad, and M. S. Brown, “Training-based spectral reconstruction from a single rgb

image," in *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 186–201.

[15] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–34.

[16] F. H. Imai, M. R. Rosen, and R. S. Berns, "Comparative study of metrics for spectral match quality," in *Conference on Colour in Graphics, Imaging, and Vision (CGIV)*, 2002, pp. 492–496.

[17] Commission Internationale de l'Eclairage, "Improvement to industrial colour-difference evaluation," Central Bureau of the CIE, CIE Publication 142-2001, January 2001.