Machine Vision and Applications (2021) 32:9
https://doi.org/10.1007/s00138-020-01136-8

ORIGINAL PAPER

Snapshot hyperspectral imaging using wide dilation networks

Mikko E. Toivonen[1] · Chang Rajani[1] · Arto Klami[1]

Received: 4 September 2019 / Revised: 16 September 2020 / Accepted: 29 September 2020

Abstract
Hyperspectral (HS) cameras record the spectrum at multiple wavelengths for each pixel in an image, and are used, e.g., for quality control and agricultural remote sensing. We introduce a fast, cost-efficient and mobile method of taking HS images using a regular digital camera equipped with a passive diffraction grating filter, using machine learning for constructing the HS image. The grating distorts the image by effectively mapping the spectral information into spatial dislocations, which we convert into a HS image by a convolutional neural network utilizing novel wide dilation convolutions that accurately model optical properties of diffraction. We demonstrate high-quality HS reconstruction using a model trained on only 271 pairs of diffraction grating and ground truth HS images.

Keywords Hyperspectral imaging · Deep learning · Convolutional neural networks

1 Introduction

In hyperspectral imaging, one wishes to capture an image that provides for each pixel the spectrum at a continuous range of wavelengths [1,2]. Since many materials have a unique spectral signature, one can use HS images to, for example, segment images according to materials [3]. This makes HS images useful in wide range of tasks in science and industry, such as satellite surveying [4,5], food quality assurance [6], gas and oil exploration [7], and various medical applications [8].

Special devices called hyperspectral cameras are used to take HS images. These devices generally operate by scanning the scene either spatially (spatial scanning) or spectrally (spectral scanning) [9], and capture tens to hundreds of spectral channels to preserve the shape of the spectrum, as opposed to multispectral cameras that record fewer, possibly disjoint, spectral channels [3]. Capturing a single image in good lighting conditions might take tens of seconds using a scanning method, since the camera needs to capture each spatial or spectral dimension separately. Furthermore, the spatial resolution at which these cameras operate is typically low—for example, the Specim IQ, a portable HS camera, yields images of size 512 × 512 [2], and more refined stationary models yield images of 1–2 MP. These specialized devices are also expensive, currently costing in the order of tens of thousands of euros or US dollars.

In contrast to the scanning approach, *snapshot* imaging techniques capture the entire hyperspectral cube at once. They are based, for example, on prism and beam-splitter constructs [10], per-pixel filters at the image sensor [11], or tunable narrow-band optical filters [12]. These methods have the advantage of short capture time, but still require costly specialized hardware. Recently it was demonstrated that even capturing HS video at high frame-rate is possible [13], by combining high-speed RGB video with specialized HS imaging hardware operating on lower temporal frequency.

To combine low cost with fast image acquisition, we need snapshot imaging without active mechanical elements. This can be done by using a diffraction grating filter [14–16], a prism [17], or an optical diffuser [18], followed by algorithmic reconstruction of the HS image. The existing work in this direction, however, has serious limitations. The early works using a diffraction grating employ linear reconstruction models that can only produce images of extremely low resolution [14,15], whereas the more recent work using a prism requires time-consuming post-processing for creating the HS image (Baek et al. [17] report 45 min for creating a

✉ Mikko E. Toivonen
tomito@cs.helsinki.fi

Chang Rajani
chra@cs.helsinki.fi

Arto Klami
aklami@cs.helsinki.fi

[1] Department of Computer Science, University of Helsinki, Helsinki, Finland

Published online: 16 October 2020

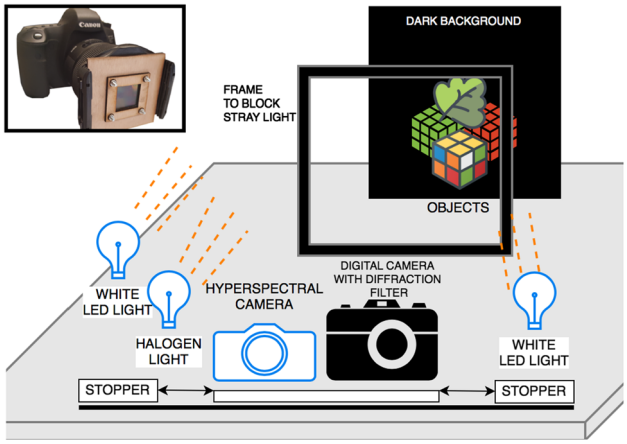

Fig. 1 (Top left) The custom diffraction grating mounted in front of a DSLR camera. (Base figure) setup for acquisition of the training image pairs. The hyperspectral camera and digital camera are placed side by side on a horizontal slide with stoppers, allowing for both cameras to capture the scene from the same location. The lighting was composed of two white led light sources and a halogen light for a more complete spectral illumination. The purpose of the frame is to stop excessive stray light from the background interfering with the diffraction

512 × 512 image), largely defeating the advantage of rapid image acquisition.

We present a method of capturing hyperspectral images by combining low-cost passive filter with deep learning. We attach a diffraction grating filter to a standard digital camera (Fig. 1, top left), in order to distribute the spectral information in the scene into the spatial dimensions in a specific—but difficult to model—manner. We then propose a novel convolutional neural network (CNN) [19] variant for inverting the diffraction and reconstructing the spectral information, essentially assigning the spatial diffraction patterns back into the spectral dimension. The technique combines fast image acquisition with a HS image construction algorithm that runs in under a second.

Our model is based on CNNs used for similar image-to-image visual tasks, such as single-image super resolution (SISR) [20]. The core novelty is the use of multiple concurrent convolutional layers with very large *dilation rates*, which allows maintaining a large spatial range with a small number of parameters. We show that such filters accurately model the underlying phenomena of diffraction, and present a way of automatically detecting the dilation rate hyperparameters based on captured image data, removing the need for a separate calibration process. The wide dilated convolutions are coupled with two other elements critical for construction of high-quality HS images: (a) residual blocks, as used in the ResNet architecture [21], for correcting for nonlinear effects not adequately modeled by the convolution layer, and (b) loss function that balances between reconstruction of the spatial structure in the image and reconstruction of the spectral characteristics of individual pixels. Furthermore, the model architecture is designed such that we can control the output resolution by adapting the dilation rate of the convolutions. This allows producing HS images of higher resolution than what is available for training the model.

By taking into account the physical properties of diffraction in the network architecture, we are able to train our model with a relatively small dataset. For training and evaluation of the model, we used pairs of hyperspectral images taken with the Specim IQ HS camera [2], and diffraction images taken with a standard digital camera and a diffraction grating filter. For this purpose, we collected a set of 271 image pairs in controlled conditions; see Fig. 1 for illustration and Sect. 4.1 for details on the experimental setup. We show the effectiveness of our technique both qualitatively, in terms of visual inspection of the of output, and quantitatively, in terms of error w.r.t. ground truth images of image pairs not used for training the model. We demonstrate high overall quality of resulting HS images and hence provide a proof-of-concept for low-cost and time-efficient high-resolution hyperspectral imaging.

2 Background: hyperspectral imaging

2.1 Dispersion and diffraction

Hyperspectral imaging is traditionally performed using a light dispersive element, such as a diffraction grating or a prism, and some scanning method [1]. The purpose of the dispersive element is to direct different wavelengths of light toward different locations in a sensor. Prisms achieve this by refraction of light and diffraction grating filters by diffraction. In both cases, the angle of dispersion—and hence the location of a light beam on the imaging sensor—depends on the wavelength of the light and physical characteristics of the dispersion element. For prisms, the dispersion is controlled by the shape and index of refraction, and for diffraction gratings by the spacing of the grating, the grating constant.

We use a diffraction grating element that consists of an array of equally spaced horizontal and vertical slits generating a grid of apertures, each of which causes diffraction. This diffraction grating causes constructive interference to be maximized at the direction of incident light, which is called the *zeroth-order diffraction component*. More diffraction components that are integer multiples of the angle of diffraction are also formed, denoted by their-order number, e.g., the *first-order diffraction component*. The intensities of the diffraction components are inversely proportional to their order number, the zeroth-order component having the highest intensity, and the following ones having diminishing intensity. We are mainly concerned with the zeroth and first-order diffraction components, because higher-order components have a much lower relative intensity.

An important observation is that the diffraction grating disperses the spectrum of each spatial area in the scene into the surrounding areas on the sensor, which may be on top of other objects in the scene. While it is in principle possible to model which part of the spectrum gets diffracted where (and apply a deconvolution to reverse it), lens curvature and the specific camera used cause additional nonlinearities. In Sect. 5, we empirically demonstrate that modeling these nonlinearities clearly improves the accuracy.

2.2 Standard hyperspectral imaging

Traditional HSI techniques capture the hyperspectral images by scanning each spectral or spatial dimension at a time. This process can involve using a mechanical device to shift a narrow aperture or a slit, like in the Specim IQ camera used in our experiments [2]. Spectral scanning can alternatively be performed using a Fabry–Perot interferometer as a tunable, narrow band bandpass filter [12], which performs optical filtering as opposed to dispersing spectral components of light. The final hyperspectral image is then formed by processing either the spectral or spatial slices.

Snapshot imaging [22] enables taking the whole HS image at once, which offers a significant advantage in terms of imaging speed. However, existing solutions are expensive and based on complex hardware [10,11].

2.3 Passive hyperspectral imaging

Our primary goal is to avoid use of expensive active elements, and hence the most closely related work is on combination of passive dispersive or diffractive elements combined with algorithmic reconstruction of the HS image.

The idea of re-constructing HS images from images taken through a diffraction grating filter was presented first by Okamoto and Yamaguchi [14]. They proposed a multiplicative algebraic reconstruction technique (MART) for generating the HS images, and Descour and Dereniak [15] provided an alternative reconstruction algorithm based on computed tomography. More recently, computed tomography was used for retinal hyperspectral imaging based on custom camera and diffraction grating filter [16]. While these early works demonstrated the feasibility of snapshot HS imaging with passive filters, their experiments were limited to images of 11×11 [15] and 72×72 pixels [14]. Modern computing hardware would help increasing the resolution, but the reconstruction algorithms would not scale to resolutions in the order of megapixels because they are based on storing and processing the whole system matrix of number of diffraction image elements times the number of hyperspectral image elements. For example, for reconstructing a $256 \times 256 \times 100$ HS image based on 1 MP diffraction image, the system matrix would consume approximately 20 TB of memory. Furthermore, the reconstruction algorithms are not robust for real-world data due to a strong linearity assumption that does not hold for most imaging setups or devices.

Another snapshot-based HS imaging method is based on combining a digital camera with an optical diffuser [18], more specifically a restricted isometry property (RIP) diffuser. The RIP diffuser diffuses the light onto the sensor, and a custom algorithm, relying on the RIP condition of the diffuser and sparsity of the reconstruction solution in some wavelet-frame domain, performs the reconstruction of the hyperspectral cube using a linear interactive split Bregman method [23]. The imaging and reconstruction method is shown to produce hyperspectral cubes of 256×256 for 33 narrow wavelength bands in the range of 400nm to 720nm. A method for diffraction-grating based hyperspectral imaging is also given by Habel et al. [24]. They use an additional lens construct together with a diffraction grating attached to a standard digital camera and present a method based on computed tomography imaging spectrometry involving spectral demosaicing and reconstruction. They are able to produce HS images of 124×124 pixels and 54 spectral channels, but the approach requires extensive camera-specific calibration.

In addition to diffraction gratings, prisms can be used as the dispersive element. Baek et al. [17] attached a prism to a DSLR lens and reconstructed HS images based on spatial dispersion of spectra over the edges in the captured image and subsequent detection of spectral edge blur. Their solution is conceptually similar to ours: both use passive add-on device and spectral reconstruction is performed computationally. However, prisms are considerably larger and heavier than diffraction grating filters, and their method of spectral reconstruction is computationally very expensive, consuming a total of 45 min for a single 512×512 pixel image on a desktop computer.

Our solution is qualitatively different from the earlier works of [14,15], producing HS images with 2–3 orders of magnitude higher spatial resolution (in each direction). The RIP diffuser method [18] suffers from high degree of blur, failing to produce sharp images even at medium resolutions. On the other hand, the newer work of Baek et al. [17] produces comparable spatial resolution, but is computationally much more expensive, relies on an alternative dispersion element, and requires presence of strong edges in the image due to the properties of the reconstruction algorithm. Similarly, the method of Habel et al. [24] achieves spatial and spectral resolutions comparable to our method, but requires more complex optical device and camera-specific calibration. Furthermore, the field-of-view of their method is also limited because of the relatively small square aperture used as a field stop.

3 Methods

3.1 CNNs for diffraction-based HS imaging

In this section, we describe the *Wide Dilation Network* model for constructing hyperspectral images. The model takes as input a RGB image $I_d \in \mathbb{R}^{w \times h \times 3}$ taken with a diffraction grating filter attached to the camera, and produces a tensor $I_{\text{hs}} \in \mathbb{R}^{w' \times h' \times N_\lambda}$, providing spectral information for each pixel. In our experiments (detailed in Sect. 4.1), we use input images of size 526×526 with three color channels and output HS images of size 384×384 with 102 spectral channels, but the approach is generally applicable for all resolutions and is limited only by the resolution of the ground truth HS images available for training the model.

The model, illustrated in Fig. 2, is an instance of convolutional neural networks. It builds on the ResNet architecture [21], but replaces standard convolutions in the first layer with a novel dilated convolution scheme designed specifically to account for the characteristic properties of light diffraction. In the following, we will first explain the convolution design and then proceed to provide a loss function optimized for re-construction of HS images. Finally, we discuss technical decisions related to the dilated convolutions and explain how simple upscaling of the dilation rates can be used for increasing the result of the output image.

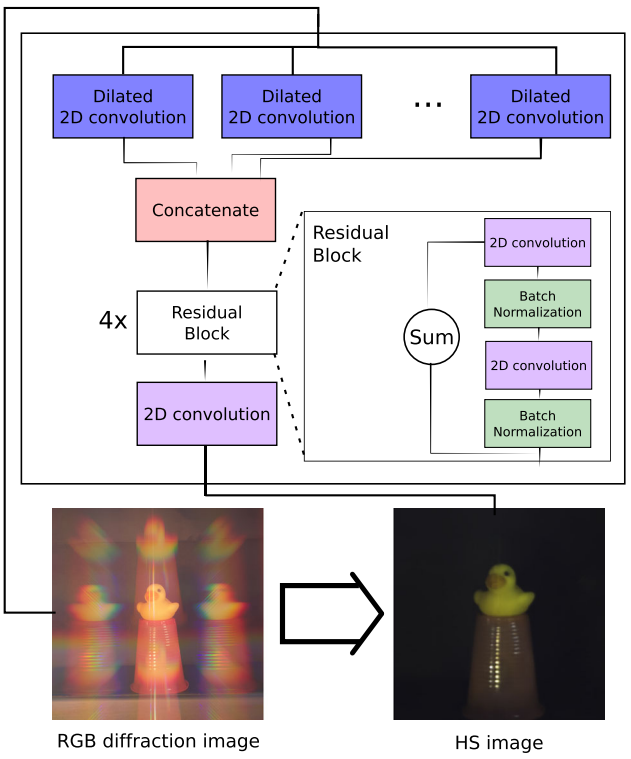

Fig. 2 Wide dilation network for HS image reconstruction. We employ dilated convolutions of different sizes along with multiple stacked residual blocks that each contains 2D convolutions and batch normalization, followed by 1×1 convolution for final reconstruction of the spectrum for each pixel. The model takes as input the diffraction image ($w \times h \times 3$ tensor), and outputs a hyperspectral image: a ($w' \times h' \times 102$) tensor, shown here as a RGB summary created using the CMF of Fig. 5. In our experiments, $w = h = 526$ and $w' = h' = 384$

3.2 Convolution design

Figure 3 (left) shows an image of a narrow-band laser projected at a dark background, taken through a diffraction grating filter. The laser is projected at a single point, but the first-order diffraction pattern of the grating filter disperses it to eight other positions as well, one in each major direction. The specific locations depend on the wavelength of the light, and for a narrow band laser are clearly localized. The first layer of our CNN is motivated by this observation. To capture the diffraction pattern, we need a convolutional filter that covers the entire range of possible diffraction distances yet retains sparsity to map the specific spatial dispersions to the right spectral wavelengths. This can be achieved with a set of *dilated convolutions* [25,26] with exceptionally wide dilation rate: this allows representing long-range dependencies with few parameters. More specifically, we use simple 3×3 kernels, with dilation rate d yielding an effective convolution window of $2d + 1$ in both directions with just 9 parameters per kernel. With 3×3 kernels we can capture the zeroth and first-order diffraction components, assuming d is selected suitably. The required d depends on the wavelength, and to cover all wavelengths within a specific range we need to introduce convolutions with varying $d \in [D_{\text{min}}, D_{\text{max}}]$. For each d in this range, we learn 5 filters, resulting in total $5\,(D_{\text{max}} - D_{\text{min}} + 1)$ filters. The required range of dilation

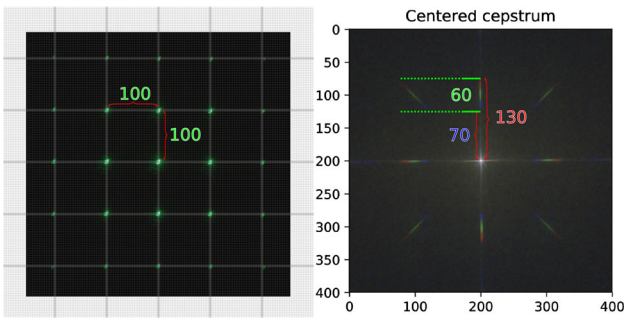

Fig. 3 (Left) Dilated convolution overlaid over a photograph of a single-point narrow-band (532 ± 10 nm) laser on a dark surface taken through a diffraction grating in a darkened room. Already a 3×3 convolution with dilation rate of 100 captures the first-order diffraction pattern. Due to the high intensity of the laser, we see also the second-order diffractions, which are typically not visible in real images. (Right) averaged, center shifted cepstrum for a random sample of 40 diffraction photographs, cropped to center where the diffraction pattern is evident. The range of dilation rates required for modeling the diffraction is revealed by the vertical (or horizontal) distance in pixels from the center to the first (D_{min}, blue) and last (D_{max}, red) maxima (color figure online)

rates can be determined based on the acquired diffraction images directly, with a process described in Sect. 3.4; for our setup we end up using 60 values for d and hence 300 filters in total.

Even though the convolutional layer can model the basic properties of diffraction using the wide dilations, the resulting linear combination is not sufficient for constructing the HS image due to nonlinearities of the imaging setup. We correct for this by forwarding the output to four consecutive layers that combine standard 2D convolutional layers with batch normalization and an additive residual connection, modeled after the residual blocks used for single-image super-resolution by Ledig et al. [20]. Each residual block consists of a sequence of 2D convolution, batch normalization, the Swish [27] activation function $\frac{y}{1+e^{-y}}$, 2D convolution, batch normalization, and a skip (identity) connection. The choice can be partially explained by the similarity of the two tasks: super-resolution techniques produce spatially higher resolution images, whereas we expand the spectral resolution of the image while keeping the spatial resolution roughly constant. For both cases, the residual connections help in retaining high visual quality. The final 1×1 convolution at the end of the network collapses the 300 channels into the desired number of spectral channels, here 102.

3.3 Loss function

To properly optimize for the quality of the reconstructed hyperspectral images, we construct a specific loss function by mixing a metric for RGB images with one for spectral distance. For high-quality HS images, we require that:

(a) The spectrum for each pixel of the output should match the ground-truth as closely as possible.
(b) Each spectral slice of the output should match the ground truth as a monochrome image.

The criterion (a) is critical for many applications of HS imaging that rely on the distinct spectral signatures of different materials [28]. Following the comprehensive study of spectral quality assessment [29], we employ the Canberra distance measure between the spectra of each pixel. Denote by $\hat{\mathbf{y}}$ and $\mathbf{y}$ the reconstructed and ground truth spectra for one pixel, and by λ a channel corresponding to a narrow wavelength band of the spectrum. The Canberra distance between the two spectra is then given by

$$d_{\mathrm{Can}}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{\lambda} \frac{|\hat{\mathbf{y}}_\lambda - \mathbf{y}_\lambda|}{\hat{\mathbf{y}}_\lambda + \mathbf{y}_\lambda}, \tag{1}$$

which should be small for all pixels of the image.

To address criterion (b), we employ the structural similarity measure SSIM [30], frequently used for evaluating similarity of RGB or monochrome images. To compute the similarity between the reconstructed and ground truth images, we slide a Gaussian window of size 11×11 through both images, and for each window compute the quantity

$$\mathrm{S}_{\hat{\mathbf{w}},\mathbf{w}} = \frac{(2\mathrm{E}[\hat{\mathbf{w}}]\mathrm{E}[\mathbf{w}] + c_1)(2\mathrm{Cov}[\hat{\mathbf{w}}, \mathbf{w}] + c_2)}{(\mathrm{E}[\hat{\mathbf{w}}]^2 + \mathrm{E}[\mathbf{w}]^2 + c_1)(\mathrm{Var}[\hat{\mathbf{w}}] + \mathrm{Var}[\mathbf{w}] + c_2)},$$

where $\hat{\mathbf{w}}$ and $\mathbf{w}$ are windows of the two images, c_1, c_2 are constants added for numerical stability, and $\mathrm{E}[\cdot]$, $\mathrm{Var}[\cdot]$, and $\mathrm{Cov}[\cdot]$ denote expectation (mean), variance and covariance, respectively. This quantity is computed for each spectral channel separately, and averaged to produce the SSIM index

$$\mathrm{SSIM}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N_{\hat{\mathbf{y}}} N_{\mathbf{y}}} \sum_{\hat{\mathbf{w}} \in \hat{\mathbf{y}}} \sum_{\mathbf{w} \in \mathbf{y}} S_{\hat{\mathbf{w}},\mathbf{w}}.$$

Here the sums loop over all the windows $\hat{\mathbf{w}} \in \hat{\mathbf{y}}$ and $\mathbf{w} \in \mathbf{y}$ and where $N_{\hat{\mathbf{y}}}$ and $N_{\mathbf{y}}$ are the number of windows in $\hat{\mathbf{y}}$ and $\mathbf{y}$, respectively.

Our final loss (to be minimized) simply combines the two terms by subtracting the SSIM index from the Canberra distance:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{h} \sum_{w} d_{\mathrm{Can}}(\hat{\mathbf{y}}, \mathbf{y}) - \mathrm{SSIM}(\hat{\mathbf{y}}, \mathbf{y}). \tag{2}$$

While we could here add a scaling factor to balance the two terms, our empirical experiments indicated that the method is not sensitive to the relative weight and hence for simplicity we use unit weights.

The model is trained using straightforward stochastic gradient descent with Adamax [31] as the optimization algorithm, using a single Nvidia Tesla P100 GPU. The model training consumed approximately 10 h.

3.4 Selection of dilation rates

The dilation range $[D_{\min}, D_{\max}]$ of the filters needs to be specified based on the range of the diffraction. This range depends on the imaging setup—mainly on the camera, lens, chosen resolution, and on the specific diffraction grating used. Importantly, however, it does not depend on the distance from the image target, or other properties of the scene. The dilation range needs to be wide enough to cover the first-order diffraction pattern, but not too wide as not to introduce excess parameters.

The range can be determined based on a diffraction photograph of a broadband but spatially narrow light source. A suitable lamp, ideally an incandescent filament lamp, placed behind a small opening would reveal the extent of diffraction. The range would then be determined by the pixel differences

between the light source and first and last diffraction components of the first-order diffraction pattern.

Alternatively, we could estimate the range by using two lasers corresponding to extreme wavelengths pointed at the camera.

It turns out, however, that the dilation range can also be determined without a separate calibration step, which makes the approach less sensitive to the imaging setup. We use the log magnitude of the cepstrum $\log(|\mathcal{C}(I)|)$, where $\mathcal{C}(I) = \mathcal{F}_{2D}^{-1}(\log(|\mathcal{F}_{2D}(I)|))$ and $\mathcal{F}$ is the Fourier transform, to extract periodic components from the frequency domain. To reduce the noise and for easy visual identification of the dilation range, we average the log magnitude of the cepstrum over multiple photographs.

Figure 3 (right) shows the averaged cepstrum for randomly selected 40 diffraction photographs, revealing the diffraction range that corresponds to the dilation rate range required for modeling the diffraction pattern.

To see why this works, we can think of diffraction photographs to have been formed by shifted and attenuated copies of infinitesimally narrow wavelength bands of the scene summed onto the scene. For the first-order diffraction components, a shifted copy is summed in a total of eight major directions for each narrow wavelength band. The amount of shift is a function of wavelength and assumed to be linearly proportional to the wavelength. The visible spectrum of light forms a continuum for which we wish to discover the range of the shifted and attenuated copies of the scene. To find this range, we make use of the "duplicate function" property of the cepstrum, explained in [32]. The shifted copies, duplicates of narrow wavelength bands of the original scene, will be visible in the cepstral domain as impulses, located at the shifting vector relative to the center as seen in Fig. 3 (right).

The computational cost of estimating the dilation rate range from the cepstrum is low, and in practice we only need a few images to see a clear range. This can be carried out on the same images that are used for training the model.

3.5 Dilation upscaling

Our method allows us to perform hyperspectral reconstruction on higher-resolution images than the ones the model was trained on. We achieve this by feeding in diffraction images at higher resolution (anyway available because the diffraction images are acquired with high-resolution DSLR) and increasing the dilation rates of the first layer by a constant scale factor $s \in \mathbb{N}$, so that for every dilation rate d_n we use the rate sd_n. This provides HS images s times larger spatially than the ones the model was trained on, without additional training. See Fig. 7 for a visual evaluation of the procedure.

4 Materials and evaluation

4.1 Data collection

For training and evaluating the model, we collected pairs of (a) hyperspectral images, in the spectral range of 400–1000 nm, and (b) RGB images captured with the diffraction grating element. The HS images were captured using a Specim IQ mobile hyperspectral camera, which captures 512×512 pixel images with 204 spectral bands. The integration time, the time to capture a single vertical column of pixels, for each HS image was 35 ms, resulting in total image acquisition time of 18 s (followed by tens of seconds of image storage and postprocessing). The last 102 spectral bands (corresponding to the 700–1000 nm range) of the HS images were discarded as these are in the near infrared range that our digital RGB camera filters out.

The imaging setup consists of a slide where the cameras are mounted (Fig. 1). The slide enables alternating the location of the HS camera and the camera for diffraction imaging so that images are captured from the same location. The RGB images were captured using a Canon 6D DSLR with a 35-mm normal lens together with a custom made diffraction grating filter mounted in front of the lens.

We used a transmitting double axis diffraction grating. Photographs were captured at a resolution of 5472×3648, but were cropped to 3000×3000 because the diffraction grating mount has a square aperture which causes vignetting. Each photograph was captured with an exposure time of 1/30 s, aperture value of 9.0 f and 500 ISO speed setting. The aperture value was selected to reduce the blurring effect caused by the vignetting of the window in the diffraction grating.

The hyperspectral images and the diffraction photographs were preprocessed by cropping and aligning. The hyperspectral images were center cropped to remove some of the unwanted background. The diffraction photographs were first downsampled to match the hyperspectral images' scale and then slightly rotated (with common rotation angle for every photograph) to account for slight bending of the camera assembly, caused by the weight of the cameras in the opposite ends of the camera assembly. The bending is estimated to have caused a shift and rotation about the imaging axis of at most 5 mm and 2°, respectively. Finally, the diffraction photographs were cropped to 526×526 for training, matching the scene with the HSI.

Finally, pairs of hyperspectral images and photographs were translated with respect to each other using template matching [33], where RGB reconstructions of the hyperspectral images were used as the templates. Compensation for distortion by means of transforming image pairs using camera extrinsic calibration was not necessary, because only center parts of images were used resulting in mostly distortion free images. We collected 271 pairs of diffraction and

hyperspectral images, of which 32 were used only for evaluation. The images were taken indoors under multiple artificial light sources. The subject of the images is a collection of toys, different colored blocks, books, leaves, and other small items against a dark background. The objects were placed mainly in the lower center area of the images.

4.2 Model variants

Our model employs three separate elements that are required for constructing high-quality HS images: (a) the wide dilation layer in the beginning of the network, (b) the residual blocks for modeling nonlinearities, and (c) the loss function ensuring good spatial and spectral characteristics.

To verify the importance of the residual blocks for correcting nonlinearities in the imaging setup, we compare the proposed model against one without the residual blocks, directly connecting the convolutional layers to the output. For the full model, the number of residual blocks was selected between 1 and 8 using standard cross-validation, resulting in the final choice of four blocks. Similarly, to demonstrate the importance of modeling both the spatial reconstruction quality using SSIM and the spectral reconstruction quality using Canberra distance in the combined loss Eq. (2), we compare against the proposed model optimized for each term alone.

Finally, we could also consider alternatives for the convolutional layer, which needs to access information tens or hundreds of pixels away (130 for the specific diffraction grating used in our experiments) to capture the first-order diffraction grating pattern. One can imagine two alternative ways of achieving this without wide dilated convolutions. One option would be to use extremely large (up to 260×260) dense convolutions, computed using FFT [34]. However, this massively increases the number of parameters and the model would not be trainable in practice. The other option would be to stack multiple layers with small convolution filters and use pooling to reduce the receptive field, but maintaining high spatial resolution would be tremendously difficult. Consequently, we did not conduct experiments with alternative convolution designs.

4.3 Evaluation metrics

We evaluate our method using the dataset collected as described in Sect. 4.1. We split the dataset into two parts, 239 images for training and 32 for evaluation. All results presented are computed on the test set. We evaluate the reconstruction quality using the two parts of our loss function Eq. (2), SSIM for visual quality and Canberra distance for the spectral quality, and additionally with three independent metrics not optimized for: mean square error (MSE) and mean absolute error (MAE) for overall quality, and spectral angle to compare spectral similarity [29].

5 Results and discussion

We compare the proposed model against the baselines described in Sect. 4.2, summarizing the results in Table 1. The "SSIM" and "Canberra" rows correspond to optimizing for only SSIM or only Canberra. The proposed wide dilation network clearly outperforms the baselines that omit critical components. The effect of omitting residual blocks is clear, but not dramatic, and the model trained to minimize Eq. (2) outperforms variants optimizing for SSIM and Canberra alone, even when the quality is measured on the specific loss itself. This is strong indication of the combination being a useful learning target.

Besides the numerical comparisons, we explore the quality of the reconstructed HS images visually. Figure 4 demonstrates both spatial and spectral accuracy for a randomly selected validation image. For validating spatial accuracy, we collapse the HS image back into RGB image by weighted summation over the spectral channels with weights show in Fig. 5). The accuracy of spectral reconstruction is studied by comparing spectra associated with individual pixels against the ground truth. In summary, the visualization reveals the method accurately reconstructs the HS image, but not without errors. The RGB summaries and individual spectral slices represent the spatial characteristics of the scene well, and the main spectral characteristics are correct even though the actual intensities deviate somewhat from the ground truth.

Comparing the properties of narrow wavelength band monochromatic images (Fig. 4) with the results presented in [18], we note that our method produces distinctly less blurry images and the spectra of individual pixels follow much more closely the ground truth. Visually the narrow wavelength band monochromatic images in [17] appear on par with images produced by our method, although our method produces over four times the number of channels. In contrast

Table 1 The proposed wide dilation network outperforms all three baselines (Sect. 4.2) according to five different metrics: the mean squared error (MSE, scale 10^{-10}), mean absolute error (MAE), Canberra distance, spectral angle of each pixel), and SSIM

Method	MSE	MAE	Canberra	Angle	SSIM
No residuals	0.89	0.014	0.050	0.070	0.9611
Only SSIM loss	0.88	0.013	0.048	0.057	0.9716
Only Canberra loss	1.05	0.010	0.032	0.048	0.9718
Full method	0.49	0.008	0.028	0.045	0.9800

For all metrics lower is better, except for SSIM where higher is better. The methods listed are (1) no residual blocks, (2) only SSIM loss, (3) only Canberra loss, and (4) all of the previous combined. Importantly, the proposed method outperforms direct optimization of SSIM and Canberra distance also when measured using the optimization criterion itself

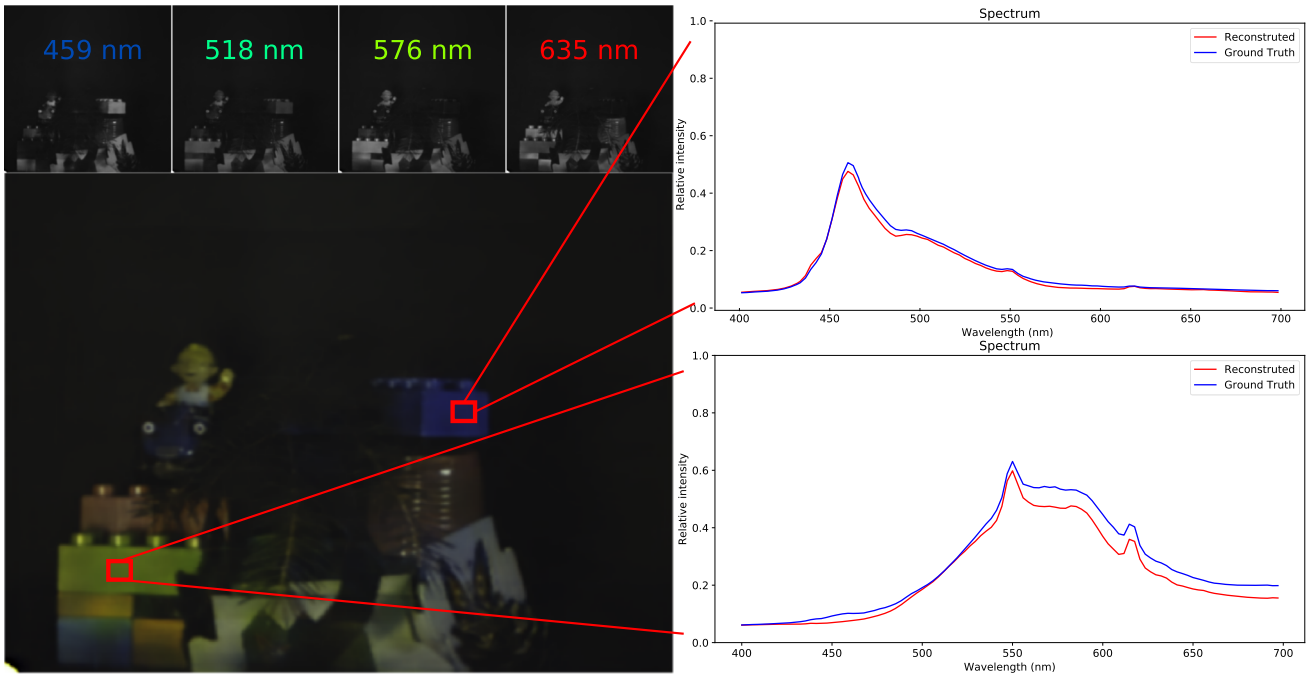

Fig. 4 (Left): illustration of a prototypical reconstructed HS image, represented as RGB summary and as four individual channels in top row, demonstrating high fidelity of both the RGB summary and individual channels. (Right): comparison of pixel-wise spectra of the reconstruction and ground truth. While the reconstruction is not exact, it matches the key characteristics well for the whole spectral range

to [17], the diffraction grating required by our method weighs less and our method is computationally much faster.

We further analyze the reconstruction quality by error analysis, separately for the spectral and spatial characteristics. Figure 6 (top) presents the average spectrum over all validation images, indicating good match between the reconstruction and the ground truth, with a slight bias toward longer wavelengths.

For analyzing the spatial distribution of errors, we divide the images into 15×15 areas of 26×26 pixels each, and compute the mean errors for those [Fig. 6 (bottom)]. The errors are larger in the lower bottom half of the image, which is where the objects were mostly located. Consequently, we note that the quantitative evaluation in Table 1 characterizes the quality of the images only in the area where the objects were placed, because we cannot accurately evaluate the output of the network for images in areas of constant background in all available images. The largest objects in the synthesized images are approximately 286×338, resulting in $858 \times 1,014$ high-quality synthesized images with threefold dilation upscaling. The neural network itself is agnostic of the image content and could be re-trained on images covering the whole area. Both training and evaluation time would remain the same, and we expect the accuracy to remain similar.

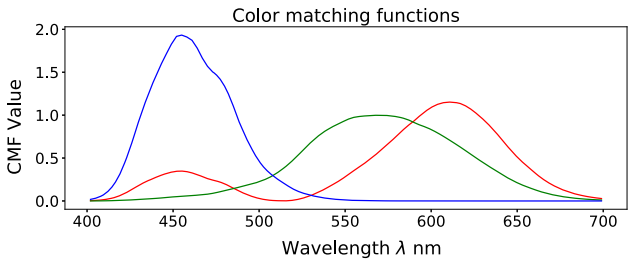

Fig. 5 Color matching function (CMF) values used for red, green and blue for the visible spectrum, used for collapsing a hyperspectral image into RGB image for visual comparison. The CMF values are based on the intensity perception to a particular wavelength for particular cone cell type (long, medium and short) for an typical human observer (color figure online)[36]

Finally, we demonstrate reconstruction of higher resolution HS images using dilation upscaling and high-resolution diffraction images. Figure 7 presents an example with $2\times$ and $3\times$ increase in resolution in both directions. The scaled-up images are clearly more sharp, but start to exhibit artifacts such as color bleeding. This is in part due to slight rotation present in the original image pairs, and in part due to the residual blocks being trained on lower-resolution images.

We also note that our data were collected under constant lighting conditions, and hence the model would not directly generalize for arbitrary environments. However, this can be remedied by training the model on a larger data set with more heterogeneity in the lighting spectrum. Collecting such data is feasible since our results indicate that already a relatively small number of images taken in each context is sufficient.

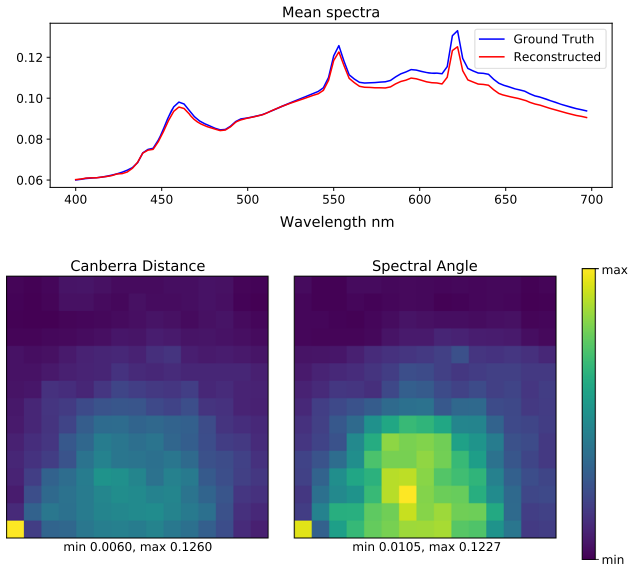

Fig. 6 (Top) Average spectrum for the test set images for ground truth (blue) and reconstructed (red) HS images shows that there is a slight bias in average intensity toward the end of the spectrum. The peaks correspond to the spectrum of the light sources. (Bottom) distribution of error by spatial location, summarized for square blocks of 26 by 26 pixels. The bottom left corner is a minor artifact of the imaging setup, and otherwise the error correlates with the placement of the objects; the top of the images was mostly background (color figure online)

Further, our experimental setup is limited to the visible light spectrum and does not account for the near-infrared wavelengths that the Specim IQ camera [2], and most other HS cameras, capture. This is because the camera we used, as most digital cameras, filters out the infrared wavelengths. There is no reason to believe our method would not generalize to the near-infrared range (approximately 700-1000nm by simply using a modified DSLR with the infrared filter removed. Extending the approach for ultraviolet range (below 400 nm) would, however, require using different sensors, since standard digital cameras have extremely low sensitivity in that range. Finally, we have not studied how the choice of the lens affects results, but we suspect that the residual network could learn to compensate for effects of, for example, chromatic aberration.

6 Conclusion

We have presented a practical, cost-effective method for acquiring hyperspectral images using a standard digital camera accompanied with a diffraction grating and a machine learning algorithm trained on pairs of diffraction grating and ground truth HS images. Our solution can be applied to almost any type of digital camera, including smartphones. Even though the idea of reconstructing hyperspectral images by combination of a computational algorithm and a passive filter is not new [14,17,18,24], our approach is the first one that can provide snapshot images of sufficient spatial and spectral dimensions in less than a second.

We showed that it is possible to generate high-quality images based on a very small data set, thanks to a model inspired by physical properties of diffraction yet trained end-to-end in a data-driven manner. The resulting images capture the spatial details and spectral characteristics of the target faithfully, but would not reach the accuracy required for high-precision scientific measurement of spectral characteristics. This is perfectly acceptable for a wide range of applications of HS imaging; tasks such as object classification, food quality control [6] or foreign object detection[35] would not be affected by minor biases and noise of our reconstruction algorithm. Hence, our solution provides tangible cost benefits in several HS imaging applications, while opening up new ones due to high spatial resolution (with further up-scaling with built-in super-resolution) and portability. Further, many

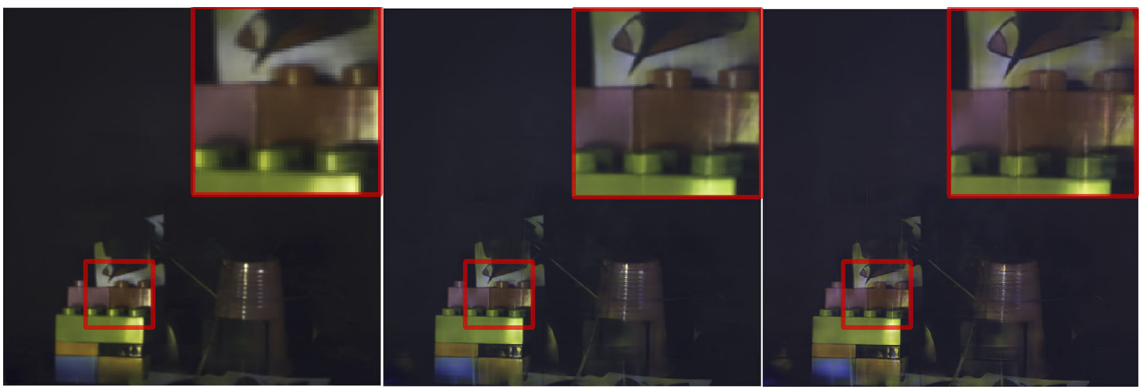

Fig. 7 Illustration of dilation upscaling for producing HS images of resolution higher than what was available for training. (Left) 384 × 384 image corresponding to the size of the training images, presented as RGB summary of the reconstruction. (Middle and right) the same image in resolutions of 768 × 768 and 1152 × 1152, respectively. The upscaled images are sharper, but start to suffer from artifacts and bleeding

existing machine learning methods that have been developed for satellite images, such as [4,5], can now be used on scenes taken on the ground.

Acknowledgements We thank the Academy of Finland (Grant 1266969) for partial funding, and the Finnish Grid and Cloud Infrastructure (urn:nbn:fi:research-infras-2016072533) for computational resources.

Funding Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

1. Grahn, H., Geladi, P.: Techniques and Applications of Hyperspectral Image Analysis. Wiley, New York (2007)
2. Behmann, J., Acebron, K., Emin, D., Bennertz, S., Matsubara, S., Thomas, S., Bohnenkamp, D., Kuska, M., Jussila, J., Salo, H., et al.: Specim IQ: evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. Sensors **18**(2), 441 (2018)
3. Manolakis, D., Shaw, G.: Detection algorithms for hyperspectral imaging applications. IEEE Signal Process. Maga. **19**(1), 29–43 (2002)
4. Makantasis, K., Karantzalos, K., Doulamis, A., Doulamis, N.: Deep supervised learning for hyperspectral data classification through convolutional neural networks. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4959–4962. IEEE (2015)
5. Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y.: Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**(6), 2094–2107 (2014)
6. Gowen, A., O'Donnell, C., Cullen, P., Downey, G., Frias, J.: Hyperspectral imaging-an emerging process analytical tool for food quality and safety control. Trends Food Sci. Technol. **18**(12), 590–598 (2007)
7. Ellis, J.M., Davis, H., Zamudio, J.A.: Exploring for onshore oil seeps with hyperspectral imaging. Oil Gas J. **99**(37), 49–58 (2001)
8. Liu, Z., Yan, J.-Q., Zhang, D., Li, Q.-L.: Automated tongue segmentation in hyperspectral images for medicine. Appl. Opt. **46**(34), 8328–8334 (2007)
9. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. J. Biomed. Opt. **19**(1), 010901 (2014)
10. Wong, G.: Snapshot hyperspectral imaging and practical applications. In: Journal of Physics: Conference Series, vol. 178, no. 1, p. 012048. IOP Publishing (2009)
11. Geelen, B., Tack, N., Lambrechts, A.: A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. In: Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII, vol. 8974, p. 89740L. International Society for Optics and Photonics (2014)
12. Guo, B., Näsilä, A., Trops, R., Havia, T., Stuns, I., Saari, H., Rissanen, A.: Wide-band large-aperture Ag surface-micro-machined MEMS Fabry–Perot interferometers (AgMFPIs) for miniaturized hyperspectral imaging. In: MOEMS and Miniaturized Systems XVII, vol. 10545, p. 105450U. International Society for Optics and Photonics (2018)
13. Wang, L., Xiong, Z., Huang, H., Shi, G., Wu, F., Zeng, W.: High-speed hyperspectral video acquisition by combining Nyquist and compressive sampling. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 857–870 (2019)
14. Okamoto, T., Yamaguchi, I.: Simultaneous acquisition of spectral image information. Opt. Lett. **16**(16), 1277–1279 (1991)
15. Descour, M., Dereniak, E.: Computed-tomography imaging spectrometer: experimental calibration and reconstruction results. Appl. Opt. **34**(22), 4817–4826 (1995)
16. Johnson, W.R., Wilson, D.W., Fink, W., Humayun, M.S., Bearman, G.H.: Snapshot hyperspectral imaging in ophthalmology. J. Biomed. Opt. **12**(1), 014036 (2007)
17. Baek, S.-H., Kim, I., Gutierrez, D., Kim, M.H.: Compact single-shot hyperspectral imaging using a prism. ACM Trans. Graph. (TOG) **36**(6), 217 (2017)
18. Golub, M.A., Averbuch, A., Nathan, M., Zheludev, V.A., Hauser, J., Gurevitch, S., Malinsky, R., Kagan, A.: Compressed sensing snapshot spectral imaging by a regular digital camera with an added optical diffuser. Appl. Opt. **55**(3), 432–443 (2016)
19. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: et al., Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Hagen, N.A., Kudenov, M.W.: Review of snapshot spectral imaging technologies. Opt. Eng. **52**(9), 090901 (2013)
23. Goldstein, T., Osher, S.: The split Bregman method for L1-regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
24. Habel, R., Kudenov, M., Wimmer, M.: Practical spectral photography. In: Computer Graphics Forum, vol. 31, no. 2pt2, pp. 449–458. Wiley Online Library (2012)
25. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions (2015). ArXiv preprint arXiv:1511.07122
26. Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning (2016). ArXiv preprint arXiv:1603.07285
27. Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function, vol. 7 (2017). ArXiv preprint arXiv:1710.05941
28. Heinz, D.C., et al.: Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **39**(3), 529–545 (2001)
29. Deborah, H., Richard, N., Hardeberg, J.Y.: A comprehensive evaluation of spectral distance functions and metrics for hyperspectral image processing. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**(6), 3224–3234 (2015)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
31. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). ArXiv preprint arXiv:1412.6980
32. Rom, R.: On the cepstrum of two-dimensional functions (corresp.). IEEE Trans. Inf. Theory **21**(2), 214–217 (1975)

33. Briechle, K., Hanebeck, U.D.: Template matching using fast normalized cross correlation. In: Optical Pattern Recognition XII, vol. 4387, pp. 95–102. International Society for Optics and Photonics (2001)
34. Mathieu, M., Henaff, M., LeCun, Y.: Fast training of convolutional networks through FFTS (2013). ArXiv preprint arXiv:1312.5851
35. Guo, J., Ying, Y., Li, J., Rao, X., Kang, Y., Shi, Z.: Detection of foreign materials on surface of ginned cotton by hyper-spectral imaging. Trans. Chin. Soc. Agric. Eng. **28**(21), 126–134 (2012)
36. Fairchild, M.D.: Color Appearance Models. Wiley, New York (2013)

Mikko E. Toivonen is a PhD student at the Department of Computer Science, University of Helsinki, Finland. He received his MSc in Communications Engineering from the Helsinki University of Technology, Finland, in 2007. He is researching machine learning applied to hyperspectral imaging.

Chang Rajani is a PhD student at the Department of Computer Science, University of Helsinki. He received his MSc in Computer Science from the University of Helsinki in 2018. He works on Bayesian deep learning with high-dimensional physical data.

Arto Klami received MSc (2003) and PhD (2008) degrees (with distinction) in computer science from Helsinki University of Technology, Department of Computer and Information Science, Finland, and the title of docent in Information and Computer Science at Aalto University, Finland, in 2013. He is currently Assistant Professor at University of Helsinki, Department of Computer Science, and a member of Helsinki Institute for Information Technology HIIT and Finnish Center for Artificial Intelligence FCAI. His main research area is statistical machine learning and Bayesian inference, with applications in computational physics, spectral imaging, and modelling human behaviour. He has also contributions in data integration, computational biology, human–computer interaction, and computational neuroscience. He has authored more than 60 scientific articles (H-index 23 GS) on these topics.