Int J Comput Vis (2014) 110:141–155
DOI 10.1007/s11263-013-0690-4

Acquisition of High Spatial and Spectral Resolution Video with a Hybrid Camera System

Chenguang Ma · Xun Cao · Xin Tong · Qionghai Dai · Stephen Lin

Received: 28 February 2013 / Accepted: 30 November 2013 / Published online: 13 December 2013

Abstract We present a hybrid camera system for capturing video at high spatial and spectral resolutions. Composed of an red, green, and blue (RGB) video camera, a grayscale video camera and a few optical elements, the hybrid camera system simultaneously records two video streams: an RGB video with high spatial resolution, and a multispectral (MS) video with low spatial resolution. After registration of the two video streams, our system propagates the MS information into the RGB video to produce a video with both high spectral and spatial resolution. This propagation between videos is guided by color similarity of pixels in the spectral domain, proximity in the spatial domain, and the consistent color of each scene point in the temporal domain. The propagation algorithm, based on trilateral filtering, is designed to rapidly generate output video from the captured data at frame rates fast enough for real-time video analysis tasks such as tracking and surveillance. We evaluate the proposed system using both simulations with ground truth data and on real-world scenes. The accuracy of spectral capture is examined through comparisons with ground truth and with a commercial spectrometer. The utility of this high resolution MS video data is demonstrated on the applications of dynamic white balance adjustment, object tracking, and separating the appearance contributions of different illumination sources. The various high resolution MS video datasets that we captured will be made publicly available to facilitate research on dynamic spectral data analysis.

Keywords Multispectral video capture · Hybrid camera system

1 Introduction

Within the human eye, light from the real world is sensed in color. The human eye perceives color from three types of cone cells sensitive to different parts of the light spectrum, namely those corresponding to red, green, and blue (RGB). Conventional cameras record RGB measurements that adequately replicate colors for human viewing, but in fact the spectrum of visible light may contain a profusion of detail that is lost in the coarse three-channel sensing of RGB. A multispectral (MS) imager, by contrast, measures at each pixel a finely sampled spectrum that may extend beyond the visible range. The details within a high resolution spectrum can reveal much about the objects and lighting in the scene, and computer vision algorithms may thus have much to gain by measuring tens or hundreds of color samples over the spectrum of each scene point.

MS image capture has drawn much attention in the past several years. For imaging of static scenes, systems have been designed to record high spectral resolution at the expense of acquisition time. Spectrometers (James 2007) employ dispersive optical elements that require scanning over the scene to capture a full image. Other methods utilize a sequence of

C. Ma · Q. Dai
Department of Automation, Tsinghua University, Beijing, China
e-mail: ChenguangMa2011@gmail.com

Q. Dai
e-mail: qionghaidai@tsinghua.edu.cn

X. Cao (✉)
School of Electronic Science and Engineering,
Nanjing University, Nanjing, China
e-mail: caoxun@nju.edu.cn

X. Tong · S. Lin
Microsoft Research Asia, Beijing, China
e-mail: xtong@microsoft.com

S. Lin
e-mail: stevelin@microsoft.com

bandpass color filters in front of a camera to measure different spectral components (Schechner and Nayar 2002; Yamaguchi et al. 2006). While both of these approaches provide direct measurements of scene spectra, they are unsuitable for MS imaging at video rates. To record dynamic scenes, reconstruction-based imaging techniques based on computed tomographic imaging spectrometry, CTIS (Descour and Dereniak 1995; Mooney et al. 1997) and coded aperture imaging (Brady and Gehm 2006; Wagadarikar et al. 2009) have been presented. These methods allow for MS imaging at video rates, but the considerable post-processing needed to reconstruct MS videos makes them unsuitable for real-time vision applications such as tracking and surveillance. Recently, a system based on a prism and occlusion mask was proposed for direct, real-time capture of MS video (Du et al. 2009). However, due to the sacrifice of spatial resolution for additional spectral resolution, the image frames contain little spatial detail, which may limit the ability to perform image analysis.

In this paper, we propose a solution for video capture with real-time MS output at both high spatial and spectral resolution. This solution is based on a hybrid camera system and a method to integrate the data from the two cameras to produce the output video, as illustrated in Fig. 1. The hardware configuration, shown in Fig. 2a, consists of a regular RGB video camera, a grayscale video camera and a few off-the-shelf optical elements including a prism, beam splitter, occlusion mask and planar mirror. Incoming light from the scene is first equally divided towards two directions by the beam splitter. In one direction, the light passes through the occlusion mask and prism to form dispersed spectra on the grayscale sensor, similar to the system in Du et al. (2009). While high spectral resolution is obtained in this manner, it is obtained with a significant loss in spatial resolution. Light directed along the other path is reflected by the mirror and captured by the RGB camera, which has low spectral resolution but high spatial resolution.

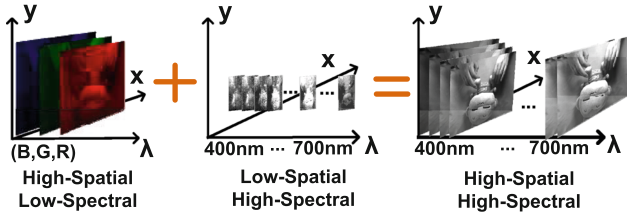


Fig. 1 Basic idea of multispectral video capture using the hybrid camera system. We integrate the high-spatial, low-spectral resolution frames from an RGB camera and the high-spectral, low-spatial resolution frames from a multispectral camera to generate videos with both high-spectral and high-spatial resolution. Temporal consistency in scene point *color* is also used, but not illustrated in this figure

The different types of data recorded by the two cameras are integrated as shown in Fig. 2b. Frames from the two cameras are aligned so that each scene point measured by the MS camera has a known corresponding pixel in the RGB camera. Captured MS data is then transferred to corresponding pixels in the RGB frames, and is propagated to other pixels according to spectral similarity, spatial proximity, and temporal color consistency of scene points tracked by optical flow (Brox et al. 2004). This spectral propagation is efficiently processed by trilateral filtering, whose fast implementation enables MS video generation in real time.

Unlike systems that tradeoff either temporal or spatial resolution for additional spectral information, the proposed approach does not require such sacrifices while maintaining high spectral accuracy. Different from reconstruction based systems, our capture device can be constructed from widely available components and is much simpler to calibrate in practice. The effectiveness of this system is demonstrated with experiments on different computer vision applications including dynamic white balance adjustment and object tracking.

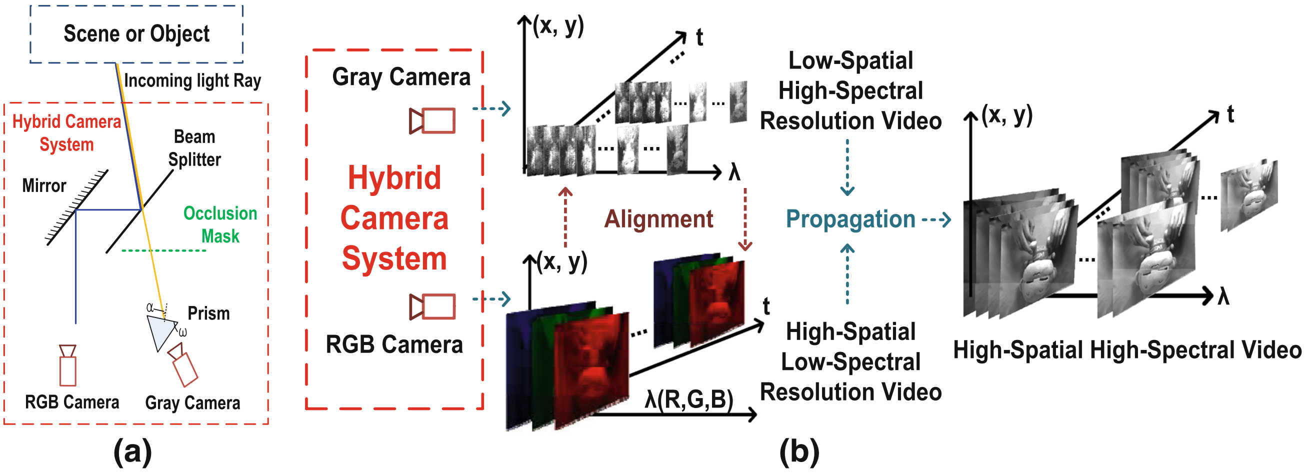


Fig. 2 Overview. (**a**) System configuration of the proposed hybrid camera system. (**b**) Pipeline for high-resolution multispectral video generation

An early version of this work appears in Cao et al. (2011b). In this paper, we include the following extensions and additional content:

- The spectral propagation algorithm is reformulated as a trilateral filter (TF) that propagates spectral information over the spatial, color and temporal domains. In comparison to the bilateral filter (BF) used in Cao et al. (2011b), the TF leads to higher accuracy in spectral propagation.
- Techniques are described to accelerate the computation of output video (by three orders of magnitude) and reduce data bandwidth, including the use of a principal components analysis (PCA) representation of spectra and graphics processing unit (GPU)-based parallelization and processing. Further details on system calibration and hybrid camera alignment are also presented.
- The accuracy of our MS camera is evaluated through comparisons to ground truth and to measurements obtained using a commercial scanning spectrometer.
- The application of user-assisted separation of mixed illumination is introduced, in which the spectral measurements from our system are used to separate the appearance contributions from different illumination sources.
- We build a dataset of high resolution MS videos, and propose a new file format based on our work for compact representation of MS images and video. Software is also developed for viewing and processing such files. Both the video dataset and software will be made publicly available to facilitate future research on processing dynamic spectral data.

2 Related Work

Depending on the target application, most existing methods for MS imaging forgo high spatial and/or temporal resolution to increase spectral sampling. For example, the spectrometer in James (2007) obtains a very high spectral resolution of 0.1 nm but measures only a single pixel at a time. Methods that employ multiple color filters, such as a rotating filter wheel in front of the camera (Yamaguchi et al. 2006) or different filters distributed over the sensor (Schechner and Nayar 2002), require multiple exposures to record different parts of the spectrum at each pixel. While such methods are effective for static scenes, they lack the efficiency needed to capture video of dynamic scenes.

Instead of directly measuring the spectral data of each scene point, CTIS systems (Volin 2000; Descour and Dereniak 1995; Johnson et al. 2006) and coded aperture methods (Brady and Gehm 2006; Wagadarikar et al. 2008, 2009) treat MS imaging as a reconstruction problem. They regard the two-dimensional (2D) spatial information plus one spectral dimension as a 3D datacube, and reconstruct the 3D datacube from a set of 2D projections. CTIS systems utilize 2D projections that integrate spectral signals from different scene positions on the detector. The multiple projections are recorded on the sensor in a single snapshot, which gives this method the potential to be used for MS video acquisition. However, CTIS systems need custom-made optical elements and are sensitive in practice to calibration noise. Only simulations and snapshots of very simple scenes have been reported in the literature (Hagen and Dereniak 2008; Vandervlugt et al. 2007; Habel et al. 2012). Coded aperture methods take multiplexed 2D projections of the 3D datacube using a specially designed aperture such as in the coded aperture snapshot imager (CASSI). The key idea of CASSI is to use apertures with certain patterns to code and decode the optical field. Recently, Wagadarikar et al. (2009) presented a video-rate CASSI system. A video of a lit candle was demonstrated, though with some reported reconstruction error. The design and performance of a coded aperture spectral imager with a wide spectral range and high spatial resolution is also described in Kittle et al. (2012). Noise and classification metrics for aperture codes in CASSI systems are compared in Mrozack et al. (2012) and multiplexing codes are also shown to be advantageous. In Kim et al. (2012), an end-to-end measurement system is introduced for capturing the spectral data of 3D objects, which is then used to facilitate spectral rendering and the study of avian vision. Although CTIS and CASSI systems have high spectral and temporal resolution, the spatial resolution is relatively low. Also, a time-consuming reconstruction step is necessary for both CTIS and CASSI, which makes them unsuitable for use in real-time video processing applications.

Other types of systems have also been presented for MS imaging with high temporal resolution. A device based on optical fiber bundles and a diffraction grating was proposed in Fletcher-Holmes and Harvey (2005) for real-time video capture, but provides only a 14 × 14 image resolution due to physical limitations on minimum fiber thickness. A few techniques have been proposed for recovering MS reflectance based on active illumination. Park et al. (2007) developed a system that captures six-channel video at 30 fps by using multiplexed coded illumination. Chi et al. (2010) used optimized wide band filtered illumination, and Han et al. (2010) employed a DLP projector for taking spectral measurements at 100 Hz. Illumination based capture of spectra, however, is limited in practice by the need for controlled lighting. The MS video capturing system presented in Darling et al. (2011) uses two different filters to extend RGB to a six-channel spectrum in real time, providing a limited increase in spectral resolution. In Du et al. (2009), a prism-based system for MS video acquisition was demonstrated with relatively low spatial resolution. Our acquisition system differs from previous methods in that it achieves high resolution in spectral, spatial and temporal domains, with real-time video generation.

3 System Overview

In this section, we first introduce the basic principles and configuration of the proposed hybrid camera system. The entire MS capture pipeline is then described including the propagation algorithm and its acceleration. Finally, image distortions and design tradeoffs are discussed.

3.1 System Configuration

A diagram of the hybrid camera system is shown in Fig. 2a. Incoming light from a scene point first reaches the beam splitter, which reflects half of the light along the blue path, while transmitting the remainder along the yellow path. The reflected light is again reflected at a mirror before arriving at the RGB camera. An alternative configuration for this light path is to remove the mirror and have the RGB camera directly facing the beam splitter, but this results in a less compact system. The light on this path is measured in RGB at a high spatial resolution.

The light transmitted through the beam splitter is dispersed by the prism onto the grayscale sensor, which measures numerous channels of its spectra. The resolution of the recorded spectrum is determined by the size of the sensor pixels and the width of the dispersed spectrum on the CCD plane, with a broader dispersion resulting in a higher spectral resolution. To avoid overlap on the sensor of spectra from different scene points, we employ an occlusion mask as in Du et al. (2009) to subsample the scene radiance. Because of this subsampling, the MS imaging at the grayscale camera is obtained at a low spatial resolution.

Triggering of the two cameras is synchronized, such that a low resolution MS video frame is captured simultaneously with each high resolution RGB video frame. The two cameras are aligned to capture the same view. Each sample point of the MS imager has a counterpart pixel in the RGB camera that shares the same incoming light ray, and these correspondences are determined as described in Sect. 4. The correspondences are used by the spectral propagation algorithm to produce high resolution MS video.

3.2 Spectral Propagation

The correspondence between MS samples and RGB pixels gives high resolution spectral values for certain pixels in each RGB frame, as illustrated in Fig. 3a for an RGB frame with a resolution of $p \times q$ pixels and a corresponding MS frame of resolution $m \times n$ $(m < p,\ n < q)$. To obtain high spectral resolution over the entire frame, we propagate the MS data to the other pixels according to color similarity, spatial proximity, and temporal consistency of scene points.

In our previous work (Cao et al. 2011b), spectral propagation was performed using a BF (Yang et al. 2009) that replaces the color of each pixel in the RGB frame (shown as blue in Fig. 3a) by a weighted average of spectra from nearby MS pixels (shown as orange). For a pixel $(i,\ j)$, its MS information is calculated as

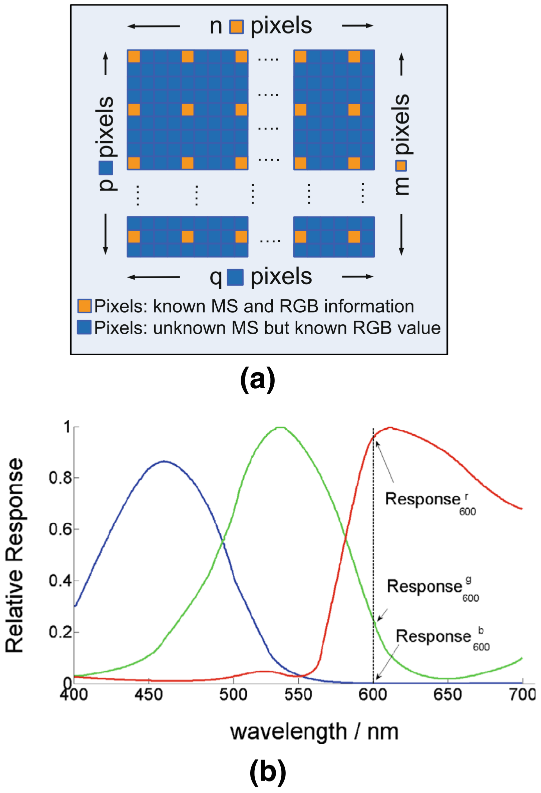


Fig. 3 (**a**) Distribution of multispectral pixels over an RGB frame. (**b**) Response curves of RGB filters

$$\boldsymbol{ms}_{ij} = \sum_{c\in R,G,B} \frac{\sum_{k\in\Omega} \mathcal{G}_{\sigma_r}(d_k^{RGB})\mathcal{G}_{\sigma_s}(d_k^{xy})\cdot\rho_k\cdot\boldsymbol{ms}_k^c}{\sum_{k\in\Omega}\mathcal{G}_{\sigma_r}(d_k^{RGB})\mathcal{G}_{\sigma_s}(d_k^{xy})}, \quad (1)$$

where $\boldsymbol{ms}_{ij}$ denotes the MS vector for pixel $(i,\ j)$, $k \in \Omega$ indexes the orange pixels within a neighborhood centered on $(i,\ j)$, $\mathcal{G}_\sigma(\cdot)$ represents the Gaussian operator with zero mean and variance σ, and d_k^{RGB} and d_k^{xy} denote the Euclidean distance between the pixels $(i,\ j)$ and k in RGB space and $(x,\ y)$-image space, respectively. In computing d_k^{RGB}, the RGB values of k in the original RGB frame are used. The factor ρ_k represents the ratio of a given color channel value at k to the corresponding value at $(i,\ j)$ (e.g., $\rho_k = R_{i,j}/R_k$ for the red channel) and is included for intensity matching.

In Eq. 1, we propagate MS data between pixels with similar RGB values, a strong indicator of similarity in scene radiance spectra. The propagation is done per-channel, so that slight intensity differences between source and target pixels in different parts of the spectrum can be accounted for using factor ρ. For this per-channel propagation, the MS data

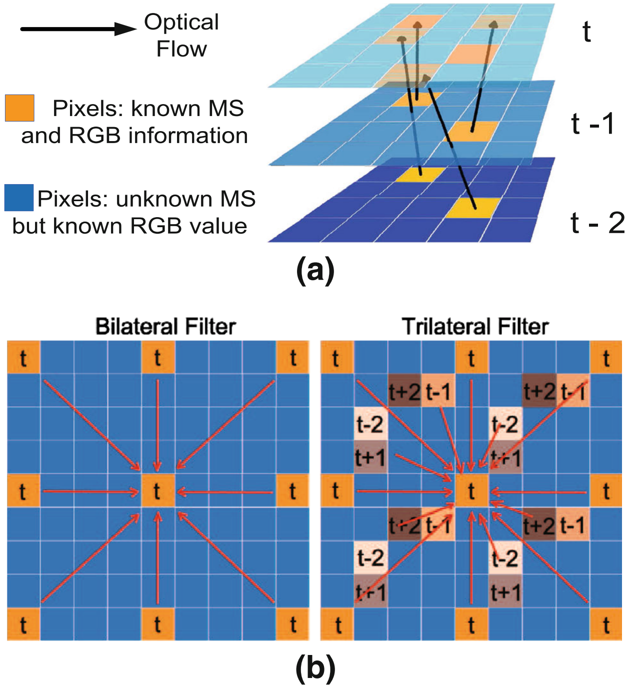

Optical Flow
🟧 Pixels: known MS and RGB information
🟦 Pixels: unknown MS but known RGB value
t
t -1
t - 2
(a)
Bilateral Filter
Trilateral Filter
t
t+2
t-1
t-2
t+1
(b)

Fig. 4 Temporally-based enhancement of spectral propagation. (**a**) For a frame at time instant t, we use temporal correspondences computed by optical flow from preceding frames to provide more multispectral data. (**b**) Our trilateral filter directly accounts for spectral information along the temporal dimension, in contrast to a 2D bilateral filter

is separated into RGB according to the relative responses of the camera's RGB filters at each wavelength (see Fig. 3b):

$$\boldsymbol{ms}_k^c = \boldsymbol{ms}_k \otimes \boldsymbol{w}^c \quad \text{for} \quad c = R,\ G,\ B, \tag{2}$$

where

$$w_\lambda^c = \frac{q_\lambda^c}{\sum_c q_\lambda^c}, \tag{3}$$

and q_λ^c denotes the response of filter c at wavelength λ. After propagating the per-channel MS data, the three channels are summed to obtain the full spectrum. An important difference between this propagation algorithm and traditional bilateral filtering is our use of intensity modulated spectra (i.e., from the ρ factor in Eq.1). This modulation is critical for our system to accurately handle subtle variations such as shading using the very sparsely sampled MS points.

Since our system captures video streams, MS data can be propagated temporally as well as spatially. The MS data recorded for a scene point in one frame is approximately valid for the same scene point in other frames captured close in time. As illustrated in Fig. 4a, optical flow (Brox et al. 2004) is adopted to track scene points for which MS data has been measured. The MS data of these points can then be propagated to other frames to increase the density of MS samples.

This temporally-based enhancement can lead to improved accuracy, especially for objects of interest in a video which are generally in motion and thus benefit more from spectral propagation between video frames. In Cao et al. (2011b), this temporal propagation was done simply by copying the MS data along the optical flow trajectories from a fixed number of previous frames. Since optical flow becomes less reliable over a larger number of frames, the number of previous frames considered in Cao et al. (2011b) was limited to only two. However, we note that temporal information can be more effectively incorporated by using a larger number of previous frames and downweighting more temporally distant frames to reflect the lower correspondence accuracy from optical flow. We formulate this as a TF algorithm that extends the BF of Cao et al. (2011b) to include a temporal dimension, as illustrated in Fig. 4b.

With the TF, the MS data $\boldsymbol{ms}_{ijt}$ of each pixel $(i,\ j)$ in the current frame t is calculated by extending Eq. 1 to the following:

$$\boldsymbol{ms}_{ij} = \sum_{c \in R,\,G,\,B} \frac{\sum_{k \in \Omega'} \mathcal{G}_{\sigma_r}(d_k^{RGB})\mathcal{G}_{\sigma_s}(d_k^{xy})\mathcal{G}_{\sigma_t}(d_k^t) \cdot \rho_k \cdot \boldsymbol{ms}_k^c}{\sum_{k \in \Omega'} \mathcal{G}_{\sigma_r}(d_k^{RGB})\mathcal{G}_{\sigma_s}(d_k^{xy})\mathcal{G}_{\sigma_t}(d_k^t)}, \tag{4}$$

where Ω' includes pixels within a neighborhood centered on $(i,\ j)$ in the current frame as well as corresponding pixels from prior frames as computed from optical flow. d_k^t denotes the distance between $(i,\ j)$ and k in the time domain. ρ_k and per-channel propagation are computed in the same manner as done for the BF.

To evaluate the effectiveness of this TF, we first generate ground truth datasets by taking all of the *.mat* and *.aix* images in the Joensuu MS dataset (2005) and adding artificial linear motion to create MS video sequences. We simulate our system by computing the RGB values of each pixel using the RGB filter responses in Fig. 3b and by sampling the MS data at pixels using a pattern and density (about 0.3 % of the total image pixels) similar to our physical system. Accuracy is computed from differences of spectral intensity from the ground truth over all the wavelengths. The parameters are fixed to $\sigma_s = 16$, $\sigma_r = 16$ and $\sigma_t = 8$, and improvements in accuracy over our previous bilateral filtering method are shown for each of the video sequences in Table 1.

3.3 Acceleration and GPU Implementation

With trilateral filtering for temporal enhancement, a larger number of video frames can be effectively utilized in spectral propagation. While this leads to an improvement in accuracy, it also increases the computational burden. A major reason for the substantial computation is the large number of spectral

Table 1 Accuracy evaluation on ground-truth datasets with different propagation algorithms, where Sampling Res denotes the spatial resolution of sampled multispectral data, *Acc* means accuracy, BF and TF stand for bilateral filter and trilateral filter, PBF and PPTF stand for parallelized BF and parallelized PCA-accelerated TF

Index	Names	Image Res	Sampling Res	BF *Acc*	TF *Acc*	PPTF *Acc*	BF time (s)	PBF time (s)	TF time (s)	PPTF time (s)
1	Winding card	905 × 1, 087	45 × 45	0.8963	0.9106	0.9082	47.41	0.2478	485.1	0.7273
2	Yarn Palette	905 × 1, 087	45 × 45	0.8137	0.8387	0.8355	49.35	0.3021	467.9	0.8259
3	Shire	1,029 × 1,292	50 × 50	0.8923	0.9097	0.9068	53.92	0.4083	457.3	1.0239
4	House	1,029 × 1,292	50 × 50	0.8398	0.8607	0.8561	55.83	0.3891	465.8	1.1307
5	Colors	1,027 × 1,054	50 × 50	0.9521	0.9714	0.9693	58.13	0.4106	476.3	0.9865
6	Hsdbimga	520 × 696	40 × 40	0.9214	0.9388	0.9369	27.45	0.2054	283.2	0.4472
7	Chestnuts	242 × 242	30 × 30	0.9376	0.9512	0.9507	13.89	0.1517	151.5	0.3253
8	Face	152 × 91	10 × 10	0.9692	0.9870	0.9838	1.656	0.031	16.36	0.029
9	Girl	147 × 87	10 × 10	0.9668	0.9854	0.9806	1.651	0.033	13.62	0.023
10	Hand	160 × 148	10 × 10	0.9536	0.9670	0.9621	2.283	0.045	24.93	0.039
11	Reading	147 × 87	10 × 10	0.9820	0.9929	0.9913	1.509	0.035	16.11	0.029

The median accuracy rates of BF, TF and PPTF are 93.7, 95.1 and 95.0 %, respectively

channels that are processed. As shown in Table 1, over 20 s is required just for a single 160 × 148 frame with 150 spectral channels, which is much too long for real time applications. To accelerate the algorithm as well as reduce storage costs, we simplify the spectral representation with the use of PCA. Through PCA, we reduce the dimensionality of the spectral domain from 150 to 15, which significantly decreases the number of spectral channels to be propagated and brings a ×9.5 speedup. The cost of this dimensionality reduction is a small loss in accuracy of about 0.3 % on average over the ground truth datasets.

Besides PCA dimensionality reduction, we also take advantage of parallel computation to accelerate the propagation algorithm. We developed implementations of both the bilateral filtering and trilateral filtering algorithms for spectral propagation on the GPU using the CUDA programming model from Nvidia (2007). Global memory and shared memory on the GPU are used for efficient computation. Furthermore, as the TF is defined in a discrete domain, some of the time-consuming numerical computation is pre-calculated and stored in look-up tables to expedite processing. By trading space for time, the parallel algorithm can remarkably increase speed. On an Nvidia GTX 550, our TF implementation can generate nearly one frame per second with 97.0 % accuracy on 1, 027 × 1, 054 images, as detailed in Table 1. The implementation for bilateral filtering is about twice as fast, with a lower accuracy of 95.2 %. Though below full video frame rates, MS video generated at these rates can nevertheless be used in real-time vision applications. In such applications, our system only handles captured video frames in accordance to its processing rate and discards the rest.

3.4 Image Distortions and Design Tradeoffs

The settings of the hybrid camera system can be adjusted for different spectral resolutions. The spectral resolution depends on several system parameters, namely the focal length *f* of the camera lens, the refraction index *n* of the prism, the prism angle ω, and the incident angle α of light on the prism, as shown in Fig. 2a. The most effective way to increase spectral resolution is to use a larger focal length lens with a suitable occlusion mask that prevents overlaps of spectra on the sensor. Unlike the device in Du et al. (2009) and Cao et al. (2011a), higher spectral resolution does not require a corresponding loss of spatial resolution, which always remains at the resolution of the RGB camera in our system. However, when spectral resolution is increased using a larger focal length, our system samples fewer MS scene points per frame, which can reduce propagation accuracy. Another tradeoff when using a large focal length to gain spectral resolution is that less light energy is collected on the grayscale sensor. If an increase in exposure time is needed for sufficient brightness, the frame rate of the camera would need to be reduced.

Different camera settings also affect the image distortions introduced by the system optics. For example, a larger aperture increases light throughput but decreases the depth of field (DOF). A smaller DOF leads to greater keystone distortion (see Sect. 4) as well as spectral samples that are less in-focus. Greater distortion is also introduced by increasing the incident angle α. Though higher spectral resolution can be gained with a larger α, in practice a smaller aperture is needed to reduce the distortions to a more manageable level. Various factors need to be considered when setting the camera parameters, but we note that calibration for the resulting distortions need only to be performed once after fixing the configuration.

4 System Implementation

The hybrid camera system, exhibited in Fig. 5a, was implemented using a PointGrey® GRAS-50S5M for the greyscale camera, which can capture 15 fps video at a maximum resolution of $2,448 \times 2,048$. The focal length of the greyscale camera lens is set according to the spectral resolution required in a given application. The RGB camera is a PointGrey® Flea2-08S2, which has a maximum resolution of $1,024 \times 768$ at 25 fps. The two cameras are synchronized using the PointGrey® MultiSync program. A half-reflect, half-pass beam splitter provides the two cameras with equal light energy. The occlusion mask is configured as shown in Fig. 5b, with rectangular holes through which the prism disperses light in the horizontal direction. Each hole represents one MS sample, and the projected spectrum from the hole is averaged vertically to reduce noise. The prism is made of *BK7* glass with known refraction indices for wavelengths 400–1,000 nm. An optional filter may be placed in the optical path to isolate a certain band of the spectrum. The spectrum, geometry and radiance distortions in the MS imager are calibrated as in Du et al. (2009) and Cao et al. (2011a). Examples of captured spectra are shown for an example scene in Fig. 5c. The scene consists of a stuffed animal against a white backdrop with two candles placed on the right and fluorescent lighting coming from the left side. The captured spectra exhibit the gradual change in spectral distribution in transitioning between the two forms of lighting. The shading from the two types of illumination can be roughly approximated by constructing image frames at wavelengths of 450 and 710 nm.

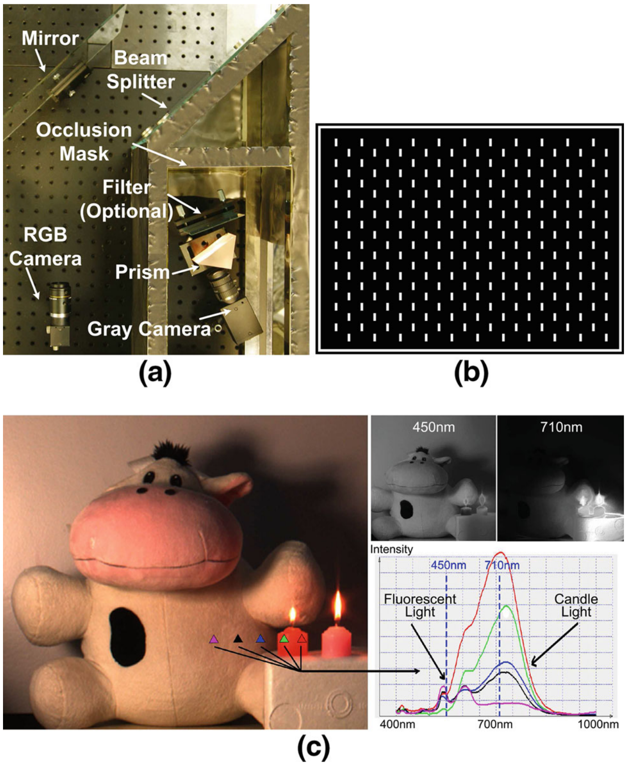


Fig. 5 (**a**) System prototype. (**b**) Occlusion mask configuration. (**c**) Real scene capture with the proposed system

As previously mentioned, there are distortions among the spectral samples (e.g., keystone and smile distortions shown in Fig. 6a) and within each spectral sample (shown in Fig. 6b). These distortions must be calibrated and undone to maximize the quality of the spectral measurements and to improve the correspondence of MS samples to pixels in the RGB camera. For spectral calibration, we take advantage of the characteristics of fluorescent light, whose spectrum has two peaks at 546.5 and 611.6 nm (Fig. 6c), by linearly unwarping and aligning each spectral sample so that the two peaks correspond to those in the standard fluorescent spectrum. This is done with an image captured of a white surface under fluorescent light. For geometric distortions among the spectral samples, Delaunay triangulation is used to warp the image so that the sample locations match those of the known mask holes.

To register the two cameras, we place an LCD screen in front of the capture system and display a calibration video of horizontal and vertical line sweeps. The lines are displayed on the LCD in white, which has a spectrum with peaks similar to those of fluorescent light, as shown in Fig. 7a. When a horizontal line sweeps past a row sampled by the MS sensor, it produces a sharp response as in Fig. 7b and its corresponding row position on the RGB sensor is identified. Correspondences between columns are similarly obtained using a vertical line sweep. The information from the two sweeps jointly determines the correspondences of individual MS samples to pixels on the RGB sensor. This registration and alignment is a one-time operation, and does not need to be repeated when capturing another scene.

Once the spectrum and geometry distortion are calibrated, we then recover the true spectral radiance at each wavelength up to a constant scale factor as described in Cao et al. (2011a), using camera response curves and sensor spectral sensitivities generally obtainable in sensor documentation.

5 Experimental Results

In this section, we use the proposed system to capture several videos and demonstrate the effectiveness and utility of high spectral and spatial resolution measurement on various applications. The high-spectral, high-spatial resolution

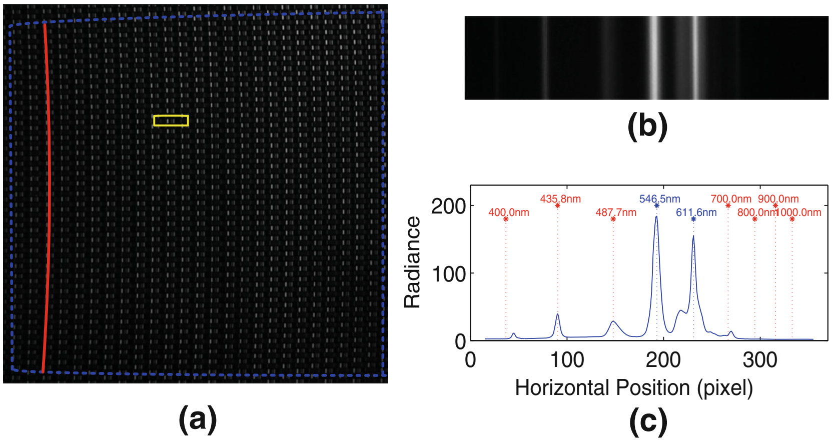


Fig. 6 (**a**) Geometric distortions: smile distortion (*red curve*) and keystone distortion (*blue trapezoid*). (**b**) A spectral sample from the *yellow box* in (**a**). (**c**) Corrected spectral curve of (**b**), in which the two peaks are aligned with their actual wavelengths at 546 and 611 nm

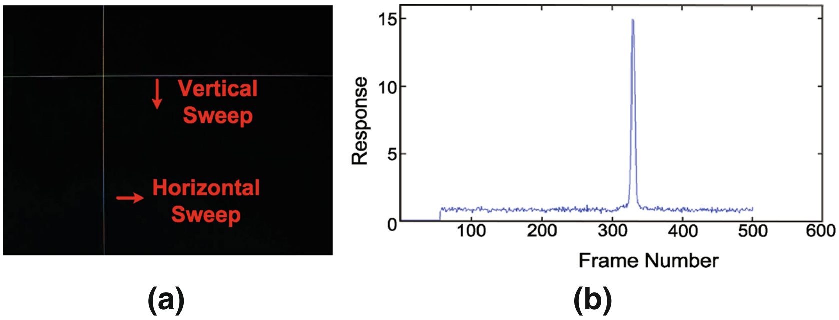


Fig. 7 Alignment of hybrid cameras. (**a**) LCD line sweeps. (**b**) Multispectral sample response in a line sweep

measurements provide considerably more data than traditional RGB or monochrome cameras for identifying and distinguishing different materials and objects. The rapid generation of output video from the captured streams makes our system practical for real-time video applications.

5.1 Resolution Comparison

We demonstrate the increase in spatial resolution that can be gained with our approach over using just a prism-mask MS video system as presented in Du et al. (2009) and Cao et al. (2011a). In this example, the MS imager is configured so that each spectral sample spans about 60 pixels on the grayscale sensor over the wavelength range of 400–700 nm, giving a spectral resolution of about 5 nm (61 channels) and a spatial resolution of 116 × 24. Video frames captured by the MS sensor are displayed in Fig. 8c, h using RGB values computed from the color filter response curves of the RGB camera. At this resolution, details on the objects are unclear. The characters on the monkey and book become visible at the $1,024 \times 768$ resolution produced with our hybrid camera system as exhibited in Fig. 8d, i. These images also show that the RGB reproduction from propagated MS data is consistent with the captured RGB frames. Figure 8e, j show examples of the additional spectral detail that can be acquired with our device in comparison to the coarse spectra reconstructed from RGB values using the method of Smits (1999).

Besides an increase in spatial resolution, this example illustrates the ability to distinguish illumination types based on MS measurements. As shown in Fig. 8e, the characteristic spectral peaks of fluorescent light are present in the reflected radiance, which is in contrast to the spectra in Fig. 8j captured of a scene under tungsten illumination. This distinction of scene lighting can facilitate automatic white balance in a manner not possible with RGB cameras.

5.2 Ground-Truth Verification

To test the accuracy of our proposed MS capture system, we compared our measurements to ground-truth values of a color calibration target with known illumination. The color chart, shown in Fig. 9a, contains 24 painted chips with spectral reflectances that are representative of natural objects. Measurements from a scanning spectrometer are also presented for comparison. With the spectrometer, we capture a reference white object and a reference black object with known spectral reflectances for calibration of the signal intensity. Typical comparison results are shown in Fig. 9. It can be seen that the spectral measurements of our system

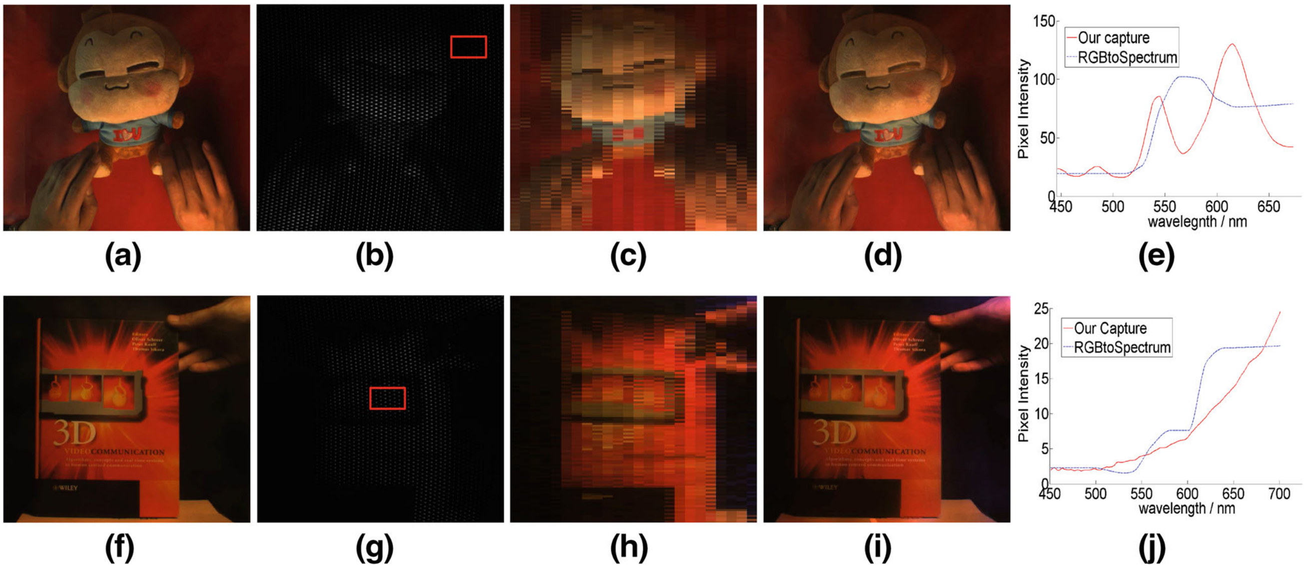


Fig. 8 Resolution comparison. (**a**), (**f**) Video frames captured by the RGB sensor. (**b**), (**g**) Video frames captured by the *grayscale* sensor. (**c**), (**h**) Reconstructed RGB in the *grayscale* frames. (**d**), (**i**) High resolution RGB video using the propagation algorithm. (**e**), (**j**) Comparison of spectral detail from an RGB camera and our multispectral system

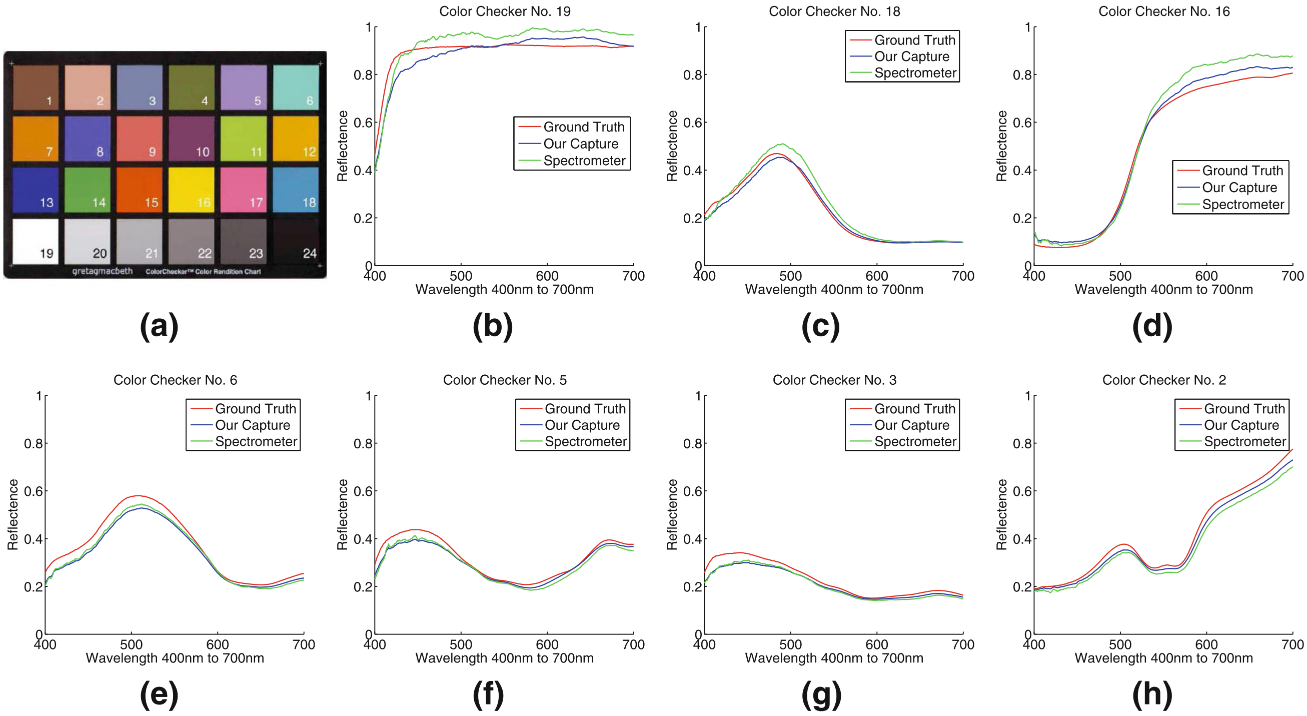


Fig. 9 Ground-truth verification. (**a**) Standard *color chart* consisting of an arrangement of 24 painted squares. (**b**)–(**h**) Spectral curves from our system, a scanning spectrometer, and ground truth

closely approximate both the ground truth and the measurements of the spectrometer. The spectrometer used in this experiment is the Isuzu Optics Spectral Camera HS with a 2.73 nm spectral resolution and a 400–1,000 nm spectral range.

5.3 System Evaluation

The overall error of the proposed system is the product of errors in the spectral measurement and the propagation algorithm. The former is due to calibration errors for the spectrum,

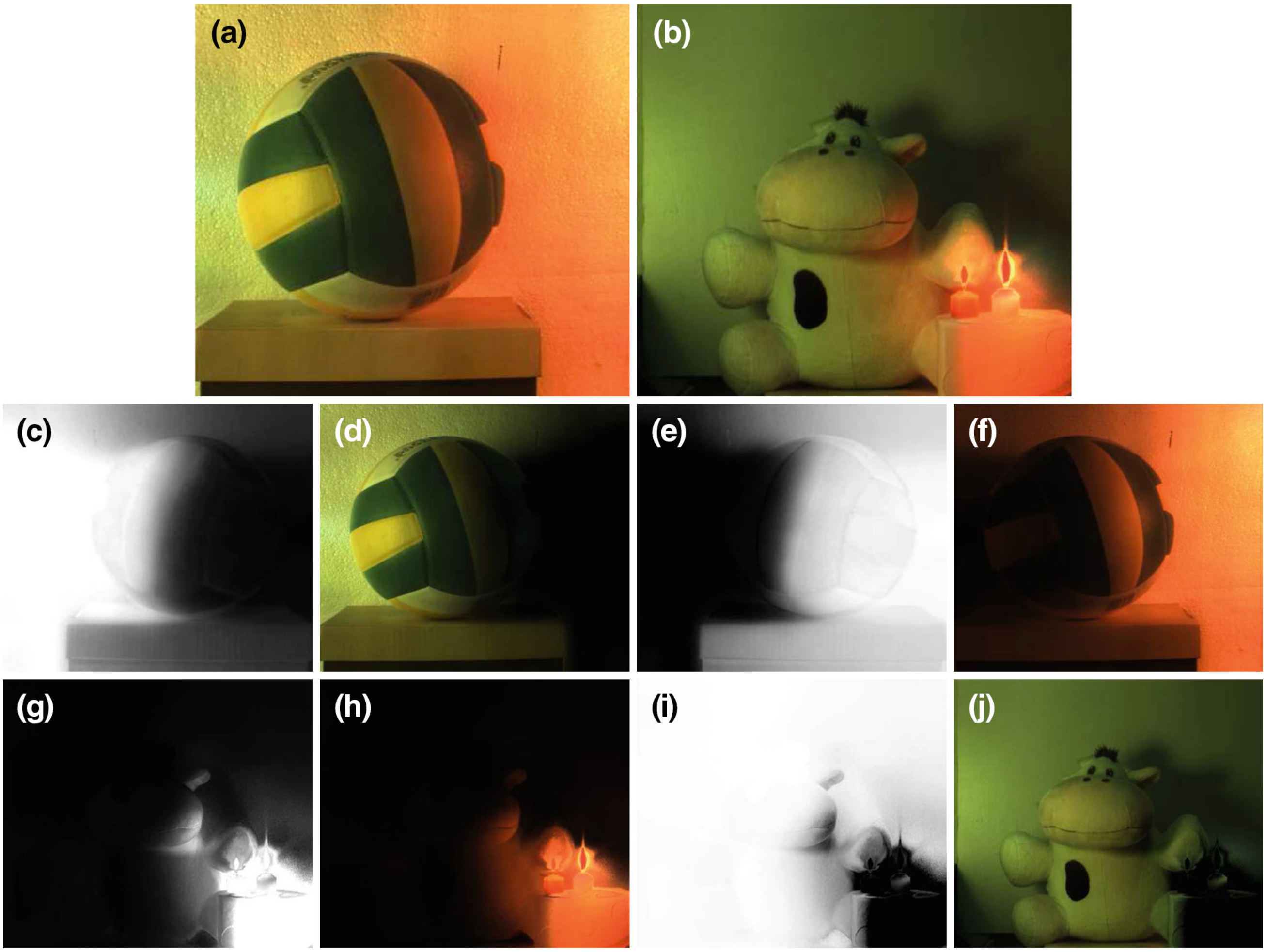


Fig. 10 Separation of mixed illumination. (**a**), (**b**) Scenes with mixed illumination. (**c**), (**e**), (**g**), (**i**) Alpha maps for respective illumination. (**d**), (**f**), (**h**), (**j**) Results of illumination separation

geometry and spectral radiance, which are discussed in detail in Cao et al. (2011a). The reported root-mean-squared (RMS) errors of the spectrum calibration and spectral radiance calibration are 0.62 and 2.28 %, respectively. The median accuracy of our propagation algorithm as examined in Sect. 3 is 95.0 %. The overall error of our proposed system was evaluated through the aforementioned ground-truth verification experiment. The RMS error of our system with respect to ground truth and the spectrometer are 6.86 and 8.30 %, respectively.

5.4 User-Assisted Separation of Mixed Illumination

Scene illumination is a fundamental element in problems such as image relighting (Wenger et al. 2005) and intrinsic image decomposition (Shen et al. 2008). As a highly underdetermined problem, separation of mixed illumination cannot be solved with only RGB information in general. Illumination often has a strong effect on measured spectra, and by using spectral information, different illumination are more distinguishable. Here we present an application of our camera system for user-assisted separation of appearance contributions from different light sources in a scene. Figure 10 displays a volleyball in front of a white backdrop, with fluorescent lighting from the left side and tungsten lighting from the right. Gradual transitions between the two types of lighting are present in the scene, but shading from the two illuminations is difficult to separate from each other based on RGB data. From MS data captured by the hybrid camera system, the contributions of the two light sources become easier to distinguish. Their alpha maps can be recovered as shown in Fig. 10c, e, and from these alpha maps we can infer what the scene would look like with one or the other light source removed, as displayed in Fig. 10d, f. Illumination separation results are also shown for the scene in Fig. 10b lit by candles and fluorescent lighting. The sharp border artifacts in Fig. 10h are caused by the over-exposure of regions surrounding the candle light. The candle light dominates the reflected

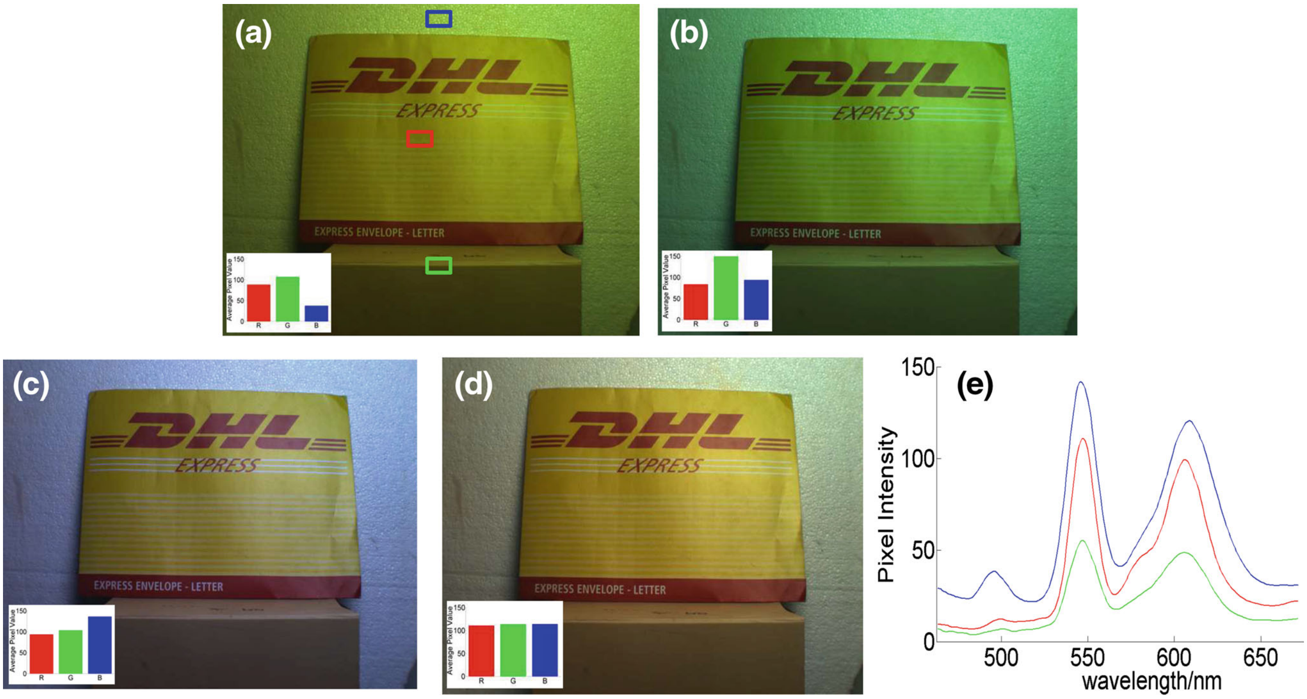


Fig. 11 Automatic *white* balance. (**a**) Raw frame captured by the RGB camera. (**b**) *White* balance result based on tungsten light. (**c**) Result of *grey* world *white* balance algorithm. (**d**) Result using correct illumination information (fluorescent light). (**e**) Multispectral curves of the *marked* areas in (**a**), which were captured by the proposed system. The peaks of these curves reveal the actual illumination to be fluorescent light

spectrum in these regions, which results in deep contrast in the alpha map.

A brief description of the illumination separation procedure is presented in the following. The spectrum $\mathbf{I}_{ij}$ of an image pixel $(i,\ j)$ can be expressed as

$$\mathbf{I}_{ij} = \mathbf{R}_{ij}\left[\mathbf{L}_1\ \mathbf{L}_2\right]\begin{bmatrix}\alpha_{ij}\\ 1-\alpha_{ij}\end{bmatrix}, \tag{5}$$

where $\mathbf{R}_{ij}$ denotes the reflectance at $(i,\ j)$, $\mathbf{L}_1$ and $\mathbf{L}_2$ are the spectra of the two illumination sources, and α_{ij} denotes alpha map values for the first illumination. With help from the user to identify the types of illumination sources, we can determine $\mathbf{L}_1$ and $\mathbf{L}_2$ from a list of common illumination spectra. We then calculate intermediate quantities $x_1,\ x_2$ as the least squares solution of

$$\mathbf{I}_{ij} = \left[\mathbf{L}_1\ \mathbf{L}_2\right]\begin{bmatrix}x_1\\ x_2\end{bmatrix}. \tag{6}$$

From $x_1,\ x_2$, the alpha map can be derived as

$$\alpha_{ij} = \frac{x_1}{x_1 + x_2}. \tag{7}$$

We note that at wavelengths where $\mathbf{I}_{ij}$ is low, the signal-to-noise ratio also becomes low and $x_1,\ x_2$ cannot be estimated reliably. We thus discard wavelengths with low intensity when solving Eq. 6. From the calculated alpha map, reflectance $\mathbf{R}$ can be solved from Eq. 5, and the illumination separation results are then computed as $\alpha\mathbf{R}\mathbf{L}_1$ and $(1-\alpha)\mathbf{R}\mathbf{L}_2$.

5.5 Automatic White Balance

High spectral resolution can benefit techniques for automatic white balance. In Fig. 11 we show an example of a warm colored object under fluorescent lighting, captured by a regular RGB camera and by our system. The warm colors of the object can mislead a white balance algorithm to infer a warm illumination color like tungsten light, giving the incorrect result in Fig. 11b. Figure 11c is the white balance result using the grey world hypothesis (Buchsbaum 1980), a standard approach that assumes the average scene radiance to be achromatic. In reality, the scene is illuminated by a fluorescent light, obvious from the spectra in Fig. 11e captured by our system. Fluorescent lighting is detected by the two characteristic peaks at wavelengths 546 and 611 nm. Figure 11d shows the white balance result with the correct illumination. Since the background color is white in this scene, we can

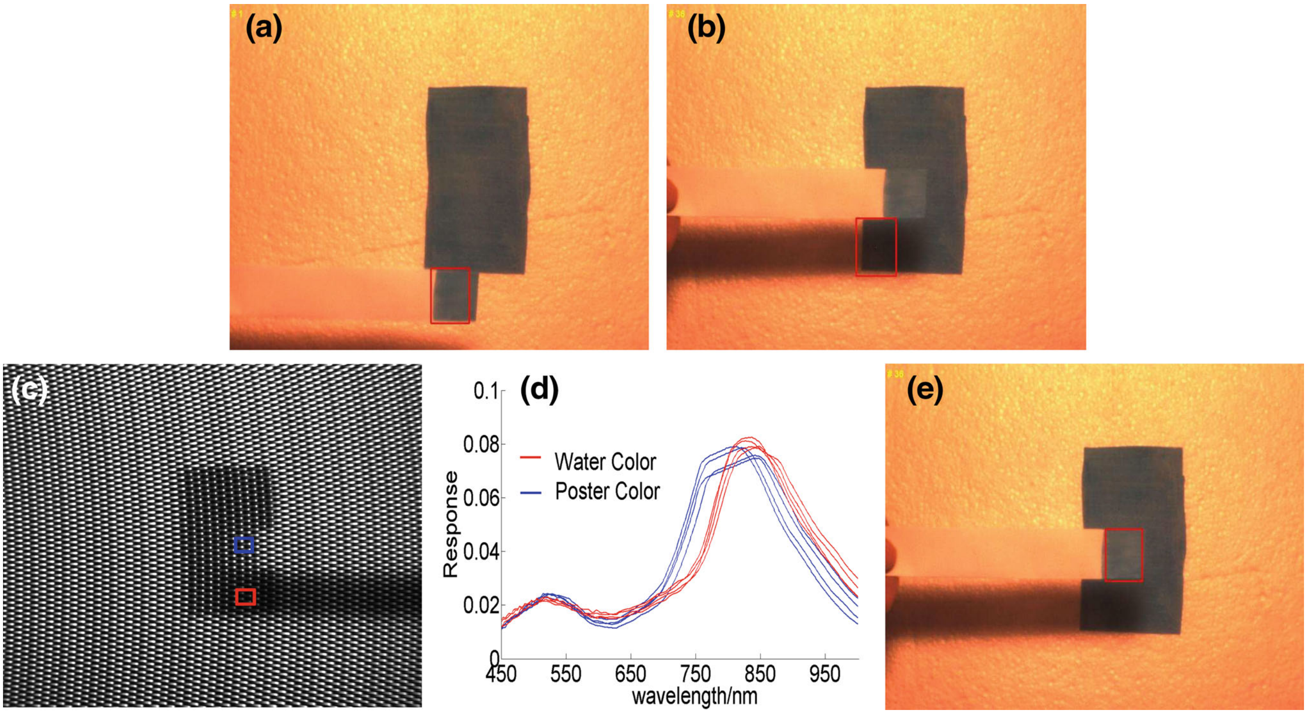


Fig. 12 Object tracking. (**a**) Tracking target, highlighted by a *red rectangle*. (**b**) Unsuccessful tracking using an RGB camera. (**c**) Multispectral frame from the *grayscale* sensor. (**d**) Spectra of poster *color* and water *color* in areas *marked* in (**c**). (**e**) Successful tracking using the hybrid camera system

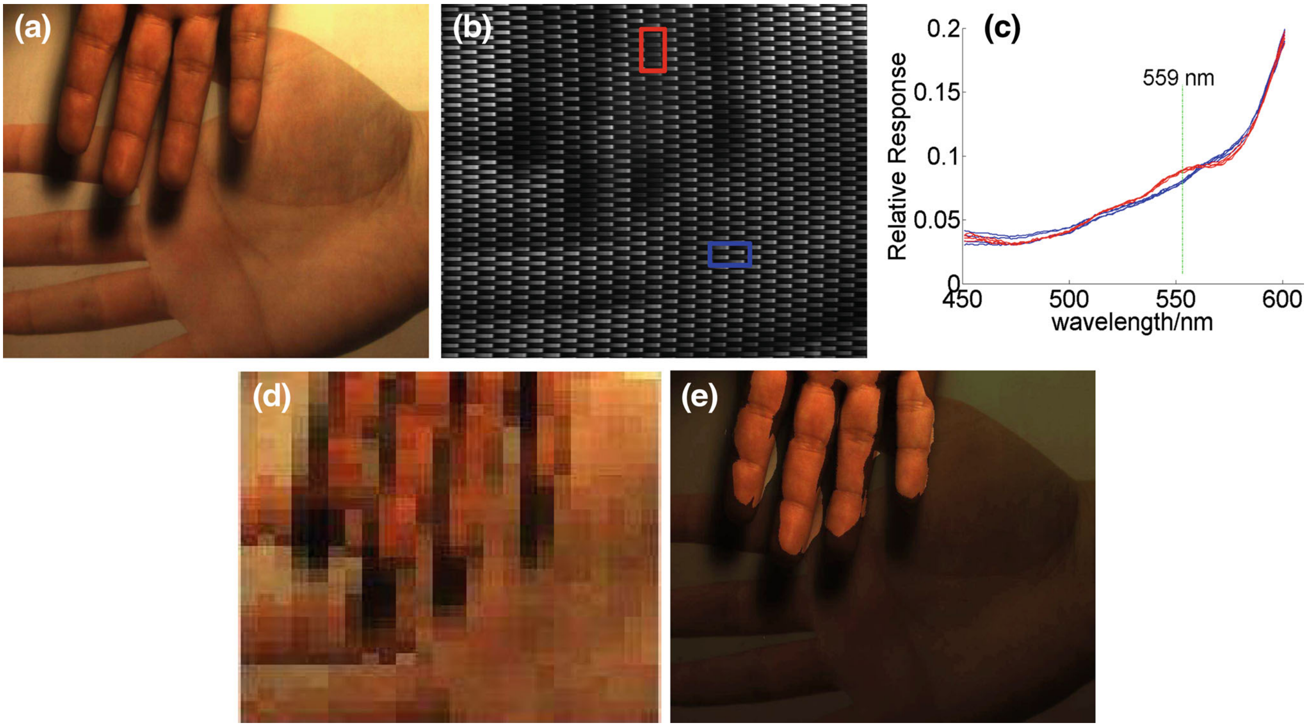
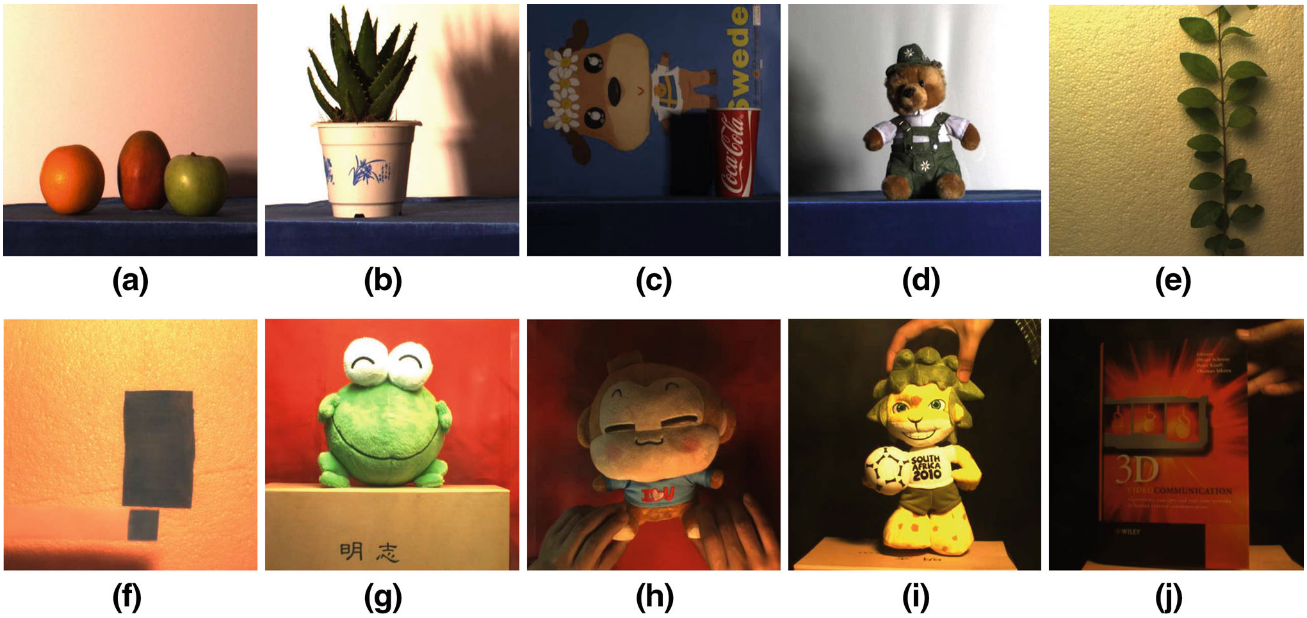


Fig. 13 Human skin tracking. (**a**) An RGB frame containing an actual hand and a fake hand. (**b**) A multispectral frame captured by the *grayscale* sensor. (**c**) Spectral curves from real human skin (*red*) and printed skin (*blue*). (**d**) A low resolution frame without multispectral propagation, where image details are unclear. (**e**) A high resolution frame with multispectral propagation, where the detected human skin pixels are *highlighted*



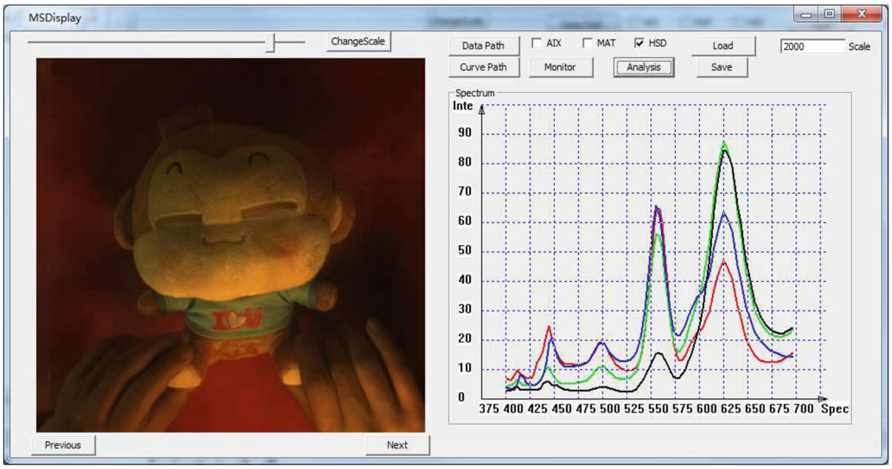


Fig. 14 (**a**)–(**j**) Images from our high spatial and spectral resolution video dataset. (**k**) User interface for our multispectral viewing software

evaluate the white balance result based on the average color of the background, shown at the bottom-left of each figure.

5.6 Tracking

In tracking, features of the target object need to be distinguished from the background. This problem is particularly challenging when the foreground object and background have a similar color. Figure 12 illustrates such an example, with a moving foreground object of blue poster color and part of the background in blue water color. With a traditional RGB camera, the two colors are difficult to distinguish. Tracking with a state-of-the-art algorithm (Ross et al. 2008) fails when the object moves over the blue background. With the hybrid camera system, the spectral differences between poster color and water color become evident, allowing for accurate tracking. Here, the tracking algorithm uses responses at wavelengths 550, 625 and 700 nm as color features for the MS input video. The video was captured with a spatial resolution of 800 × 600 under tungsten illumination. We used a 25 mm lens in this example, which gives a spectral resolution of about 6 nm.

We also demonstrate tracking of human skin, a task important in many surveillance applications. A special characteristic of human skin spectra is a small hump centered on wavelength 559 nm, a physical feature caused by melanin and hemoglobin in skin (Angelopoulou 2001). Detection of this hump using a MS camera can distinguish human skin from other materials with similar colors in RGB space. Figure 13 shows an example in which we capture a real human hand in front of a fake hand printed on paper. In this experiment, we identify true skin pixels by thresholding the quantity $(r(559) - 0.5(r(540) + r(580)))$ to detect the 559 nm peak, where $r(\lambda)$ denotes the spectral radiance at wavelength λ. We use a 50 mm lens and a 400–620 nm bandpass filter in this example, which gives a spectral resolution of about 1 nm. We capture a video clip of the hand performing different gestures. One frame is shown in Fig. 13a. Figure 13b shows the corresponding MS frame captured by the grayscale camera. Figure 13c plots the spectral curves of samples from the actual human hand and fake hand. The spectral hump around 559 nm is evident in measurements of the real hand. Figure 13e shows the detection result in a video frame, with the detected skin pixels highlighted. In comparison to the low resolution frame in Fig. 13d, the high spatial resolution of our system in Fig. 13e allows for a clearer analysis of hand gestures.

5.7 MS Video Dataset

Existing MS datasets (e.g., the Joensuu MS dataset 2005) contain only static images. To facilitate spectral research on dynamic scenes, we have captured several high resolution MS videos of scenes shown in Fig. 14. This dataset will be made publicly available at http://media.au.tsinghua.edu.cn/msdataset.html.

We additionally developed a novel MS image and video representation based on our work. This representation, called Hybrid Spectral Data format (*.hsd*), follows the basic structure of our captured data as shown in Fig. 1, by consisting of sparse MS samples together with full-resolution RGB data. As this representation does not explicitly include hundreds of channels of spectral data for each pixel, it has the advantage of reduced transmission bandwidth and storage space. With our propagation algorithm, the MS data of any pixel can be obtained from an *.hsd* file in real time.

To view *.hsd* files, we have built software shown in Fig. 14k. Users can click on a pixel in the image viewing window (rendered in RGB) and its corresponding spectrum will be shown in a plot. The spectra of multiple image points can be displayed simultaneously for comparison, and other MS image formats (namely *.aix* and *.mat*) are supported. Spectrum analysis tools can be incorporated into this software, which will also be made publicly available.

6 Conclusions and Future Work

MS imaging at video rates has long been a challenging problem. In this paper, we addressed this problem with a hybrid camera system that generates video with both high spectral and high spatial resolution. Our experiments demonstrate the utility of such video in different computer vision applications. System implementation details are included in this paper together with a discussion of design tradeoffs. Furthermore, a dataset of MS videos has been collected, and software has been developed for viewing MS video files in a new compact format based on our work.

A limitation of this system is the relatively large DOF needed to reduce optical distortions from the prism and to avoid blurring of the occlusion mask. While this can be accomplished with a small aperture size, the light throughput of our system is limited as a result. Furthermore, the ability of our hybrid camera system to handle high-frequency color variations is limited by its relatively low MS sampling rate and its reliance on similar pixels for MS data propagation. This problem may potentially be addressed by using dynamic masks with temporal multiplexing.

Another issue of the presented system is its large physical size. The size could be reduced by reconfiguring the optics in a manner with a small occlusion mask placed at the aperture. For future work, we are interested in incorporating a temporally-varying mask that can increase MS sampling and thus further improve spectral propagation accuracy.

Acknowledgments The authors thank Yanxiang Lan, Hao Du and Moshe Ben-Ezra for helpful discussions on implementation issues, Tao Yue for assistance in experimentation, the National Basic Research Project (No. 2010CB731800) and the Key Project of NSFC (No. 61035002), The NSFC Grant No. 61371166, and the reviewers for their constructive comments.

References

Angelopoulou, E. (2001). Understanding the color of human skin. *SPIE Conference on Human Vision and Electronic Imaging*, *4299*, 243–251.

Brady, D. J., & Gehm, M. E. (2006). Compressive imaging spectrometers using coded apertures. In *SPIE* (Vol. 6246).

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *ECCV*.

Buchsbaum, G. (1980). A spatial processor model for object colour perception. *Journal of the Franklin Institute*, *310*, 1–26.

Cao, X., Du, H., Tong, X., Dai, Q., & Lin, S. (2011a). A prism-mask system for multispectral video acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(12), 2423–2435.

Cao, X., Tong, X., Dai, Q., & Lin, S. (2011b). High resolution multispectral video capture with a hybrid camera system. In *2011 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Colorado Springs, CO, pp. 297–304.

Chi, C., Yoo, H., & Ben-Ezra, M. (2010). Multi-spectral imaging by optimized wide band illumination. *International Journal of Computer Vision*, *86*(2), 140–151.

Darling, B. A., Ferwerda, J. A., Berns, R. S., & Chen, T. (2011). Realtime multispectral rendering with complex illumination. In *Color and Imaging Conference* (Vol. 2011, pp. 345–351). Springfield, VA: Society for Imaging Science and Technology.

Descour, M., & Dereniak, E. (1995). Computed-tomography imaging spectrometer: Experimental calibration and reconstruction results. *Applied Optics*, *34*, 4817–4826.

Du, H., Tong, X., Cao, X., & Lin, S. (2009). A prism-based system for multispectral video acquisition. In *Proceedings of the ICCV*.

Fletcher-Holmes, D. W., & Harvey, A. R. (2005). Real-time imaging with a hyperspectral fovea. *Journal of Optics A: Pure and Applied Optics*, *7*, S298–S302.

Habel, R., Kudenov, M., & Wimmer, M. (2012). Practical spectral photography. *Computer Graphics Forum*, *31*, 449–458. Wiley Online Library.

Hagen, N., & Dereniak, E. L. (2008). Analysis of computed tomographic imaging spectrometers. I. Spatial and spectral resolution. *Applied Optics*, *47*, F85–F95.

Han, S., Sato, I., Okabe, T., & Sato, Y. (2010). Fast spectral reflectance recovery using DLP projector. In *Computer Vision: ACCV*, *2010* (pp. 323–335).

James, J. (2007). *Spectrograph design fundamentals*. Cambridge, MA: Cambridge University Press.

Johnson, W. R., Wilson, D. W., & Bearman, G. (2006). Spatial–spectral modulating snapshot hyperspectral imager. *Applied Optics*, *45*, 1898–1908.

Kim, M. H., Rushmeier, H., Dorsey, J., Harvey, T. A., Prum, R. O., Kittle, D. S., et al. (2012). 3D imaging spectroscopy for measuring hyperspectral patterns on solid objects. *ACM Transactions on Graphics (TOG)*, *31*(4), 38.

Kittle, D. S., Marks, D. L., & Brady, D. J. (2012). Design and fabrication of an ultraviolet-visible coded aperture snapshot spectral imager. *Optical Engineering*, *51*(7), 071403-1.

Mooney, J. M., Vickers, V. E., An, M., & Brodzik, A. K. (1997). High throughput hyperspectral infrared camera. *Journal of Optical Society of America A*, *14*, 2951–2961.

Mrozack, A., Marks, D. L., & Brady, D. J. (2012). Coded aperture spectroscopy with denoising through sparsity. *Optics Express*, *20*(3), 2297–2309.

Nvidia, C. (2007). *Compute unified device architecture programming guide*. Santa Clara, CA: NVIDIA Corp.

Park, J. I., Lee, M. H., Grossberg, M. D., & Nayar, S. K. (2007). Multispectral imaging using multiplexed illumination. In *ICCV*.

Ross, D., Lim, J., Lin, R. S., & Yang, M. H. (2008). Incremental learning for robust visual tracking. *IJCV*, *77*, 125–141.

Schechner, Y. Y., & Nayar, S. K. (2002). Generalized mosaicing: Wide field of view multispectral imaging. *IEEE PAMI*, *24*(10), 1334–1348.

Shen, L., Tan, P., & Lin, S. (2008). Intrinsic image decomposition with non-local texture cues. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR 2008)*, Anchorage, AK, pp. 1–7.

Smits, B. (1999). An RGB to spectrum conversion for reflectances. *Journal of Graphics Tools*, *4*(4), 11–22.

University of Joensuu Color Group. (2005). *Spectral database*. http://spectral.joensuu.fi/.

Vandervlugt, C., Masterson, H., Hagen, N., & Dereniak, E. L. (2007). Reconfigurable liquid crystal dispersing element for a computed tomography imaging spectrometer. In *SPIE* (Vol. 6565).

Volin, C. (2000). *MWIR spectrometer operating theory*. Tucson, AZ: University of Arizona Press.

Wagadarikar, A., John, R., Willett, R., & Brady, D. (2008). Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics*, *47*, B44–B51.

Wagadarikar, A., Pitsianis, N., Sun, X., & Brady, D. (2009). Video rate spectral imaging using a coded aperture snapshot spectral imager. *Optics Express*, *17*(8), 6368–6388.

Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T., & Debevec, P. (2005). Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)*, *24*(3), 756–764.

Yamaguchi, M., Haneishi, H., Fukuda, H., Kishimoto, J., Kanazawa, H., Tsuchida, M., et al. (2006). High-fidelity video and still-image communication based on spectral information: Natural vision system and its applications. In *SPIE/IS&T Electronic Imaging* (Vol. 6062).

Yang, Q., Tan, K. H., & Ahuja, N. (2009). Real-time o(1) bilateral filtering. In *CVPR*.