# High-speed Hyperspectral Video Acquisition with a Dual-camera Architecture

Lizhi Wang[1*] Zhiwei Xiong[2] Dahua Gao[3] Guangming Shi[1] Wenjun Zeng[2] Feng Wu[4]
[1]Xidian Univ. [2]Microsoft Research [3]Air Force Eng. Univ. [4]Univ. of Sci. & Tech. of China

## Abstract

*We propose a novel dual-camera design to acquire 4D high-speed hyperspectral (HSHS) videos with high spatial and spectral resolution. Our work has two key technical contributions. First, we build a dual-camera system that simultaneously captures a panchromatic video at a high frame rate and a hyperspectral video at a low frame rate, which jointly provide reliable projections for the underlying HSHS video. Second, we exploit the panchromatic video to learn an over-complete 3D dictionary to represent each band-wise video sparsely, and a robust computational reconstruction is then employed to recover the HSHS video based on the joint videos and the self-learned dictionary. Experimental results demonstrate that, for the first time to our knowledge, the hyperspectral video frame rate reaches up to 100fps with decent quality, even when the incident light is not strong.*

## 1. Introduction

Hyperspectral imaging, which collects and processes scene information by dividing the whole spectrum into tens or hundreds of bands, has gained increasing attention from both academic and industrial communities. Thanks to its capability for detailed scene representation, this technique has been widely adopted in many fields, including medical diagnosis, health care, remote sensing, and military operations [17, 24]. Recently, it is found that various computer vision tasks, *e.g.*, recognition, classification, and tracking, can benefit from incorporating the spectral information in tens or hundreds of bands [11, 20, 28].

Unfortunately, conventional spectrometers have to confront a tradeoff between spatial/spectral and temporal resolution, as they need to scan the scene along either spatial or spectral dimension to capture a full hyperspectral image [19, 31, 7, 26]. Therefore, conventional spectrometers are not suitable for measuring dynamic scenes. To enable hyperspectral video acquisition, snapshot spectral imagers have been developed thanks to the flourish of computational

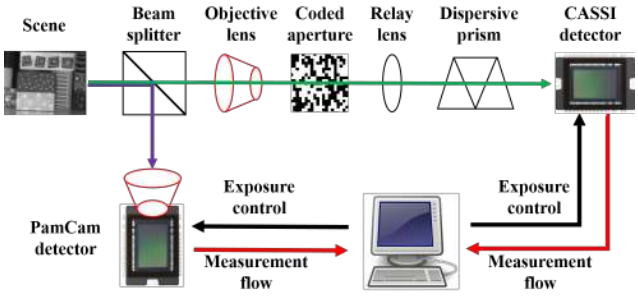


Figure 1. Dual-camera architecture for HSHS video acquisition.

reconstruction. Techniques in this category, including computed tomographic imaging spectrometry (CTIS) [29, 12], coded aperture snapshot spectral imager (CASSI) [3, 34], and hybrid spectral video imaging system (HVIS) [10, 25], have the ability to recover a full hyperspectral image with a single shot. Nevertheless, these systems need to employ additional optical elements to encode and/or disperse the scene information. Inevitably, the incident light will go through a long optical path and certain light-blocking elements, leading to considerable light intensity attenuation. As a result, the video frame rate that can be achieved with these snapshot spectral imagers is usually limited compared to the RGB/panchromatic cameras equipped with the same detector, especially when the incident light is not strong. So far, the highest frame rate reported in the literature is 30fps by CASSI, for a bright scene with burning candles [35].

In this paper, we propose a novel dual-camera design to acquire 4D high-speed hyperspectral (HSHS) videos, which leverages the high spatial and spectral resolution of the compressive spectral imaging and high light efficiency of the panchromatic camera (PanCam for short hereafter). This new design, as shown in Fig. 1, comprises a beam splitter, a high-speed PanCam, a suite of CASSI (inside the CASSI there is an objective lens, a coded aperture, a relay lens, a dispersive prism, and a detector). Specifically, the incident light from the scene is equally divided by the beam splitter into two parts, which are then captured by the PanCam and the CASSI, respectively. Light in the PanCam branch will go through a shorter path and less optical elements compared with the CASSI branch. Therefore, the

*This work was performed at Microsoft Research.

dual-camera system can simultaneously capture a panchromatic video at a high frame rate and a hyperspectral video at a low frame rate, which jointly provide reliable projections for the underlying HSHS video.

Meanwhile, the panchromatic video is further exploited to learn an over-complete 3D dictionary to represent each band-wise video sparsely. This is motivated by the observation that a 4D HSHS video can be treated as a concatenation of multiple band-wise videos which often have similar structural content (*e.g.*, edges) as the panchromatic video. Therefore, the dictionary learned from the panchromatic video yields high sparsity when representing the band-wise videos. A robust computational reconstruction is then employed to recover the HSHS video based on the joint videos and the self-learned dictionary.

With the enhanced overall light efficiency provided by the dual-camera design and the effective sparse representation provided by the self-learned dictionary, for the first time to our knowledge, it is possible to acquire 4D HSHS videos using a low-cost system as we developed in this paper. Experimental results demonstrate that the hyperspectral video frame rate reaches up to 100fps with decent quality, even when the incident light is not strong.

## 2. Related work

The fundamental problem for hyperspectral video acquisition is how to capture 4D data (2D spatial + 1D spectral + 1D temporal) in a 3D real world where the imaging sensor exists. Conventional spectrometers simply trade temporal resolution for spatial/spectral resolution, and thus lose the ability to record dynamic scenes [7, 19]. For example, pushbroom or whiskbroom based methods capture the spectral information of a slit or a single point of the scene, and spatially scan the whole scene to obtain a full hyperspectral image [4, 30]. Filter wheel or tunable filter based methods integrate multiple color bandpass filters to select one band for each exposure, and multiple exposures are required to capture different spectral information of the scene [16, 31]. All these systems actually cut off or block a large portion of light, and thus are inefficient in terms of light utilization.

To overcome the limitation of conventional spectrometers and make it possible to capture dynamic scenes, snapshot spectral imagers have been developed in the last decade. CTIS multiplexes the 3D spectral information onto a 2D detector with customized optical elements and reconstructs the underlying information by solving a linear problem [29, 12]. However, this method sacrifices the spatial resolution to achieve the snapshot property and also suffers from the missing cone problem. Another snapshot solution is the prism-mask spectral video imaging system (PMVIS) [15], which directly trades spatial resolution for spectral resolution using a customized occlusion mask. This solution is later upgraded to HVIS [10, 25], which uses an additional RGB camera to enhance the spatial resolution. A distinct advantage of PMVIS and HVIS is real-time reconstruction. However, the underlying problem is that the occlusion mask only allows a small portion of light to pass (the overall throughput is one out of the total band number), which limits both the spectral resolution and the video frame rate that can be achieved.

Relying on the compressive sensing theory, CASSI has made a significant breakthrough towards hyperspectral video acquisition. CASSI employs one or two dispersers and a coded aperture to optically encode the 3D spectral information onto a 2D detector, and a full hyperspectral image is then recovered through computational reconstruction [18, 34, 3]. CASSI has been demonstrated to capture dynamic scenes with high spatial and spectral resolution [35]. As a modification of CASSI, a dual-coded compressive spectral imager (DCSI) is proposed recently [22], which separately encodes spatial and spectral dimensions using a digital micromirror device (DMD) and a liquid crystal on silicon (LCOS). Still, the video frame rate of the above systems is limited due to light intensity attenuation caused by the extended light path and the light-blocking elements (*e.g.*, coded aperture, DMD, and LCOS), especially when the incident light is not strong.

Our work is built on CASSI and also motivated by the hybrid imaging systems that use a high-speed, low-resolution camera and a low-speed, high-resolution camera for motion deblurring [5, 32], as well as the coded exposure high-speed imaging [23]. Compared to CASSI and other snapshot spectral imagers, our system not only enhances the overall light efficiency, but also carves out a new way to rely on multiple cameras for the challenging task of capturing HSHS videos. The contribution of this paper goes beyond the system. We also propose to learn an over-complete 3D dictionary from the panchromatic video to represent each band-wise video sparsely, which copes well with a robust computational reconstruction to recover the underlying HSHS video. The improvement over simply using the joint videos further justifies the dual-camera design. Note that the dual-camera design for CASSI was first investigated in our previous work [36], but it was limited to single-frame image reconstruction directly using the measurements.

## 3. System principles

A schematic of our proposed system is shown in Fig. 1 and the data flow in this system is detailed in Fig. 2. As can be seen, there are two branches after the beam splitter. In the PanCam branch, there is simply an objective lens in front of the detector and thus the light path is short and unobstructed. In the CASSI branch, light is first encoded by a coded aperture and then dispersed by a dispersive prism before reaching the detector, which results in considerable

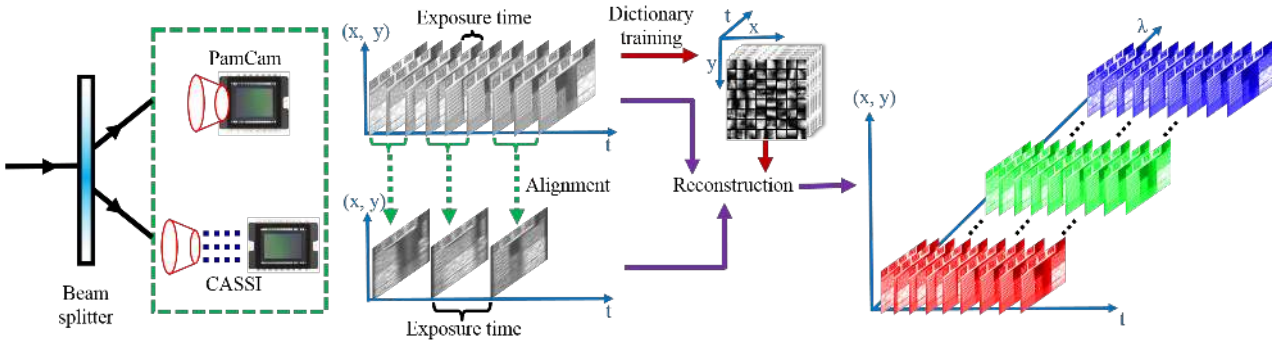


Figure 2. Data flow in the dual-camera system. The CASSI branch captures a low frame-rate video while the PanCam branch captures a high frame-rate video simultaneously. The PanCam measurements are then used to train an over-complete 3D dictionary, together with which the underlying 4D HSHS video is reconstructed from the joint videos.

light intensity attenuation. Suppose the detectors in the two branches are identical, the PanCam branch can work at a much higher frame rate than the CASSI branch in practice, due to higher efficiency of light utilization. That is to say, the PanCam branch lacks in spectral resolution while the CASSI branch lacks in temporal resolution. Therefore, it is possible to recover 4D HSHS videos by jointly using the measurements from the two branches, under elaborate calibration and synchronization.

### 3.1. Light efficiency analysis

We start with a light efficiency analysis of our system in comparison with existing hyperspectral video acquisition systems. It is an increasing trend to optically encode the spectral information and then recover it through computational reconstruction. Besides elegant interpretation with the mathematical model, light efficiency plays a crucial role in determining the speed performance of the system. As the representative work in this direction, we evaluate the light efficiency of CASSI [35], PMVIS [15], HVIS [10], and DCSI [22] along with our system. Specifically, We measure the light efficiency by the overall light transmission percentage of the whole system, which is mainly determined by the light-blocking elements in the optical path. Note that we only consider ideal optical elements here, which may suffer from some deviation in practice.

Considering to capture a hyperspectral video with $\Omega$ spectral bands, the light efficiency of different systems is summarized in Table 1. CASSI involves one coded aperture with 50% light transmission. PMVIS incorporates an occlusion mask which downsamples the spatial resolution by a factor of $\Omega$ to obtain the desired spectral resolution and thus sacrifices considerable light intensity. In two-branch HVIS, one branch shares the same light efficiency as in PMVIS and the other branch employing an RGB camera contributes $1/3$ ligh transmission due to the Bayer pattern. DCSI provides an overall light efficiency of 0.25 due to the DMD and LCOS employed, each with 50% light transmission. In our system, the CASSI branch provides 50% light transmission and the PamCam branch reaches 100% light transmission. Considering the employed beam splitter, our system achieves an overall light efficiency of 0.75. This is the highest among existing hyperspectral video acquisition systems, making it most efficient for capturing HSHS videos.

Table 1. Light efficiency comparison of different hyperspectral video acquisition systems.

| CASSI [35] | PMVIS [15] | HVIS [10] | DCSI [22] | Ours |
|---|---|---|---|---|
| 0.5 | $1/\Omega$ | $0.5(1/\Omega+1/3)$ | 0.25 | 0.75 |

### 3.2. Formulation

Let $f(x, y, \lambda, t)$ denote the scene information of a 4D HSHS video clip in its discrete form, where $1 \leq x \leq W$ and $1 \leq y \leq H$ index the spatial coordinates, $1 \leq \lambda \leq \Omega$ indexes the spectral coordinate, and $1 \leq t \leq K$ indexes the temporal coordinate. Without loss of generality, we assume the PanCam has $K$ exposures and the CASSI has only one exposure corresponding to this video clip. In other words, the PanCam captures $K$ frames while the CASSI captures one frame, equivalent to an acceleration rate of $K$. Since the beam splitter equally divides the incident light, the high frame-rate PanCam image captured at time $t$ can be written as

$$g^p(x, y, t) = 0.5 \sum_{\lambda=1}^{\Omega} w(\lambda) f(x, y, \lambda, t), \tag{1}$$

where $w(\lambda)$ is the spectral response function of the detector. This equation can be rewritten in a linear matrix form as

$$G_t^p = \Phi^p F_t, \tag{2}$$

where $G_t^p$ and $F_t$ are the vectorized representation of $g^p$ and $f$ at time $t$, and $\Phi^p$ is the time-invariant observation matrix of the PanCam (determined by $w(\lambda)$).

On the other hand, the low frame-rate CASSI image captured during the whole clip can be written as

$$g^c(x,y) = 0.5\sum_{t=1}^{K}\sum_{\lambda=1}^{\Omega} w(\lambda)S(x,y-\phi(\lambda))f(x,y-\phi(\lambda),\lambda,t), \tag{3}$$

where $S(x,y)$ denotes the transmission function of the coded aperture and $\phi(\lambda)$ denotes the wavelength-dependent dispersion function of the prism. (Please refer to [35] for a detailed formulation of the CASSI measurement.) Similar to the PanCam branch, the output of the CASSI branch can be rewritten as

$$G^c = \Phi^c F, \tag{4}$$

where $G^c$ is the vectorized representation of $g^c$, $F = (F_1, F_2, \ldots, F_K)^T$ is the temporal concatenation of the original HSHS video, and $\Phi^c$ is the observation matrix of the CASSI (jointly determined by $w(\lambda)$, $S(x,y)$, and $\phi(\lambda)$).

The dual-camera system model can then be expressed as

$$\begin{pmatrix} G^c \\ G_1^p \\ G_2^p \\ \vdots \\ G_K^p \end{pmatrix} = \begin{pmatrix} \Phi^c & & & \\ \Phi^p & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \Phi^p & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \Phi^p \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_K \end{pmatrix}. \tag{5}$$

A more simplified expression will be

$$G = \Phi F, \tag{6}$$

where $G$ includes all the measurements and $\Phi$ is a sparse matrix representing the overall system forward operation.

## 4. Dictionary-based reconstruction

The total number of measurements from the PanCam branch and the CASSI branch is $W \times H \times K + W \times (H + \Omega - 1)$[1], which is significantly smaller than the dimension of the unknown data $W \times H \times \Omega \times K$. Recovering the original HSHS video $F$ from its incomplete measurements $G$ thus remains an ill-posed inverse problem. Thanks to the compressive sensing theory [8, 14], $F$ can be recovered by seeking an over-complete dictionary $D$ on which it can be sparsely represented. We can then solve the following minimization problem instead

$$\hat{\alpha} = \arg\min_{\alpha} ||G - \Phi D \circ \alpha||_2^2 + \tau||\alpha||_0, \tag{7}$$

where $\alpha$ is the concatenation of the sparse coefficients of all patches in $F$ when represented on $D$, the operation $\circ$ derives $F$ from $D$ and $\alpha$, and $\tau$ is a regularization parameter balancing the data fidelity and the prior sparsity[2].

[1]Please refer to [34] for the number of measurements in CASSI.

[2]Mathematically, $\alpha$ needs to be reorganized to multiply with $D$ and their product needs to be reorganized again to give $F$. Please refer to [13] for details on this operation.

Following the recent advances in the information theory community, a proper dictionary indicates that a good reconstruction can be obtained with a high probability [9]. Meanwhile, we notice that the 4D HSHS video can be treated as a spectral concatenation of multiple band-wise videos which often have similar structural content (*e.g.*, edges) to the panchromatic video. To validate this observation, we conduct a statistical experiment on a hyperspectral database [38]. For each 6x6 image patch in all band-wise images, we calculate the normalized root mean squared error with its most similar patch from the corresponding panchromatic image. There are 98.1% and 93.4% patches with less than 0.04 and 0.02 errors respectively, which reveals a high degree of similarity. The same observation applies to the panchromatic video and the band-wise videos as long as they are well calibrated and synchronized. Therefore, the output video of the PanCam in our system can be readily used to train an over-complete 3D dictionary to sparsely represent each band-wise video.

To train the over-complete dictionary, we randomly sample a large number of 3D patches sized $m = w \times h \times k$ from the panchromatic video. The dictionary $D \in \mathbb{R}^{m \times n}$ can then be derived by KSVD [2], where $n\,(n > m)$ is the number of atoms (vectorized 3D patches) remaining in the dictionary. Note that the data we use to train the dictionary is self-provided and highly correlated to the underlying HSHS video, which ensures that the dictionary yields high sparsity when representing the band-wise videos. Once the dictionary is learned, the 4D HSHS video can be sparsely represented as

$$F = (F_1, F_2, \ldots, F_\Omega)^T = D \circ (\alpha_1, \alpha_2, \ldots, \alpha_\Omega)^T = D \circ \alpha, \tag{8}$$

where $F_\lambda (1 \le \lambda \le \Omega)$ denotes a band-wise video, $\alpha_\lambda$ denotes the sparse coefficient vector that represents $F_\lambda$ on $D$, and $\alpha$ can be regarded as the concatenation of $\alpha_\lambda$. Substituting Eq. 8 into Eq. 7, the optimization problem of our system can be efficiently solved by employing the orthogonal matching pursuit algorithm [33].

## 5. Simulation

In this section, simulations are conducted to evaluate the performance of the proposed approach in principle. The test data we use come from the synthetic hyperspectral video reported in [27]. We remove some deteriorated bands with heavy noise and take out the moving region of the scene. Specifically, a 4D hyperspectral video clip with the dimension of $256(W) \times 256(H) \times 20(\Omega) \times 24$ is selected. Three different acceleration rates $K = 2, 4, 8$ are tested, respectively. For example, when $K = 8$, there are 24 PanCam frames and 3 CASSI frames synthesized as measurements, from which a total of $20 \times 24$ hyperspectral frames need to be recovered. The spectral response function of the detector

Table 2. Quantitative evaluation of three reconstruction methods.

| $K$ | Method | Spatial Metric | | Spectral Metric | |
|---|---|---|---|---|---|
| | | PSNR | SSIM | RMSE | SAM |
| 2 | CASSI-TI | 26.99 | 0.611 | 0.196 | 0.071 |
| | TwIST | 32.94 | 0.951 | 0.089 | 0.059 |
| | DBR | 34.25 | 0.961 | 0.075 | 0.050 |
| 4 | CASSI-TI | 26.45 | 0.573 | 0.214 | 0.081 |
| | TwIST | 32.15 | 0.932 | 0.093 | 0.061 |
| | DBR | 33.06 | 0.940 | 0.084 | 0.057 |
| 8 | CASSI-TI | 23.08 | 0.447 | 0.317 | 0.124 |
| | TwIST | 28.56 | 0.912 | 0.136 | 0.089 |
| | DBR | 29.17 | 0.922 | 0.126 | 0.085 |

is borrowed from the real one used in our experiments. The transmission function of the coded aperture is generated as a random Bernoulli distribution with $p = 0.5$. The dispersion function of the prism utilizes a linear distribution for simplicity.

The parameters used in the dictionary-based reconstruction (DBR) are chosen empirically as below. For training the dictionary, the 3D patch size is set as $m = 6 \times 6 \times K$, and 60000 patches are randomly sampled from the panchromatic video as the input of KSVD. After KSVD, there are $n = 4m$ atoms remaining in the dictionary. The maximum iteration number of DBR is set to 30 and $\tau$ is set to 0.01. For comparison, we generate the reconstruction results using the two-step iterative shrinkage/thresholding (TwIST) algorithm along with the total variation (TV) regularizer [6], which is generally used for CASSI reconstruction. The parameters for TwIST are properly tuned. To provide a direct comparison with traditional CASSI, we also generate the temporal interpolation results of CASSI reconstruction (CASSI-TI) using a publicly available tool Twixtor in Adobe After Effects [1].

We use four quantitative image quality metrics to evaluate the performance of the reconstruction results, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [37], root mean squared error (RMSE), and spectral angle mapping (SAM) [21]. PSNR and SSIM are calculated based on each 2D spatial image, which measure the spatial fidelity between the reconstruction results and the original hyperspectral video. A larger value of these two metrics indicates a higher fidelity reconstruction. RMSE and SAM are calculated based on each 1D spectrum vector, which measure the spectral fidelity of the reconstruction. A smaller value of these two metrics suggests a better reconstruction. All metrics are averaged across the remaining dimensions.

The quantitative results for three different acceleration rates are shown in Table 2. It can be seen that the CASSI-TI results suffer from large deviation from the original test data. In contrast, both TwIST and DBR decently recover the 4D hyperspectral video even under $K = 8$, which validates the superiority of our proposed dual-camera design. On the other hand, with respect to all the four metrics, our proposed DBR outperforms TwIST under all acceleration rates, which demonstrates the effectiveness of the self-learned dictionary.

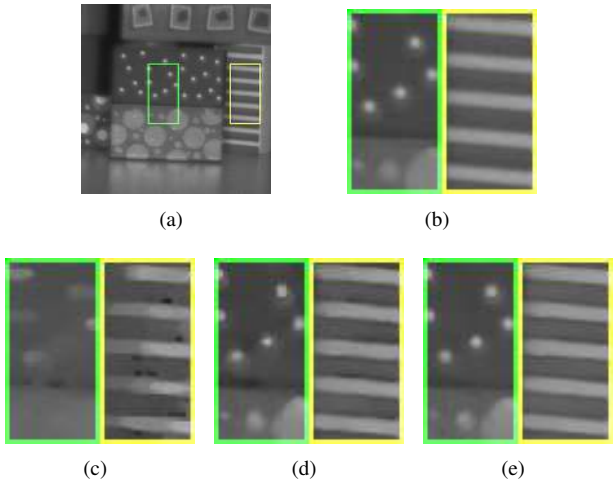


Figure 3. Reconstruction results of one selected band at a certain temporal location under $K = 2$. (a) Original frame. (b) Cropped regions from (a). (c) CASSI-TI. (d) TwIST. (e) DBR. (Please see the electronic version for better visualization.)

To further demonstrate the performance of the proposed approach, Fig. 3 shows the reconstruction results of one selected band at a certain temporal location under $K = 2$. While CASSI-TI introduces noticeable artifacts, the original hyperspectral frame is well recovered with the dual-camera design through either TwIST or DBR. Still, the latter using the self-learned dictionary achieves better perceptual quality, especially for the object details.

## 6. Experiments

### 6.1. System setting

**System components.** Fig. 4 demonstrates the prototype system we have developed for 4D HSHS video acquisition. The incident light is equally divided by a beam splitter and captured by two branches. In the Pan-Cam branch, a panchromatic detector (PointGrey FL3-U3-13Y3M-C) equipped with an 8mm objective lens is used, which can capture up to 150fps video at a maximum resolution of 1280×1024 pixels. In the CASSI branch, an 8mm objective lens is used to project the scene onto a coded aperture. The manufactured coded aperture is a random binary pattern with 300×300 elements and each element has a size

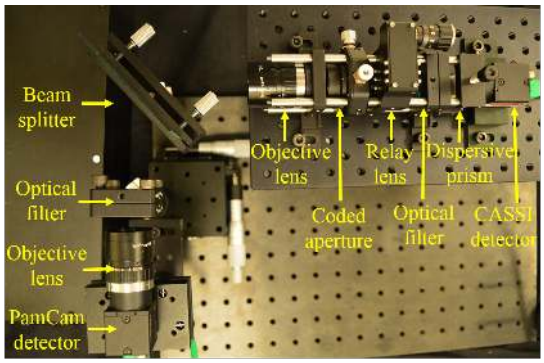


Figure 4. Prototype system for HSHS video acquisition.

of 10um $\times$ 10um. A doublet-Amici prism vertically disperses the spectral information with the center wavelength at 550nm. The CASSI detector has same model as the one in the PanCam branch. In our experiments, each element on the coded aperture is mapped to 2$\times$2 pixels on the detector by a relay lens (Edmund 45-762), so the effective spatial resolution is $600 \times 600$ pixels. In addition, an optical filter with a passband from 450nm to 650nm is used in each branch to restrict the spectrum to a certain range.

**System calibration.** There are two parts of calibration: the CASSI itself and that between the two branches. The CASSI calibration has already been well studied in [35] and will not be detailed here. Following the same procedures, we can obtain the observation matrix of the CASSI, and the whole spectrum spanning over the passband of the optical filter is discretized into 28 bands with different intervals. The calibration between the PanCam and the CASSI is essential for our system. Owing to the CASSI calibration, we only need to align the PanCam image with the projection of one wavelength on the CASSI image plane, and the alignment with other wavelengths can then be easily deduced. To this end, we place an auxilliary coded aperture in front of the beam splitter to act as an objective scene. This auxilliary coded aperture is illuminated by monochromatic light and captured by the two detectors. Once the optical elements in the CASSI branch are fixed, we fine tune the position of the PanCam so that the auxilliary coded aperture occupies an area with the same resolution on the two detectors. Then, under the illumination at one specified wavelength, the two captured images jointly determine the correspondence between the measurements of the two branches.

**System synchronization.** Synchronization of the CASSI and the PanCam is also essential to temporally align the output sequences of the two branches. Given an acceleration rate $K$, there should be $K$ exposures of the PamCam during one exposure of the CASSI, and their starting time should be synchronized exactly. To this end, we use a signal generator (RIGOL 1022D) to trigger the two branches with the same impulse signal. The frame rate of the two branches is preset by software.

Table 3. RMSE of spectral signatures in the center of three shapes.

| Shape | $K$=5 | $K$=10 | $K$=20 | CASSI |
|---|---|---|---|---|
| Circle | 0.037 | 0.042 | 0.043 | 0.060 |
| Rectangle | 0.031 | 0.034 | 0.044 | 0.070 |
| Diamond | 0.043 | 0.044 | 0.045 | 0.085 |

## 6.2. Qualitative and quantitative evaluation

To evaluate the performance of the proposed approach, we first test a simple scene that consists of three fast moving shapes with distinct colors displayed on an LCD screen. Three sets of experiments are conducted under different acceleration rates $K = 5, 10, 20$. Since the screen brightness is limited, to reach a proper exposure, the exposure time of the CASSI needs to be 200ms (*i.e.*, 5fps) under the maximum aperture of the objective lens. Correspondingly, the exposure time of the PanCam is set to 40ms (25fps), 20ms (50fps), and 10ms (100fps), under different apertures. Two sets of CASSI and PanCam measurements along with the color scene are shown in Fig. 5(a)-(b). The PanCam measurements are then used to train an over-complete dictionary for the reconstruction, where the parameters are chosen empirically as in the simulation. The reconstruction results of three selected bands at one temporal location under different acceleration rates are shown in Fig. 5(c). It can be seen that the proposed approach faithfully recovers the scene content even under $K = 20$. On the other hand, Fig. 5(d) shows the reconstruction results of one selected band at five temporal locations under $K = 20$, from which we can see that the high-speed motion is well recovered by the proposed approach. For comparison, we also generate a 5fps hyperspectral video directly from the CASSI measurements through the TwIST reconstruction. As shown in Fig. 5(c)-(d), this low frame-rate reconstruction suffers from blurring artifacts caused by the fast motion[3].

For a quantitative evaluation on the reconstruction quality, we further compare the averaged spectral signatures in the center areas of the three moving shapes. The reference values are measured by a scanning spectrometer (StellarNet BLK-CXR-SR-50 with 1.3nm spectral resolution). The spectral signature is normalized by the total energy of each area. Fig. 6 shows the comparison results in three areas. We can see the spectral signatures well match the reference under all acceleration rates, which indicates the high spectral fidelity of our reconstruction. In contrast, there is larger deviation for the direct CASSI reconstruction. Meanwhile, we calculate the RMSE of these signatures with respect to the reference in Table 3. It can be seen, as the acceleration rate gets higher, the reconstruction error also increases due

[3]In fact, a single CASSI without the beam splitter should achieve 10fps hyperspectral video for this scene, so this comparison is just for reference.

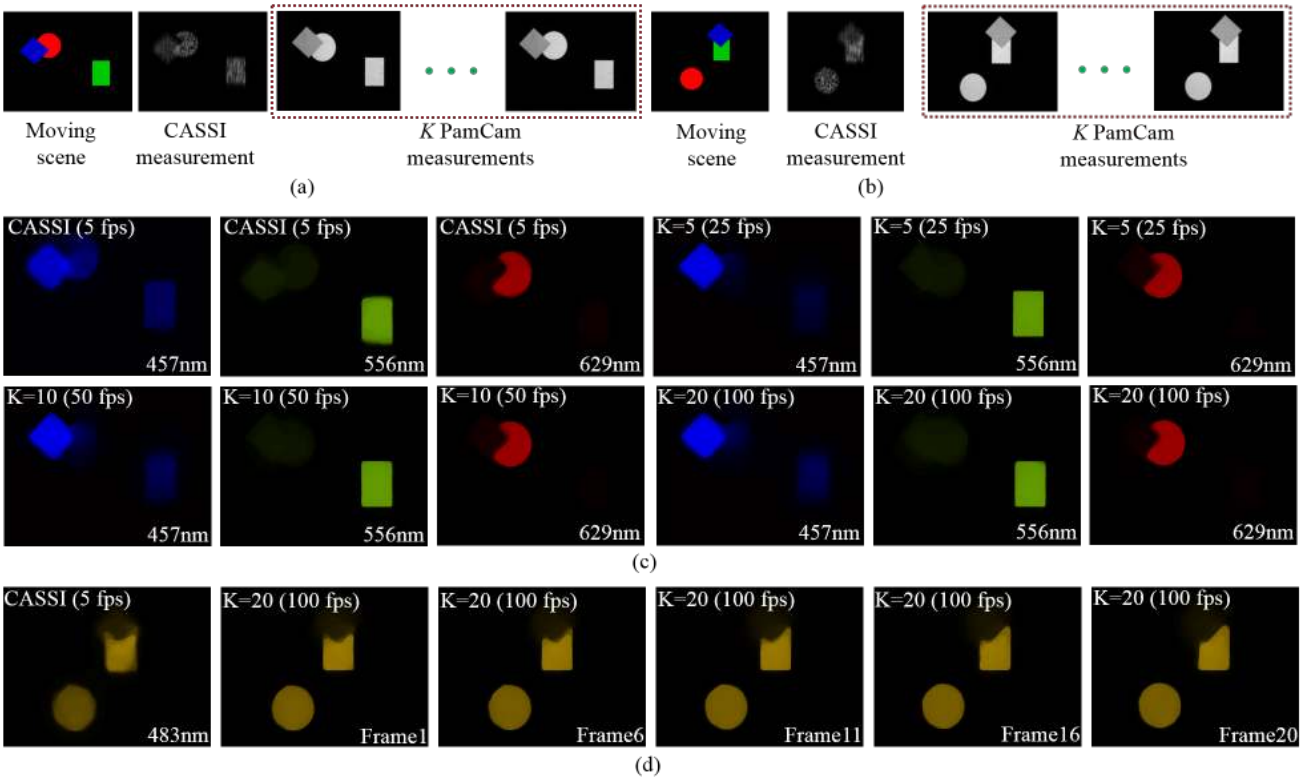


Figure 5. HSHS video reconstruction results of a simple scene consisting of three fast moving shapes with distinct colors displayed on an LCD screen. (a)-(b) Two sets of CASSI and PanCam measurements along with the color scene. (c) Results of three selected bands at one temporal location corresponding to (a). Direct CASSI reconstruction and ours under $K = 5, 10, 20$ are compared. (d) Results of one selected band at different temporal locations corresponding to (b). Direct CASSI reconstruction and ours under $K = 20$ are compared. (Please see the electronic version for better visualization.)

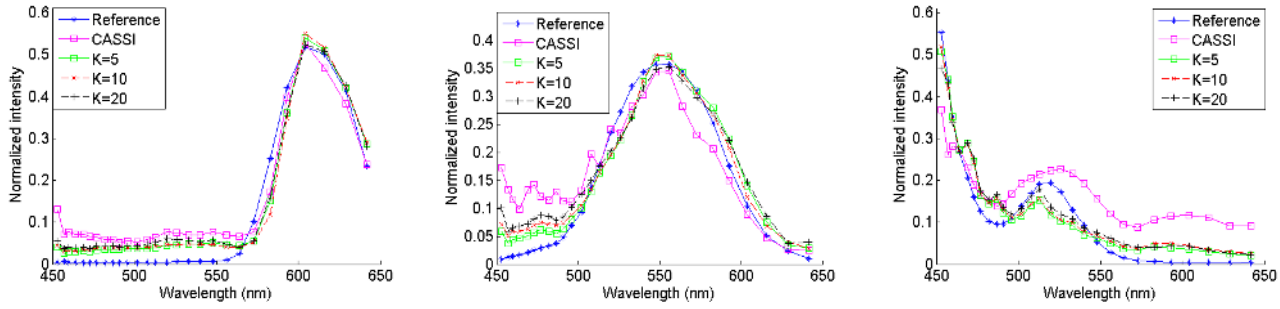


Figure 6. Spectral signature comparison in the center areas of the three moving shapes: (a) circle, (b) rectangle, and (c) diamond.

to the increasing number of unknowns in the reconstruction. Still, the RMSE values are fairly small even under $K = 20$ compared to the direct CASSI reconstruction, which validates the superior performance of the proposed approach.

### 6.3. Comparison with temporal interpolation

We then test the proposed approach on a doll with rich details moving fast on a stage, under ordinary indoor illumination. The exposure times for the CASSI and the PanCam are set to 100ms (10fps) and 10ms (100fps) respectively, equivalent to an acceleration rate of 10. As mentioned above, the CASSI measurements can be used to reconstruct a 10fps hyperspectral video. However, this low frame-rate video will inevitably be deteriorated by motion blur due to the long exposure time. To provide a baseline for evaluating our reconstruction results, we temporally in-

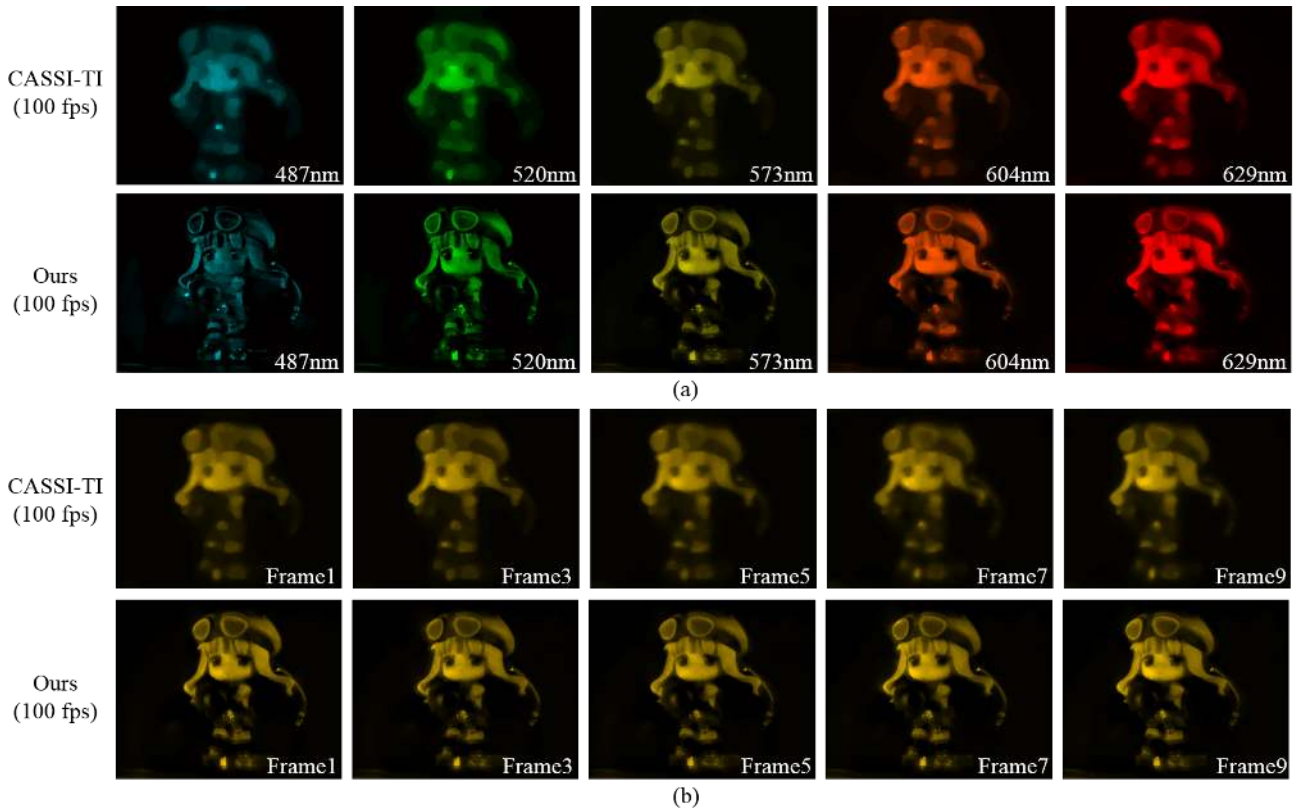


Figure 7. HSHS video reconstruction results of a fast moving doll under ordinary indoor illumination. Exposure times for CASSI and PanCam are 100ms and 10ms, respectively. (a) Results of different selected bands at one temporal location. (b) Results of one selected band at different temporal locations. (Please see the electronic version for better visualization.)

terpolate the 10fps reconstruction of CASSI to 100fps using Twixtor.

Fig. 7 shows a part of the comparison results. In Fig. 7(a), the reconstruction results of several selected bands at one temporal location are displayed. It can be seen that the CASSI-TI results suffer from blurring artifacts, while our reconstruction results contain more detailed scene information (*e.g.*, clearer edges). In Fig. 7(b), the reconstruction results of one selected band at several temporal locations are compared, which again demonstrates the superior performance of the proposed approach over CASSI-TI.

## 7. Conclusion and discussion

In this paper, we have made the first effort in 4D HSHS video acquisition. Specifically, we have designed a novel dual-camera system with enhanced overall light efficiency and developed a robust computational reconstruction by using a self-learned dictionary. Simulation and experimental results validate the superiority of the proposed approach.

Our current system has some limitations, which are considered as the future work. First, since the two branches have different exposure times and potentially different apertures, the difference of PSFs may become noticable and thus influence the performance when target scenes have a large diversity in depth. This requires more effort for the calibration. Second, as can be observed from the experimental results, there are still some reconstruction errors especially around the edges and spectral discontinuities. Further exploiting the correlation among different spectral bands may help improve the reconstruction quality. Last, compared with PMVIS and HVIS, one shortcoming of our system is that the reconstruction cannot be performed in realtime. We plan to investigate using parallel computation to improve the reconstruction speed.

## Acknowledgments

This work was partially supported by the National Science Foundation of China (Nos. 61227004 and 61425026), and the Fundamental Research Funds for the Central Universities of China (No. WK3490000001).

## References

[1] https://www.adobe.com/products/aftereffects.html.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse

representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, 2006.

[3] G. Arce, D. Brady, L. Carin, H. Arguello, and D. Kittle. Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Process. Mag.*, 31(1):105–115, 2014.

[4] R. W. Basedow, D. C. Carmer, and M. E. Anderson. Hydice system: Implementation and performance. In *Proc. SPIE*, 1995.

[5] M. Ben-Ezra and S. K. Nayar. Motion deblurring using hybrid imaging. In *CVPR*, 2003.

[6] J. Bioucas-Dias and M. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. Image Process.*, 16(12):2992 – 3004, 2007.

[7] D. Brady. *Optical Imaging and Spectroscopy*. John Wiley, Sons Inc., 2008.

[8] E. Candes. Compressive sampling. In *Int. Congr. Math.*, 2006.

[9] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[10] X. Cao, X. Tong, Q. Dai, and S. Lin. High resolution multispectral video capture with a hybrid camera system. In *CVPR*, 2011.

[11] A. Chakrabarti and T. Zickler. Statistics of real-world hyperspectral images. In *CVPR*, 2011.

[12] M. Descour and E. Dereniak. Computed-tomography imaging spectrometer: experimental calibration and reconstruction results. *Appl. Opt.*, 34(22):4817–4826, 1995.

[13] W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.*, 22(4):1620–1630, 2013.

[14] D. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[15] H. Du, X. Tong, X. Cao, and S. Lin. A prism-based system for multispectral video acquisition. In *CVPR*, 2009.

[16] N. Gat. Imaging spectroscopy using tunable filters: a review. In *Proc. SPIE*, 2000.

[17] N. Gat, S. Subramanian, J. Barhen, and N. Toomarian. Spectral imaging applications: remote sensing, environmental monitoring, medicine, military operations, factory automation, and manufacturing. In *Proc. SPIE*, 1997.

[18] M. Gehm, R. John, D. Brady, R. Willett, and T. Schulz. Single-shot compressive spectral imaging with a dual-disperser architecture. *Opt. Express*, 15(21):14013 – 14027, 2007.

[19] J. F. James. *Spectrograph design fundamentals*. Cambridge University Press, 2007.

[20] M. H. Kim, T. A. Harvey, D. S. Kittle, H. Rushmeier, J. Dorsey, R. O. Prum, and D. J. Brady. 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. In *SIGGRAPH*, 2012.

[21] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz. The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2):145–163, 1993.

[22] X. Lin, G. Wetzstein, Y. Liu, and Q. Dai. Dual-coded compressive hyperspectral imaging. *Opt. Lett.*, 39(7):2044–2047, 2014.

[23] D. Liu, J. Gu, Y. Hitomi, M. Gupta, T. Mitsunaga, and S. K. Nayar. Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):248–260, 2014.

[24] W. M. Borengasser and R. Watkins. *Hyperspectral romote sensing: principles and applications*. CRC, 2008.

[25] C. Ma, X. Cao, R. Wu, and Q. Dai. Content-adaptive high-resolution hyperspectral video acquisition with a hybrid camera system. *Opt. Lett.*, 39(4):937–940, 2014.

[26] A. Manakov, J. Restrepo, O. Klehm, R. Hegedus, E. Eisemann, H.-P. Seidel, and I. Ihrke. A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. In *SIGGRAPH*, 2013.

[27] A. Mian and R. Hartley. Hyperspectral video restoration using optical flow and sparse coding. *Opt. Express*, 20(10):10658 – 10673, 2012.

[28] H. V. Nguyen, A. Banerjee, and R. Chellappa. Tracking via object reflectance using a hyperspectral video camera. In *CVPRW*, 2010.

[29] T. Okamoto and I. Yamaguchi. Simultaneous acquisition of spectral image information. *Opt. Lett.*, 16(16):1277–1279, 1991.

[30] W. M. Porter and H. T. Enmark. A system overview of the airborne visible/infrared imaging spectrometer (aviris). In *Proc. SPIE*, 1987.

[31] Y. Schechner and S. Nayar. Generalized mosaicing: wide field of view multispectral imaging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(10):1334–1348, 2002.

[32] Y.-W. Tai, H. Du, M. S. Brown, and S. Lin. Image/video deblurring using a hybrid camera. In *CVPR*, 2008.

[33] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory,*, 53(12):4655–4666, 2007.

[34] A. Wagadarikar, R. John, R. Willett, and D. Brady. Single disperser design for coded aperture snapshot spectral imaging. *Appl. Opt.*, 47(10):B44 – B51, 2008.

[35] A. Wagadarikar, N. Pitsianis, X. Sun, and D. Brady. Video rate spectral imaging using a coded aperture snapshot spectral imager. *Opt. Express*, 17(8):6368 – 88, 2009.

[36] L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu. Dual-camera design for coded aperture snapshot spectral imaging. *Appl. Opt.*, 54(4):848–858, 2015.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[38] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Trans. Image Process.*, 19(9):2241–2253, 2010.