HAL
open science

Single Sensor Compressive Light Field Video Camera

Saghi Hajisharif, Ehsan Miandji, Christine Guillemot, Jonas Unger

To cite this version:

Saghi Hajisharif, Ehsan Miandji, Christine Guillemot, Jonas Unger. Single Sensor Compressive Light Field Video Camera. Computer Graphics Forum, 2020, pp.1-12. 10.1111/cgf.13944 . hal-02498719

HAL Id: hal-02498719

https://hal.science/hal-02498719v1

Submitted on 6 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

EUROGRAPHICS 2020 / U. Assarsson and D. Panozzo
(Guest Editors)

Volume 39 (2020), Number 2

Single Sensor Compressive Light Field Video Camera

Saghi Hajisharif[1†] and Ehsan Miandji[2] and Christine Guillemot [2] and Jonas Unger[1]

[1] Linköping University, Sweden [2] Inria, Rennes, France

Figure 1: Reconstruction of a 6D light field video (middle) from 2D sensor images (left). Each 2D image contains samples from angular and spectral dimensions convolved with a random color-coded mask that changes per-frame. The proposed algorithm reconstructs the full resolution color light field video using a novel sensing model, together with a temporally-aware learned dictionary. On the bottom right, the magnified image of one angle of the reconstructed frame is shown with the corresponding ground truth on top.

Abstract

This paper presents a novel compressed sensing (CS) algorithm and camera design for light field video capture using a single sensor consumer camera module. Unlike microlens light field cameras which sacrifice spatial resolution to obtain angular information, our CS approach is designed for capturing light field videos with high angular, spatial, and temporal resolution. The compressive measurements required by CS are obtained using a random color-coded mask placed between the sensor and aperture planes. The convolution of the incoming light rays from different angles with the mask results in a single image on the sensor; hence, achieving a significant reduction on the required bandwidth for capturing light field videos. We propose to change the random pattern on the spectral mask between each consecutive frame in a video sequence and extracting spatio-angular-spectral-temporal 6D patches. Our CS reconstruction algorithm for light field videos recovers each frame while taking into account the neighboring frames to achieve significantly higher reconstruction quality with reduced temporal incoherencies, as compared with previous methods. Moreover, a thorough analysis of various sensing models for compressive light field video acquisition is conducted to highlight the advantages of our method. The results show a clear advantage of our method for monochrome sensors, as well as sensors with color filter arrays.

CCS Concepts

• ***Computer graphics*** → *Computational photography; Image compression;*

1. Introduction

Light field imaging is a rapidly emerging technology in computational photography. By capturing both the spatial and angular variations of the light rays incident onto the sensor(s), light fields open up for a range of novel applications ranging from computer vision and industrial applications to computer graphics, cinematography, and everyday photography. As a result, there has been extensive research, and development of methods and technology for capturing light fields [LH96, GGSC96] in the past two decades.

To date, the most common techniques for capturing the angular variations in the light field have been to use multiple cameras, [WJV*05], or to place an array of micro-lenses, [NLB*05], in front of an ordinary 2D sensor as in the Lytro and Raytrix cameras. However, multi-sensor systems lead to bulky, expensive, and oftentimes impractical setups, and the micro-lens approach leads to a reduced spatial resolution since a large portion of the budget of available pixels on the sensor needs to be sacrificed to sample the

† Corresponding author: email: firstname.lastname@liu.se

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

angular domain. Another challenge is to efficiently handle the high bandwidth data streams and very large memory footprints inherent to high-resolution light field imaging.

This paper presents a novel compressed sensing (CS) framework and evaluates a camera design for single sensor light field video capture and reconstruction. Compressed sensing theory, [CRT06a, Don06], postulates that if a signal is compressible (or sparse) in some basis, then it can be reconstructed from a very small number of samples (well below the Nyquist rate). Similar to Marwah et al. [MWBR13] and Miandji et al. [MUG18], we use a coded aperture design to optically construct a sensing matrix by placing an attenuation mask with random transmittance in front of the sensor. However, in contrast to [MWBR13] and [MUG18] who only considered static light fields, we extend the compressed sensing into the temporal domain and enable light field video capture. By changing the random pattern on the color mask between each consecutive frame, our CS algorithm can reconstruct a full 4D light field for each frame from the 2D compressive measurements (images) on the camera sensor. At each pixel on the camera sensor, the random color-coded mask acts as a convolution filter, convolving the angular and spectral information of the incident light field into a single pixel.

The main contribution of this paper is a new sensing model, where, in contrast to previous work, temporal information is taken into account for CS reconstruction. Our model addresses several key requirements that exist for achieving high-quality reconstructions. First, the dictionary used for sparse coding of the light field is trained on a small set of consecutive frames to utilize the sparsity in the spatial, angular, as well as temporal domains simultaneously. Second, the compressive measurements obtained using the random color-coded mask also include temporal information. The inclusion of the temporal domain leads to an increased number of incoherent (random) compressive samples, which significantly increases the reconstruction quality compared to existing methods for light field photography, see Section 4. Indeed, there exist various sensing models based on the design of the dictionary and the sensing matrix. We present and analyze such models and propose a new model that is vastly superior in terms of the quality of the reconstructed light field videos.

The main contributions of this paper are:

- Compressive light field video camera design with a temporal signal model.
- High quality reconstruction of full resolution light field videos from a video sequence captured using a single-sensor consumer-level camera.
- A study on the effect of monochrome and color sensors in light field video reconstruction quality.
- A study on various sensing models for compressive light field video cameras.
- A dictionary learning method enabling increased sparsity and temporal coherency of light field videos.

The evaluation shows that the algorithm presented in this paper produces significantly better visual quality as compared to the state-of-the-art. To the best of the authors' knowledge, this is the first CS light field capture and reconstruction algorithm with an explicit model leveraging from temporal coherence in the data.

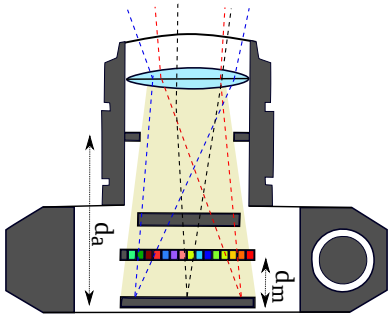


Figure 2: Light filed camera design with color-coded attenuation mask. The mask is placed between the aperture and the sensor at distance d_m from the sensor.

2. Related Work

One of the first attempts at capturing high-quality light fields was with multi-sensor systems, also known as camera arrays [WJV*05]. By utilizing camera parameters, the acquired images are re-projected to construct a light field. The angular and spatial resolution is limited by the number of cameras and their corresponding resolution. Indeed high-resolution light fields and light field videos can be captured using this setup; however, the high cost and the size of these capturing setups limit their usability. An alternative is to mount a single camera on a mechanical arm [LH96, UWH*03]. However, these light field imaging systems can only be used for static scenes.

A popular and well-established method for capturing light fields is through the utilization of micro-lens arrays. This technique was first introduced by [AW92], and was later implemented by [NLB*05]. The design is based on a dense array of small lenses that are placed in front of a sensor. Therefore, the size of each lenslet, together with the number of detector elements in the image sensor determines the angular and spatial resolution of the light field measurement. For instance, if each micro-lens covers an area of 8×8 pixels, then the spatial resolution of the light field is $1/8$th of the sensor resolution; see [GZC*06] for a more elaborate description of the spatio-angular trade-off in plenoptic cameras. Since micro-lens light field cameras are small in size and relatively cheap, they were the first to be commercialized. Both multi-sensor array and micro-lens array designs result in a massive amount of data, especially for light field videos, and require effective compression techniques after capturing [MHU19].

Recently, a new sensor technology was introduced [WGM09, WGM11], which uses an array of *angle sensitive pixels* (ASP). Each ASP is tuned to have a predefined angular response by utilizing the Talbot effect. Using computational photography, it is possible to obtain a light field from an array of ASPs [HSJ*14]. However, the reconstruction quality of these systems, at the current state, are not competitive with aforementioned techniques.

Another method for single sensor light field photography is based on *coded aperture* [VRA*07] capturing, where a transparent non-refractive random mask is placed on the aperture plane. Hence the sensor integrates randomly modulated light field views. Using the mask and the image formed on the sensor, one can use

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

various deconvolution techniques to obtain a light field. [LLW*08] achieved higher reconstruction quality by using a Liquid Crystal Array (LCA) to change the mask pattern in order to obtain multiple shots (and hence a higher number of samples from the original light field). A dual mask light field camera was proposed in [XL12] to improve the spatial resolution at the cost of reducing the light transmission to the sensor. Babacan et al. [BAL*12] proposed to place a randomly generated mask at the aperture of the camera and using a Bayesian framework, they reconstructed the light fields from the encoded images.

Deep learning methods have also been used for sparse coding and reconstruction of light fields [GL10, GJK*17]. However, the use of fully connected networks limits their capabilities to small patch sizes. In some cases, the coded mask is restricted to a fixed pattern to reduce the size of the parameter space [VCR*17]. Furthermore, these methods require large amounts of data for training which is typically not feasible in practice. Chen et al. [CC17] proposed a disparity aware dictionary learning, where first the disparity of the scene was calculated using sub-aperture scans, making the method only suitable for static scenes. Nabati et al. [NMG18] use a pre-defined coded mask and convolutional neural networks for recovering a light field from coded measurements. However, it is not clear how this method can be extended to multiple shots; moreover, it has been shown in [Mia18] that the compressive sensing leads to competitive results for a single shot, while significantly outperforming [NMG18] with two or more shots. An autoencoder network was proposed by Inagaki et al. [IKT*18] for learning a mask, based on the features of the scene, in order to reconstruct a light field image from the coded input using coded-aperture photography. Wang et al. [WZK*17] proposed a hybrid light field video capturing system consisting of a Lytro Illum and a DSLR camera that enabled high frame rate acquisition. Although their method achieves a high frame rate, it is still dependent on a two-camera system and suffers from low spatial resolution.

It has been shown that *compressive sensing* can be used to capture static light fields using a single sensor. Marwah et al. [MWBR13] introduced a compressive light field camera where the random mask is placed at a predefined small distance from the sensor. Since the mask is monochrome, each color channel of the light field is reconstructed independently. Therefore, correlations between color channels were not utilized. Miandji et al. [MUG18] proposed to place a color-coded mask in front of the color filter array (CFA) to encode the angular light rays. By utilizing multiple shots and the incoherence introduced by the color mask, significantly higher results are reported. Parallel to this work, Nabati et al. [NMG18] introduce a reconstruction algorithm based on deep neural networks for compressive light field photography using a color-coded mask. While the reconstruction quality is competitive with [MUG18] for a single shot, it is not clear how this method can be extended to multiple shot reconstruction. Compressive sensing has also been used for video acquisition [WLD*06] with coded aperture video representation [MW08] to enhance the resolution of the digital video. Hitomi et al. [HGG*11] developed a prototype imaging system with per-pixel coded aperture control and proposed to reconstruct the video by learning a sparse representation of the video frames with an over-complete dictionary. This paper extends the idea of exploiting temporal coherence in CS and combines this with a coded aperture to capture and reconstruct light field videos.

3. Compressive Light Field Video Acquisition

Since our method is based on the well-established field of compressed sensing [CRT06a, Don06], we start by a brief introduction to essential concepts related to compressed sensing in Section 3.1. This is followed by a review of compressive light field photography [MWBR13,MUG18] in Section 3.2, which utilizes compressed sensing for efficient acquisition of light field images on a single sensor. Compressed sensing comprises three main components: 1. a sensing matrix, 2. a dictionary, and 3. a reconstruction algorithm. Our goal in this paper is to design a sensing matrix and a dictionary such that we achieve high sparsity and measurement incoherence for faithful recovery of a light field video from merely the coded images that are formed on the 2D camera sensor.

Two sensing matrix designs are proposed in Section 3.3 for acquiring a light field video by modulating consecutive frames onto the sensor using a color-coded mask. These designs take into account the presence of a CFA on the sensor. In Section 3.4, we describe our dictionary training approach for light field videos. Finally, three sensing models based on three different sensing matrix configurations are presented in Section 3.5, together with their corresponding dictionary. These sensing models, called SM1, SM2, and SM3, are essentially three different approaches for reconstructing a light field video from compressive measurements. Our results in Section 4 show that SM3, where the temporal information is utilized in both the sensing matrix and the dictionary, performs the best.

3.1. Compressed Sensing

Let $\mathbf{x} \in \mathbb{R}^n$ be a deterministic vector representing a bandlimited continuous-time signal. Our goal is to sample $\mathbf{x}$ with minimal number of samples while admitting the exact recovery of $\mathbf{x}$. Let $\Phi \in \mathbb{R}^{s\times n}$, $s < n$, be a sampling operator that takes s linear samples from $\mathbf{x}$, i.e. $\mathbf{y} = \Phi\mathbf{x}$. The sampling operator is often referred to as a *measurement* or *sensing* matrix in the field of compressed sensing. It is clear that recovering $\mathbf{x}$ from the measurements $\mathbf{y}$ requires solving a linear system of equations

$$\arg\min_{\mathbf{x}} \ \|\mathbf{y} - \Phi\mathbf{x}\|_2^2. \tag{1}$$

However, equation (1) has infinitely many solutions since $s < n$. Therefore, we need to limit the space of solutions by considering a prior on the signal $\mathbf{x}$. One such prior is sparsity. Assume that $\mathbf{x}$ is sparse in a suitable dictionary $\mathbf{D} \in \mathbb{R}^{n\times k}$, i.e. we can write $\mathbf{x} = \mathbf{D}\theta$ such that $\|\theta\|_0 \leq \tau$; in other words, we require that the vector θ to have at most τ nonzero elements. Using this assumption, equation (1) becomes:

$$\arg\min_{\theta} \|\theta\|_0 \ \text{ s.t. } \ \|\mathbf{y} - \Phi\mathbf{D}\theta\|_2^2 \leq \varepsilon, \tag{2}$$

where ε is a small constant often related to the noise power. There exists a large number of algorithms for solving (2) and its ℓ_1 variant [NT09, DTDS12, SZ11, YZ11, LSQQ13, YZ11], as well as a large body of research on required conditions for obtaining the exact recovery of $\mathbf{x}$ [MEUA17, GN03, DE03, CRT06b, Tro04].

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

Compressed sensing is based on two main principles: *sparsity* and *incoherence*. Sparsity is the most important parameter in defining the required number of samples for faithful recovery of a signal and incoherence requires that if $\mathbf{x}$ is sparse in $\mathbf{D}$, it should be dense in Φ. Incoherence is closely related to the uncertainty principle [DH01]. Since $\mathbf{D}$ is deterministic, a common approach in improving incoherency is to define Φ as random matrices, e.g. with independent and identically distributed Gaussian entries. Moreover, when using random sensing matrices, another important factor for improving incoherency is the number of samples, s. One of the key contributions of this paper is to take multiple random samples along the time domain to improve the incoherency of the measurements.

3.2. Compressive Light Field Acquisition

A Light field can be described by the two-plane parameterization as $\mathcal{L}(s,t,u,v,\lambda)$ [LH96, GGSC96], where (s,t) and (u,v) denote the spatial and angular coordinates, respectively, and λ parameterizes the spectral domain. By adding the time domain, f, we obtain a light field video, which is a 6D function $\mathcal{L}(s,t,u,v,\lambda,f)$. A 2D image using a conventional photograph is captured by integrating light rays over the angular domain of the light field projected onto the camera sensor

$$\mathbf{y}(s,t,\lambda) = \int_{u,v} \mathcal{L}(s,t,u,v,\lambda)\cos^4\alpha \, du dv, \tag{3}$$

where α is the angle between the ray and the sensor and $\cos^4\alpha$ represents the vignetting effect [Ray02], which we omit in order to simplify our design methodology. Marwah et al. [MWBR13] suggested to place a monochrome coded attenuation mask Φ at a distance d_m from the sensor to optically modulate the light field and project it onto the sensor as shown in figure 2. Using this model, equation (3) can be written as

$$\mathbf{y}(s,t,\lambda) = \int_{u,v} \mathcal{L}(s,t,u,v,\lambda) a(s+\sigma(u-s), t+\sigma(v-t)) du dv, \tag{4}$$

where the function $a(.)$ defines the attenuation mask and $\sigma = d_m/d_a$ defines the shear of the mask pattern. In discrete form, equation (4) can be written as matrix-vector multiplication as follows

$$\mathbf{y} = \begin{bmatrix} \Phi^1 & \Phi^2 & \dots & \Phi^\nu \end{bmatrix} \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^\nu \end{bmatrix}, \tag{5}$$

where $\nu = |u||v|$ is the angular resolution, $\Phi^i \in \mathbb{R}^{\omega\times\omega}$ are diagonal matrices representing the mask model in (4), $\omega = |s||t|$ is spatial resolution, and $\mathbf{x}^i \in \mathbb{R}^\omega$ contains the vectorized light field view images. With a monochrome mask, the sensing matrix Φ is applied to each color channel independently, hence increasing the coherence of the measurements, which in turn reduces the quality of reconstruction, as shown in [MUG18].

3.3. Sensing Matrix Design for Light Field Videos

In this section, we describe our approach for designing a sensing matrix corresponding to a compressive light field video camera. To this end, we consider two sensing matrix designs. First, we describe the configuration of a sensing matrix corresponding to a sensor with a Color Filter Array (CFA), together with a color-coded mask placed in front of the sensor at a predefined distance. The model assumes the color measurements recorded after the CFA to be pre-demosaiced such that each measurement has three color components. Second, a sensing matrix design is presented that assumes a monochrome sensor, together with a color-coded mask at a predefined distance from the sensor.

3.3.1. CFA-equipped Sensor with Color Mask

Miandji et al. [MUG18] proposed a random color mask placed at a small distance to the aperture of a sensor equipped with color filter array (CFA); see Figure 3(a) for an example of a sensor image captured using this setup. Acquiring multiple shots from the scene further increases the incoherence in the measured light field, which leads to higher quality reconstruction. However, in the light field video, capturing multiple shots is not possible due to movements in the scene. In a compressive light field video camera based on the design of [MUG18], for each frame, a single 2D image $\mathbf{y}_i$ is formed on the sensor using a unique mask pattern. The mask pattern changes by moving the mask or the sensor using a piezo motor. The question is: How should we design Φ based on the moving mask such that we make use of the temporal coherence between frames? Indeed, since the capturing frame rate is limited by the capabilities of the camera (which exceeds hundreds of frames even on modern smartphones), we can expect significant correlations between the consecutive frames that can be utilized in the reconstruction.

Without loss of generality we assume three colors as RGB to simplify the notations. Let us define the sensing matrix for frame $j \in \{1,\dots,N\}$, where N is the total number of frames, using a color mask and sensor CFA as follows

$$\Psi^j = \begin{bmatrix} \Phi^{1,R,j}\dots\Phi^{\nu,R,j} & 0 & 0 \\ 0 & \Phi^{1,G,j}\dots\Phi^{\nu,G,j} & 0 \\ 0 & 0 & \Phi^{1,B,j}\dots\Phi^{\nu,B,j} \end{bmatrix} \tag{6}$$

This definition of the sensing matrix Ψ^j coincides with that of [MUG18]. We propose to utilize β consecutive frames and stack their corresponding measurement matrices Ψ^j, $i \in \{1,\dots,\beta\}$, vertically as follows

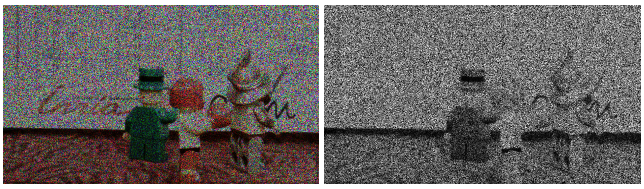
$$\Phi_{(\mathrm{I})} = \begin{bmatrix} \Psi^1 \\ \vdots \\ \Psi^\beta \end{bmatrix} \in \mathbb{R}^{\beta\omega\lambda\times\omega\nu\lambda}. \tag{7}$$

Alternatively, stacking β sensing matrices horizontally would result in the sensing matrix

$$\Phi_{(\mathrm{II})} = \begin{bmatrix} \Psi^1 \dots \Psi^\beta \end{bmatrix} \in \mathbb{R}^{\omega\lambda\times\beta\omega\nu\lambda}, \tag{8}$$

hence performing a linear combination of the input frames with a convolution filter over their angular domain into $y \in \mathbb{R}^{\omega\lambda}$. A clear advantage of (7) over (8) is that the former contains β-times more uncorrelated samples. Note that the number of samples is defined by the number of rows in Φ. Indeed if the consecutive frames are sufficiently similar to each other, the compressive random measurements obtained from β frames are highly incoherent (due to

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.



(a) Sensor with CFA (b) Monochrome Sensor

Figure 3: Raw images with a color-coded mask placed in front of (a) Sensor with CFA and (b) Monochrome Sensor.

the movement of the mask at each frame), and hence producing a higher reconstruction quality. This difference in the sensing model will be described in Section 3.5, and the superiority of (7) over (8) will be confirmed using our simulation results in Section 4.

3.3.2. Monochrome Sensor with Color Mask

Similar to [NMG18, Mia18], we also consider a camera design with a color mask placed in front of a monochrome sensor; see Figure 3(b) for an example of a sensor image captured using this setup. This setup leads to the compression of the angular domain as well as the spectral domain. Although the number of random measurements, and hence the incoherency, is reduced compared to the model using CFA, this setup is more practical in reality. This is because the light efficiency is higher when one mask is used instead of two, and the mask can be placed in a desired distance to the sensor. Another benefit of this design is the high compression ratio which can be used for fast transmission of the captured data, as well as reduced in-camera processing time due to the removal of the debayering and color correction processes, which leads to higher frame rates.

We formulate the sensing matrix for a single frame without a CFA as follows

$$\Lambda^j = \left[\Phi^{1,R,j} \dots \Phi^{\nu,R,j} \quad \Phi^{1,G,j} \dots \Phi^{\nu,G,j} \quad \Phi^{1,B,j} \dots \Phi^{\nu,B,j}\right] \quad (9)$$

Similar to sensing design with a CFA, here we can also construct two sensing matrices:

$$\Phi_{(\mathrm{III})} = \begin{bmatrix} \Lambda^1 \\ \vdots \\ \Lambda^\beta \end{bmatrix} \in \mathbb{R}^{\beta\omega \times \omega\nu\lambda}. \quad (10)$$

$$\Phi_{(\mathrm{IV})} = \left[\Lambda^1 \dots \Lambda^\beta\right] \in \mathbb{R}^{\omega \times \beta\omega\nu\lambda}, \quad (11)$$

Note that the sensing matrices in (7), (8), (10), and (11) do not require custom hardware implementations. Indeed we have assumed that the only data that is available as input to our method is the image formed on the sensor, as well as the mask values for the corresponding frame. Furthermore, the reconstruction method, see Section 3.5, works with both monochrome and CFA-equipped sensors with a color-coded mask. In Section 4, we compare the reconstruction quality of various data sets for both designs and discuss their advantages and disadvantages.

3.4. Dictionary Training for Light Field Videos

In this section, we describe our approach for training a dictionary that admits sparse representation of light field videos. Indeed the utilization of the temporal domain is of importance since it increases the sparsity due the correlation between a set of neighboring frames. The theory of compressed sensing states that a sparse signal with at most τ nonzero values can be reconstructed using Gaussian or Bernoulli random sensing matrices if $s \geq C\tau \ln(n/\tau)$, where C is a constant, n is the signal length, and s is the number of samples. Therefore, if we increase the sparsity, i.e. a smaller τ, it is expected that the required number of samples will decrease.

We use the online dictionary learning algorithm [MBPS10] on a training set $\mathbf{Z} = \{\mathbf{z}^1 \dots \mathbf{z}^t\}$ consisting of t light field video frames by solving

$$\min_{\mathbf{D}} \frac{1}{t} \sum_{i=1}^{t} \min_{\mathbf{h}^i} \frac{1}{2} \left\|\mathbf{z}^i - \mathbf{D}\mathbf{h}^i\right\|_2^2 + \lambda \left\|\mathbf{h}^i\right\|_0. \quad (12)$$

The aim of the dictionary learning algorithm is to find a dictionary $\mathbf{D}$ such that each training data $\mathbf{z}^i$ has a latent sparse representation $\mathbf{h}^i$. The non-negative coefficient λ in (12) defines a trade-off between reconstruction error and sparsity.

Solving (12) on the whole light field is not feasible, and therefore, we create smaller patches on the light field data set. The dimensionality of the patches affects the quality of the dictionary and how well it can represent the light field data. Four dimensional (4D) spatio-angular light field patches have shown to increase angular coherency in the reconstructed light field compared to 2D patches [MWBR13]. Expanding the patches to 5D to include the color information in each patch has shown superior results compared to 4D patches [MUG18]. We propose to add temporal information to the patches to include the spatial, angular, spectral, and temporal domain in our patches, which will also increase the dimensionality of the dictionary. As mentioned above, including the temporal domain in the patches would increase sparsity, and improve the reconstruction quality, as well as the temporal coherence of reconstructed light fields. With a slight abuse of notation, we use s, t, u, v, λ, and β, utilized in Section 3 for the resolution of a light field video, to denote the patch size. As a result, the dimensionality of a light field video patch is $n = s \times t \times u \times v \times \lambda \times \beta$, corresponding to the spatial, angular, spectral, and temporal resolution of the patch, respectively.

For training the dictionaries we considered two options: training a dictionary using 5D patches extracted from each individual frame, which we call a *single-frame* dictionary. Each atom of this dictionary is a basis function in spatial, angular and spectral domains. The other option is to extract 6D patches that spans the spatial, angular, spectral and temporal dimensions. We train on patches extracted from β-consecutive frames to train a *multi-frame* dictionary that has a structure as following:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_\beta \end{bmatrix}, \quad (13)$$

where $\mathbf{D} \in \mathbb{R}^{\beta\lambda\nu\omega \times \rho\beta\lambda\nu\omega}$ and ρ is the over-completeness factor. The

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

	Boxer		***Chess***	
Models	SSIM	PSNR(dB)	SSIM	PSNR(dB)
[MUG18]	0.8426	27.31	0.8832	28.75
SM1	0.9023	29.85	0.9201	30.69
SM2	0.8135	25.82	0.8620	27.28
SM3	**0.9500**	**33.18**	**0.9619**	**34.49**

Table 1: Comparison of proposed sensing models; data sets used are ***Boxer*** and ***Chess*** with non-overlapping patches, each with 5 frames. We set $\beta = 3$, $n = 7 \times 7 \times 5 \times 5 \times 3 \times 3$, batchSize: 6000, and we used 10 frames for training (distinct from the testing set).

Figure 4: Comparison of proposed sensing models of the *Boxer* data set on a monochrome sensor with color-coded mask.

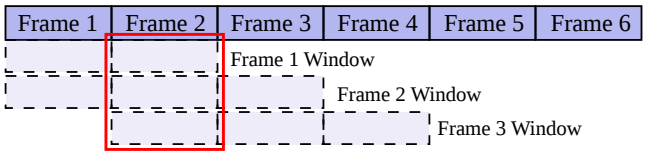


Figure 5: A window of size $\beta = 3$ for reconstruction is chosen so that the current frame is placed at the center of the window (except for corner cases). For each frame of the original light field video (Frame 2 in this example), we reconstruct three light field sequences (three rows shown with dashed lines). Therefore, we can combine three reconstructed frames (shown with a red box) to obtain a single frame corresponding to frame 2 in the original light field video.

atoms of the multi-frame dictionary contain temporal information, which improves the sparsity, and hence the reconstruction quality, compared to the single-frame dictionary, see Section 3.5.

3.5. Sparse Reconstruction of Light Field Videos

To reconstruct the measured light field video, we need to formulate a suitable reconstruction algorithm, according to (2), that takes into account the sensing matrices described in Section 3 and the multi-frame dictionary in Section 3.4. In what follows, we explain three possible sensing models for the reconstruction of light field video frames. The sensing models explained here are applicable to both camera designs of Section 3. For simplicity and without loss of generality, our explanation in this section considers the sensing matrix for monochrome sensors, as in (9). Moreover, we will occasionally refer to Figure 4 and Table 1 for quality comparison of different sensing models. Note that the main results and detailed comparisons are presented in Section 4.

3.5.1. Sensing Model 1 (SM1)

In this model, the sensed 2D RAW image of each frame, y^i are appended vertically into $y \in \mathbb{R}^{\beta\omega}$ and its corresponding sensing matrix Λ^i are stacked vertically into $\Phi_{(III)} \in \mathbb{R}^{\beta\omega \times \omega\nu\lambda}$ as explained in Section 3, and equation (10). The dictionary that we train for this sensing model is based on a single-frame dictionary learning method explained in Section 3.4, where the dictionary is $\mathbf{D} \in \mathbb{R}^{\omega\nu\lambda \times \rho\omega\nu\lambda}$. For β-consecutive incoming light fields $\{\mathbf{x}_1, \ldots, \mathbf{x}_\beta\}$, where $\mathbf{x}_i \in \mathbb{R}^{\omega\nu\lambda}$, the reconstruction is carried out by solving:

$$\underset{\theta}{\arg\min} \ \|\theta\|_0 \ \text{s.t.} \ \left\| \begin{bmatrix} \Lambda^1 \mathbf{x}^1 \\ \vdots \\ \Lambda^\beta \mathbf{x}^\beta \end{bmatrix} - \begin{bmatrix} \Lambda^1 \\ \vdots \\ \Lambda^\beta \end{bmatrix} \mathbf{D}\theta \right\|_2^2 \leq \varepsilon \tag{14}$$

As mentioned in Section 3, arranging the per-frame sensing matrices vertically will increase the number of incoherent samples, hence improving the reconstruction quality. It can be seen in Figure 4 that SM1 can reconstruct stationary objects in the background quite well, but due to lack of temporal information in the dictionary, it has many artifacts along the edges of moving objects where the foreground and the background meet.

3.5.2. Sensing Model 2 (SM2)

In this model we use a multi-frame dictionary where the patches span the time domain, i.e. there is a temporal coherency between the atoms in the dictionary. The sensing matrix is arranged horizontally as in (11) and lead to the following reconstruction problem:

$$\underset{\theta}{\arg\min} \ \|\theta\|_0 \ \text{s.t.} \ \left\| \begin{bmatrix} \Lambda^1 \ldots \Lambda^\beta \end{bmatrix} \begin{bmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^\beta \end{bmatrix} - \begin{bmatrix} \Lambda^1 \ldots \Lambda^\beta \end{bmatrix} \begin{bmatrix} \mathbf{D}_1 \\ \vdots \\ \mathbf{D}_\beta \end{bmatrix} \theta \right\|_2^2 \leq \varepsilon \tag{15}$$

Using a multi-frame dictionary trained on β frames with 6D patches, SM2 can recover each light field frame with significantly lower temporal artifacts as compared to SM1, as shown in Figure 4 and Table 1. However, arranging the sensing matrix horizontally will decrease the number of incoherent samples used in solving the BPDN problem (2). Even though the dictionary encodes multi-dimensional information, the minimization problem cannot find a suitable coefficient vector to reconstruct the signal accurately. As demonstrated in the figure, even the colors are not recovered accurately, and the resulting light field is very blurry.

3.5.3. Sensing Model 3 (SM3)

To maximize the incoherency of the measurements and at the same time the sparsity, we propose to use the multi-frame dictionary of **SM2** and the sensing matrix of **SM1**. In this way, we can have β times more incoherent samples compared to SM2 for the reconstruction algorithm while benefiting from the temporal correlations

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

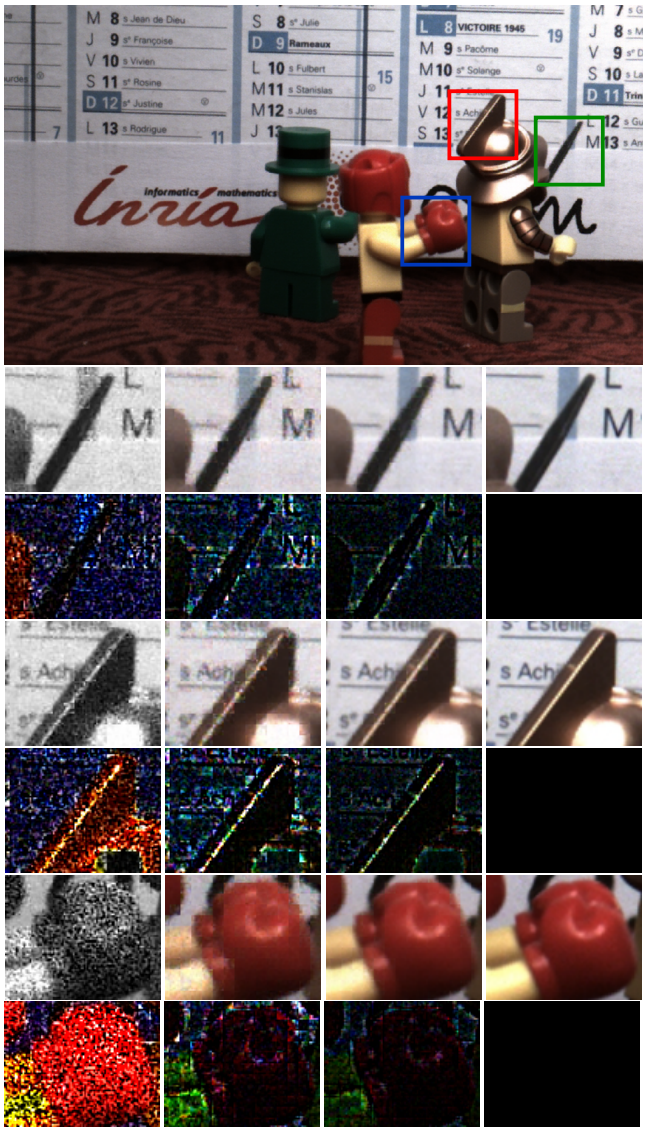
of the dictionary atoms. The efficiency of SM3 is confirmed in our results in Figure 4 and Table 1, as well as in Section 4. Since each frame of the light field video is captured individually using compressed sensing, we can re-arrange the matrix multiplication of the sensing matrix and the dictionary to obtain the following optimization problem for reconstruction

$$\underset{\theta}{\arg\min}\ \|\theta\|_0 \ \text{s.t.} \ \left\| \begin{bmatrix} \Lambda^1 \mathbf{x}^1 \\ \vdots \\ \Lambda^\beta \mathbf{x}^\beta \end{bmatrix} - \begin{bmatrix} \Lambda^1 \mathbf{D}_1 \\ \vdots \\ \Lambda^\beta \mathbf{D}_\beta \end{bmatrix} \theta \right\|_2^2 \leq \varepsilon, \tag{16}$$

where $\mathbf{D}_i \in \mathbb{R}^{\lambda\nu\omega \times \rho\beta\lambda\nu\omega}$, $i \in \{1, \ldots, \beta\}$, are sub-matrices of the multi-frame dictionary $\mathbf{D} \in \mathbb{R}^{\beta\lambda\nu\omega \times \rho\beta\lambda\nu\omega}$, defined in (13), corresponding to frame i of the captured light field. Using this sensing model will result in the recovered 4D light field $\hat{x} \in \mathbb{R}^{\beta\omega\nu\lambda}$, meaning that for each frame in the original light field video, we reconstruct β frames. As shown in Figure 5, we choose our temporal window for reconstruction such that the current frame is placed at the center of the window. Since SM3 reconstructs β frames for each frame in the original light field video, there is a possibility of combining the reconstructed frames to achieve higher quality. To this end, we use a simple average operation over the β reconstructed 4D light fields. We expect further improvement in reconstruction quality with a more sophisticated algorithm for combining the frames; for instance by considering the image structure and features present in light field views. The implementation of a more robust interpolation algorithm is left for future work.

The optimal value of β is dependent on the frame rate of the light field video and the amount of object movements of the scene between frames. In practice, for fast moving scenes, we set the value of β to a small value, e.g. $\beta = 3$ as used in our experiments in Section 4. For relatively stationary scenes, β can be set to a higher value. Since the choice of β is independent of the hardware design and only affects the reconstruction algorithm during post processing, one can choose different values for β for distinct portions of the light field sequence to achieve higher reconstruction quality. We have left this extension of our method for future work.

To solve the reconstruction problems in (14), (15), and (16), corresponding to the sensing models SM1, SM2, and SM3, we use the Smoothed-ℓ_0 (SL0) algorithm [MBZJ09]. Indeed any sparse recovery algorithm [WNF09, PRK93, NV10, NT09, KXAH15] can be used for this purpose. However, we found SL0 to have a better trade-off between reconstruction quality and speed.



[MWBR13] [MUG18] Ours (SM3) Reference

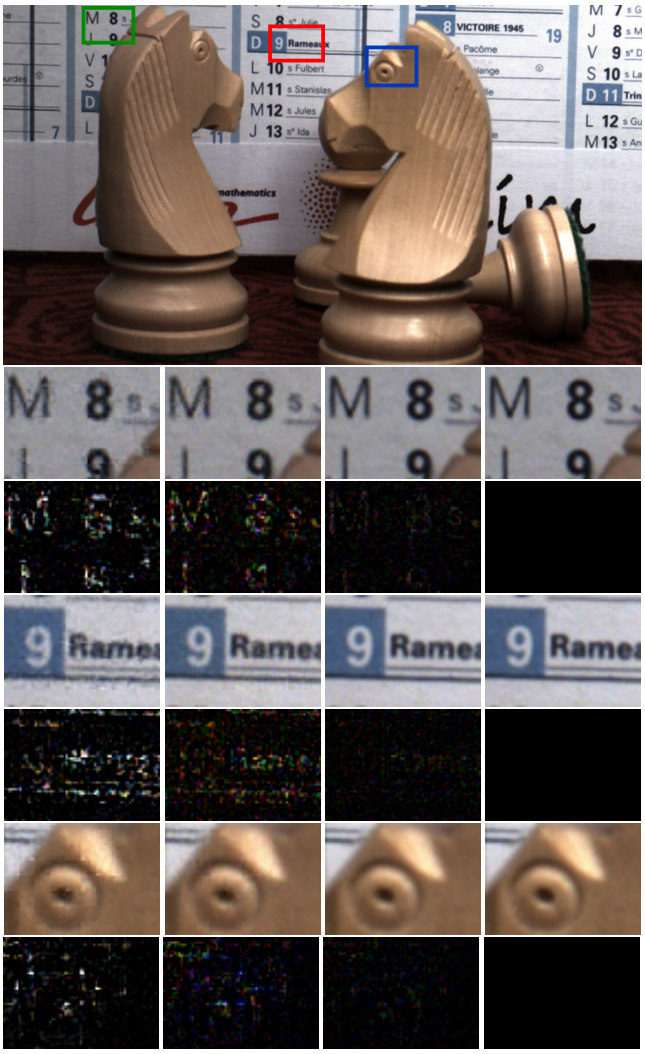
Figure 6: Reconstruction results using a monochrome sensor with a color-coded mask for the *Boxer-Gladiator-Irish* data set. The top image is the reconstruction with our method (SM3) including interpolation between the reconstructed frames as explained in Section 3.5. For quantitative results see Table 2. Error insets have a 5x intensity scaling to facilitate comparisons.

4. Results

We present our simulation results using the light field video data set of [GjLG18]. The data set consists of three light field sequences, where in two of them the camera is stationary and the objects are moving, *Boxer-Gladiator-Irish* and *Chess*, and one sequence where the objects are stationary and the camera moves around, which we call *Chess-moving*. The data sets are captured using a Raytrix R8 camera at a frame rate of 30 frames per second, where each frame consists of 5×5 light field views. For training the dictionary, we chose frames 490–500 from *Boxer-Gladiator-Irish* and frames 210–220 from *Chess*. Note that no frame from *Chess-moving* was included in the training set. The reconstruction for our method and all the methods we compare to was performed on frames 400–404 of *Boxer-Gladiator-Irish*, frames 15–19 of *Chess*, and frames 400–404 of *Chess-moving*.

The patch size for training and testing was set to $s \times t = 7 \times 7$ for the spatial domain, $u \times v = 5 \times 5$ for the angular domain, $\lambda = 3$ for the spectral domain, and $\beta = 3$ for the temporal domain. We placed the current frame in the center of the window to include backward and forward temporal movements. The size of the win-

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.



[MWBR13] [MUG18] Ours (SM3) Reference

Figure 7: Reconstruction results using a CFA-equipped sensor with a color-coded mask for the ***Chess*** data set. The top image is the reconstruction with our method (SM3) including the averaging of the reconstructed frames as explained in Section 3.5. For quantitative results see Table 2. Error insets have a 5x intensity scaling to facilitate comparisons.

dow can be adapted based on the movements in the scene. For light field sequences with rapid scene or camera movements, one should choose a smaller value for β, and vice versa. We found that $\beta = 3$ is sufficient for our data sets. The batch size for dictionary training was set to 6000 and we performed 40 iterations. Additionally, the training sparsity value was set to $\tau = 10$. We used the SPAMS library [MBPS10] to perform the training with OMP [PRK93] as the sparse coding method.

The camera is simulated with two sensor designs, as explained in Section 3, where a color-coded mask is placed at a distance from a CFA sensor or a monochrome sensor. The sensed RAW

Monochrome Sensor				
Data Set	*Boxer*		*Chess*	
Algorithm	SSIM	PSNR(dB)	SSIM	PSNR(dB)
[MWBR13]	0.4909	21.18	0.3954	19.87
[MUG18]	0.8426	27.31	0.8832	28.75
Ours (SM3)	**0.9500**	**33.07**	**0.9619**	**34.49**
CFA Sensor				
Data Set	*Boxer*		*Chess*	
Algorithm	SSIM	PSNR(dB)	SSIM	PSNR(dB)
[MWBR13]	0.9265	30.74	0.9443	32.15
[MUG18]	0.9504	34.29	0.9627	35.76
[IKT*18]	0.9608	33.08	0.9682	33.53
Ours (SM3)	**0.9824**	**40.29**	**0.9860**	**41.23**

Table 2: Reconstruction results for 5 frames of ***Boxer*** and ***Chess*** data sets using monochorome and CFA-equipped sensors with a color-coded mask. Non-overlapping patches of size $s \times t \times u \times v \times \lambda \times \beta = 7 \times 7 \times 5 \times 5 \times 3 \times 3$ were used.

Chess-moving with CFA Sensor				
Method	Ours (SM3)	[MWBR13]	[MUG18]	[IKT*18]
PSNR	**39.91dB**	35.27dB	38.14dB	37.53dB
SSIM	**0.9863**	0.9722	0.9817	0.9843
Chess-moving with Monochrome Sensor				
Method	Ours (SM3)	[MWBR13]	[MUG18]	
PSNR	**34.55dB**	19.27dB	31.25dB	
SSIM	**0.9693**	0.2421	0.9388	

Table 3: Reconstruction results for ***Chess-moving*** data set using monochorome and CFA-equipped sensors with a color-coded mask. We used non-overlapping patches of size $s \times t \times u \times v \times \lambda \times \beta = 7 \times 7 \times 5 \times 5 \times 3 \times 3$.

2D images for each setup is shown in Figure 3. For random entries in the sensing matrix, which are independent and identically distributed (i.i.d.), we use a Gaussian distribution with zero mean and a variance of one. A comparison of different distributions and their effect on the reconstruction quality is presented in [MUG18]. We tested our method on all three data sets, and the results reported here are an average over the PSNR and SSIM [WBSS04] for all reconstructed frames. We compared the result of our proposed sensing model SM3 with the previous state-of-the-art methods on compressive light field camera designs, in particular [MWBR13], [MUG18], and [IKT*18].

Table 1 represents the reconstruction result for *Boxer-Gladiator-Irish* and *Chess* data sets for both monochrome and CFA sensors. To have a fair comparison with [MWBR13], which uses a monochrome mask, we applied both our color sensing matrix (9), as well as a monochrome sensing matrix applied to each color channel, as described in [MWBR13]. For comparison with [MUG18], we used their proposed sensing matrix similar to (6) for CFA sensor and for monochrome sensor we used sensing matrix of (9). To compare with the deep learning method of Inagaki et al. [IKT*18], we trained their proposed network with a spatial patch size of 64×64 on the same training set that was used for training the dictionaries of our method, [MWBR13], and [MUG18]. We applied all three methods on each frame of the light field video individually to re-

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

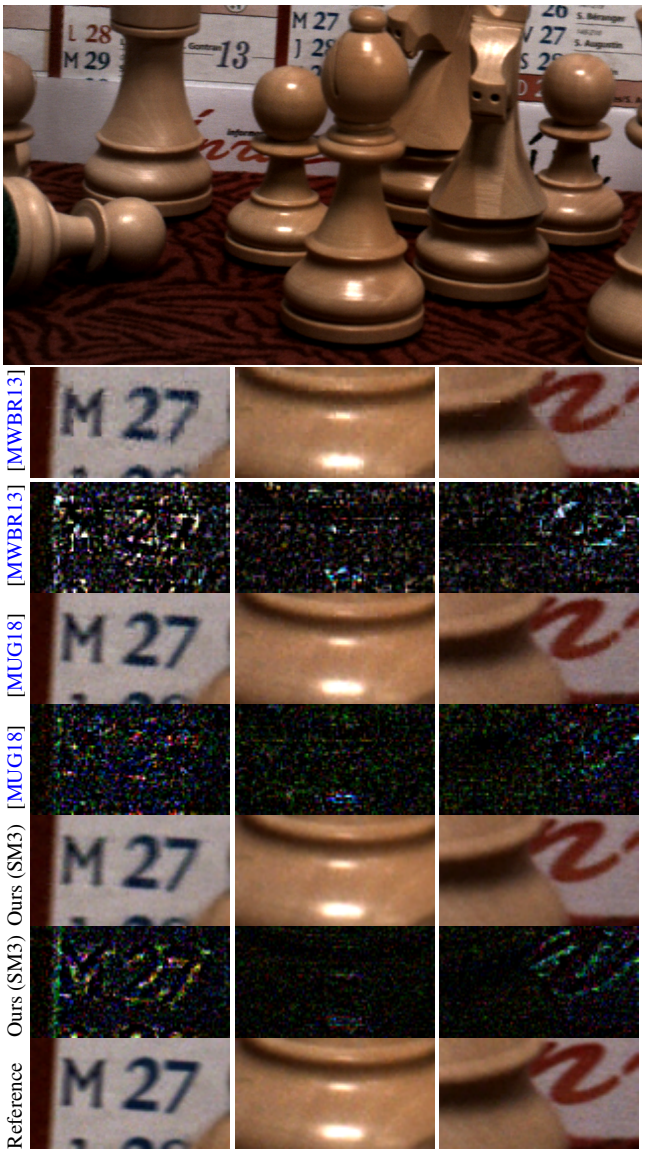


Figure 8: *Chess-moving* data set reconstructed with our proposed method (SM3) that utilizes a color-coded mask and a CFA-equipped sensor. For quantitative results see Table 3. Error insets have a 5x intensity scaling to facilitate comparisons.

construct the sequence. Note that the results reported here take into account the first and last frames of the video, where our method has fewer samples available for the reconstruction. Indeed, our results can be improved if the border frames are ignored, or if we pad the video with extra frames.

Figure 6 and Table 2 present the qualitative and quantitative results of our reconstruction from the RAW 2D image of a monochrome sensor in comparison to [MWBR13] and [MUG18]. Note the high accuracy in the reconstruction of details around the edges and high-frequency regions with reflections using our proposed method. It should be pointed out that the method of Inagaki et al. [IKT*18] does not support a monochrome sensor, hence it is not included in these results. The method of Marwah et al. [MWBR13] using a monochrome sensors with the color-coded mask cannot recover any color information as their proposed dictionary does not contain spectral information in its atoms. The method of Miandji et al. [MUG18] recovers signal reasonably well; however, the results are blurry and there exists severe color artifacts in high-frequency regions such as edges where the foreground and background meet. The PSNR of our method is on average 5.8dB higher than the state-of-the-art, a highly significant advantage. This is also confirmed with SSIM.

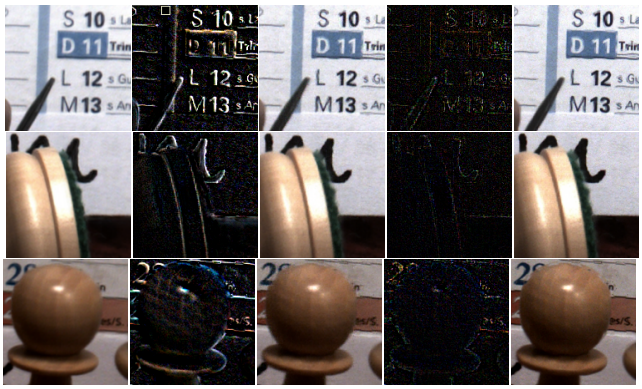


Figure 9: Visual comparison of our method using SM3 versus Inagaki et al. [IKT*18] for three data sets: *Boxer-Gladiator-Irish*, *Chess*, and *Chess-moving*; shown from top to bottom, respectively. For quantitative results see tables 2 and 3. Error insets have a 5x intensity scaling to facilitate comparisons.

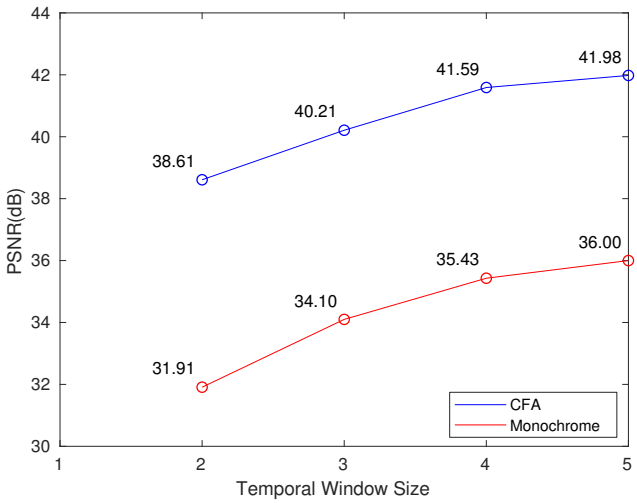


Figure 10: The effect of temporal window size, β, on the reconstruction quality of the *Boxer-Gladiator-Irish* data set for both designs using a monochrome sensor and a CFA-equipped sensor.

We also tested our proposed method using a CFA-equipped sensor, see Fig. 7 and Table 2. Our method shows sharper images without noise-like artifacts when compared to [MWBR13] and [MUG18]; see the supplementary video for temporal coherency of

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

the reconstructed light field videos for each algorithm. In this example, our method on average has 5.9dB higher PSNR than the state-of-the-art, showing the effectiveness of our method regardless of the sensor design. Although the reconstruction quality of the monochrome sensor is much lower than the CFA-equipped sensor, the compression ratio of the former design is much higher, which can be useful for e.g. fast transmission of the captured data. In particular, since the color components are convolved into a single scalar using a monochrome sensor, capturing a light field using this design leads to three times less samples than a CFA-equipped sensor design.

To evaluate the robustness of our algorithm to a fast moving scene or camera, we use the *Chess-moving* data set where the objects are stationary but the camera moves around. As a result, there are large pixel displacements on light field images, moving from one frame to the another. Table 3 summarizes the results of our reconstruction in comparison to the state-of-the-art, and Figure 8 compares the visual quality of the reconstructions. Even though this data set is very challenging, it can be seen that our method faithfully recovers the light field video. Moreover, for both monochrome and CFA sensors, our PSNR is about 2.0dB to 3.4dB higher than [MUG18]. See the the supplementary video for the advantages of our method with respect to temporal coherency of the reconstructed light field video in comparison to prior work.

Figure 9 illustrates the comparison of our method with the method of Inagaki et al. [IKT*18] on all three data sets: *Boxer-Gladiator-Irish*, *Chess*, and *Chess-moving*, as shown from top to bottom, respectively. As it can be seen in the figure, specially in the false-color error insets, the reconstruction results using [IKT*18] have blurring artifacts and pixel shifts around sharp edges, e.g. where foreground and background meet, as well as the areas in the background with text.

Figure 10 demonstrates the effect of the temporal window parameter, β, on the reconstruction quality for the *Boxer-Gladiator-Irish* data set. We changed the window size to include 2 to 5 consecutive frames in the reconstruction of the light field data set. As it can be seen, the reconstruction quality increases when more frames of the video are included. However, there is only a slight difference between the quality of reconstruction for $\beta = 4$ when compared to $\beta = 5$. Furthermore, the computational complexity increases when more frames are used in the reconstruction since the signal dimensionality increases. As a result, the temporal window size provides a trade-off between the reconstruction quality and the computational complexity. Note that since β is only used during the reconstruction, we can change the window size without modifying the camera design.

With regards to the computational complexity, on average, our algorithm takes 89 minutes to reconstruct a frame using SM3 in Eq. (16) when a monochrome sensor is used. For the same setup but with a CFA-equipped sensor, the reconstruction takes about 143 minutes. Note that since the resolution of the data sets we used is the same, the computation time for the full reconstruction of each data set is about the same. The timing results were obtained using a consumer-level desktop PC with a Ryzen 3600 CPU running at 4.0GHz.

[IKT*18] [MUG18] [MWBR13] Ours (SM3) Reference

Figure 11: Visual comparison of reconstructed *Animated Bunnies* data set using a CFA-equipped sensor.

Method	[IKT*18]	[MUG18]	[MWBR13]	Ours (SM3)
PSNR(dB)	22.95	25.04	23.93	**27.07**
SSIM	0.8100	0.7655	0.7536	**0.8444**

Table 4: Reconstruction results of the *Animated Bunnies* data set for CFA-equipped sensor.

5. Limitations and Future Work

One of the limitations of compressed sensing methods for mask-based light field photography is the requirement for a small baseline between the neighboring views. Indeed, this is not a limitation in practice since the hardware implementation of a light field camera using a coded mask does not admit a large baseline [MWBR13]. Regardless, to test the limits of our proposed reconstruction method, we also use a synthetic data set with a relatively large baseline, namely the *Animated Bunnies* data set [WLHR12]. The results are summarized in Figure 11 and Table 11. We see that our method significantly outperforms previous algorithms. However, comparing the PSNR of our method in Table 11 with those in e.g. Table 2, we see that the synthetic data set results in a much lower image quality. We also associate this with the pixel-wide sharp edges between the foreground and background, which does not happen for natural light fields.

Since we vectorize each 6D light field video patch, the size of the resulting vector is typically large, e.g. $n = 7 \times 7 \times 5 \times 5 \times 3 \times 3 = 11025$ for the light field videos we used here. If the dictionary is two times overcomplete, then $\mathbf{D} \in \mathbb{R}^{11025 \times 22050}$. Such a large dictionary negatively affects the computational complexity of the reconstruction algorithm. We propose two solutions to reduce the computation time, which are left for future work. First, optimized GPU implementations of the reconstruction algorithm, e.g. SL0 or similar techniques, can greatly reduce the reconstruction time. Up to 70x speedup has been reported for a variety of sparse recovery algorithms using a GPU implementation [BT13, FCWH11, BMU19]. Second, one can use a multidimensional dictionary, e.g. [MHU19], where an orthogonal dictionary is trained for each dimension of the light field. For instance, according to the example above, we will have two 7×7 dictionaries for the spatial domain, two 5×5 dictionaries for the angular domain, one 3×3 dictionary for the spectral, and a 3×3 dictionary for the temporal domains. This indeed greatly reduces the size of the dictionary, and hence the computational complexity.

Recovering a light field video can be challenging for scenes with extreme fast movement of the objects or the camera. One solution would be to estimate the disparity or flow information from the coded measurements formed on the sensor and use them in the reconstruction. Such information can also help us in deriving an effi-

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

cient method for combining the reconstructed frames to form a final frame of a light field video, as described in Section 3.5. Another direction for future work is to adaptively find the optimal number of consecutive frames, β, for faithful reconstruction. Since this parameter only affects the post-processing, i.e. the reconstruction, there is no need for changing the camera design based on β.

6. Conclusions

This paper presented a novel method for single-sensor compressive acquisition of light field video. A random color mask is placed in front of the sensor and moved randomly using a piezo motor prior to each frame capture. Given each captured 2D image and the corresponding mask, we formulate various sensing models to recover the full 6D light field video. We demonstrated that the use of temporal information in the dictionary training and the sensing model greatly improves the reconstruction quality with minimal temporal artifacts. Moreover, the proposed method was formulated for both monochrome and CFA-equipped sensors. We confirmed our findings by comparing our algorithm with the state-of-the-art methods and using various distinct data sets. Finally, since hardware implementation of mask-based light field photography has been successfully realized [MWBR13], and that we use the same input data as [MWBR13], we believe that our framework can be utilized in practice for efficient light field video cameras.

7. Acknowledgements

This work has been funded by Swedish Foundation for Strategic Research through grant IIS11- 0081 and in part by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

References

[AW92] ADELSON E. H., WANG J. Y. A.: Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 2 (1992), 99–106. doi:10.1109/34.121783. 2

[BAL*12] BABACAN S. D., ANSORGE R., LUESSI M., MATARAN P. R., MOLINA R., KATSAGGELOS A. K.: Compressive Light Field Sensing. *IEEE Transactions on Image Processing 21*, 12 (Dec 2012), 4746–4757. doi:10.1109/TIP.2012.2210237. 3

[BMU19] BARAVDISH G., MIANDJI E., UNGER J.: GPU Accelerated Sparse Representation of Light Fields. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - VISAPP* (2019), INSTICC, SciTePress, pp. 177–182. doi:10.5220/0007393101770182. 10

[BT13] BLANCHARD J. D., TANNER J.: GPU Accelerated Greedy Algorithms for Compressed Sensing. *Mathematical Programming Computation 5*, 3 (July 2013), 267–âĂŞ304. doi:10.1007/s12532-013-0056-5. 10

[CC17] CHEN J., CHAU L.-P.: Light field compressed sensing over a disparity-aware dictionary. *IEEE Trans. Cir. and Sys. for Video Technol. 27*, 4 (Apr. 2017), 855–865. doi:10.1109/TCSVT.2015.2513485. 3

[CRT06a] CANDES E. J., ROMBERG J., TAO T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory 52*, 2 (Feb 2006), 489–509. doi:10.1109/TIT.2005.862083. 2, 3

[CRT06b] CANDÈS E. J., ROMBERG J. K., TAO T.: Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics 59*, 8 (2006), 1207–1223. doi:10.1002/cpa.20124. 3

[DE03] DONOHO D. L., ELAD M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences 100*, 5 (2003), 2197–2202. doi:10.1073/pnas.0437847100. 3

[DH01] DONOHO D. L., HUO X.: Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory 47*, 7 (Nov 2001), 2845–2862. doi:10.1109/18.959265. 4

[Don06] DONOHO D. L.: Compressed sensing. *IEEE Trans. Inf. Theor. 52*, 4 (Apr. 2006), 1289–1306. doi:10.1109/TIT.2006.871582. 2, 3

[DTDS12] DONOHO D. L., TSAIG Y., DRORI I., STARCK J.: Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory 58*, 2 (Feb 2012), 1094–1121. doi:10.1109/TIT.2011.2173241. 3

[FCWH11] FANG Y., CHEN L., WU J., HUANG B.: GPU Implementation of Orthogonal Matching Pursuit for Compressive Sensing. In *IEEE 17th International Conference on Parallel and Distributed Systems* (Dec 2011), pp. 1044–1047. doi:10.1109/ICPADS.2011.158. 10

[GGSC96] GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.: The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1996), SIGGRAPH '96, ACM, pp. 43–54. doi:10.1145/237170.237200. 1, 4

[GJK*17] GUPTA M., JAUHARI A., KULKARNI K., JAYASURIYA S., MOLNAR A., TURAGA P.: Compressive light field reconstructions using deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (July 2017), pp. 1277–1286. doi:10.1109/CVPRW.2017.168. 3

[GjLG18] GUILLO L., JIANG X., LAFRUIT G., GUILLEMOT C.: Light field video dataset captured by a r8 raytrix camera, April 2018. 7

[GL10] GREGOR K., LECUN Y.: Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (USA, 2010), ICML'10, Omnipress, pp. 399–406. 3

[GN03] GRIBONVAL R., NIELSEN M.: Sparse representations in unions of bases. *Information Theory, IEEE Transactions on 49*, 12 (Dec. 2003), 3320–3325. doi:10.1109/TIT.2003.820031. 3

[GZC*06] GEORGEIV T., ZHENG K. C., CURLESS B., SALESIN D., NAYAR S., INTWALA C.: Spatio-angular resolution tradeoffs in integral photography. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques* (Aire-la-Ville, Switzerland, Switzerland, 2006), EGSR '06, Eurographics Association, pp. 263–272. doi:10.2312/EGWR/EGSR06/263-272. 2

[HGG*11] HITOMI Y., GU J., GUPTA M., MITSUNAGA T., NAYAR S. K.: Video from a single coded exposure photograph using a learned over-complete dictionary. In *2011 International Conference on Computer Vision* (Nov 2011), pp. 287–294. doi:10.1109/ICCV.2011.6126254. 3

[HSJ*14] HIRSCH M., SIVARAMAKRISHNAN S., JAYASURIYA S., WANG A., MOLNAR A., RASKAR R., WETZSTEIN G.: A switchable Light Field Camera Architecture with Angle Sensitive Pixels and Dictionary-based Sparse Coding. In *2014 IEEE International Conference on Computational Photography (ICCP)* (May 2014), pp. 1–10. doi:10.1109/ICCPHOT.2014.6831813. 2

[IKT*18] INAGAKI Y., KOBAYASHI Y., TAKAHASHI K., FUJII T., NAGAHARA H.: Learning to Capture Light Fields Through a Coded Aperture Camera. In *Computer Vision – ECCV 2018* (Cham, 2018), Ferrari V., Hebert M., Sminchisescu C., Weiss Y., (Eds.), Springer International Publishing, pp. 431–448. 3, 8, 9, 10

[KXAH15] KHAJEHNEJAD M. A., XU W., AVESTIMEHR A. S., HASSIBI B.: Improving the Thresholds of Sparse Recovery: An Analysis of a Two-Step Reweighted Basis Pursuit Algorithm. *IEEE Transactions on Information Theory 61*, 9 (Sept 2015), 5116–5128. doi:10.1109/TIT.2015.2448690. 7

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.

[LH96] LEVOY M., HANRAHAN P.: Light Field Rendering. In *Proceedings of SIGGRAPH 1996* (1996), ACM, pp. 31–42. doi:10.1145/237170.237199. 1, 2, 4

[LLW*08] LIANG C.-K., LIN T.-H., WONG B.-Y., LIU C., CHEN H. H.: Programmable Aperture Photography: Multiplexed Light Field Acquisition. *ACM Transactions on Graphics 27*, 3 (Aug 2008), 55:1–55:10. doi:10.1145/1360612.1360654. 3

[LSQQ13] LIU H., SONG B., QIN H., QIU Z.: An adaptive-admm algorithm with support and signal value detection for compressed sensing. *IEEE Signal Processing Letters 20*, 4 (April 2013), 315–318. doi:10.1109/LSP.2013.2245893. 3

[MBPS10] MAIRAL J., BACH F., PONCE J., SAPIRO G.: Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res. 11* (March 2010), 19–60. 5, 8

[MBZJ09] MOHIMANI H., BABAIE-ZADEH M., JUTTEN C.: A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm. *IEEE Transactions on Signal Processing 57*, 1 (2009), 289–301. doi:10.1109/TSP.2008.2007606. 7

[MEUA17] MIANDJI E., EMADI M., UNGER J., AFSHARI E.: On probability of support recovery for orthogonal matching pursuit using mutual coherence. *IEEE Signal Processing Letters 24*, 11 (Nov. 2017), 1646–1650. doi:10.1109/LSP.2017.2753939. 3

[MHU19] MIANDJI E., HAJISHARIF S., UNGER J.: A Unified Framework for Compression and Compressed Sensing of Light Fields and Light Field Videos. *ACM Trans. Graph. 38*, 3 (May 2019), 23:1–23:18. doi:10.1145/3269980. 2, 10

[Mia18] MIANDJI E.: *Sparse Representation of Visual Data for Compression and Compressed Sensing*. Linköping Studies in Science and Technology. Dissertations. Linköping University Electronic Press, 2018. 3, 5

[MUG18] MIANDJI E., UNGER J., GUILLEMOT C.: Multi-shot single sensor light field camera using a color coded mask. In *2018 26th European Signal Processing Conference (EUSIPCO)* (2018), pp. 226–230. doi:10.23919/EUSIPCO.2018.8553230. 2, 3, 4, 5, 6, 7, 8, 9, 10

[MW08] MARCIA R. F., WILLETT R. M.: Compressive coded aperture video reconstruction. In *2008 16th European Signal Processing Conference* (Aug 2008), pp. 1–5. 3

[MWBR13] MARWAH K., WETZSTEIN G., BANDO Y., RASKAR R.: Compressive Light Field Photography Using Overcomplete Dictionaries and Optimized Projections. *ACM Transactions on Graphics 32*, 4 (2013), 46:1–46:12. doi:10.1145/2461912.2461914. 2, 3, 4, 5, 7, 8, 9, 10, 11

[NLB*05] NG R., LEVOY M., BREDIF M., DUVAL G., HOROWITZ M., HANRAHAN P.: Light Field Photography with a Handheld Plenoptic Camera. *Technical Report CTSR 2005-02, Stanford University* (2005). 1, 2

[NMG18] NABATI O., MENDLOVIC D., GIRYES R.: Fast and accurate reconstruction of compressed color light field. In *2018 IEEE International Conference on Computational Photography (ICCP)* (May 2018), pp. 1–11. doi:10.1109/ICCPHOT.2018.8368477. 3, 5

[NT09] NEEDELL D., TROPP J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis 26*, 3 (2009), 301–321. doi:http://dx.doi.org/10.1016/j.acha.2008.07.002. 3, 7

[NV10] NEEDELL D., VERSHYNIN R.: Signal Recovery From Incomplete and Inaccurate Measurements Via Regularized Orthogonal Matching Pursuit. *Selected Topics in Signal Processing, IEEE Journal of 4*, 2 (April 2010), 310–316. doi:10.1109/JSTSP.2010.2042412. 7

[PRK93] PATI Y. C., REZAIIFAR R., KRISHNAPRASAD P. S.: Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers* (Nov 1993), pp. 40–44. doi:10.1109/ACSSC.1993.342465. 7, 8

[Ray02] RAY S.: *Applied Photographic Optics: Lenses and optical systems for photography, film, video, electronic and digital imaging*. 02 2002. doi:10.4324/9780080499253. 4

[SZ11] SALIGRAMA V., ZHAO M.: Thresholded Basis Pursuit: LP Algorithm for Order-Wise Optimal Support Recovery for Sparse and Approximately Sparse Signals From Noisy Random Measurements. *IEEE Transactions on Information Theory 57*, 3 (March 2011), 1567–1586. doi:10.1109/TIT.2011.2104512. 3

[Tro04] TROPP J.: Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on 50*, 10 (Oct. 2004), 2231–2242. doi:10.1109/TIT.2004.834793. 3

[UWH*03] UNGER J., WENGER A., HAWKINS T., GARDNER A., DEBEVEC P.: Capturing and Rendering with Incident Light Fields. In *Proceedings of the 14th Eurographics Workshop on Rendering* (2003), Eurographics Association, pp. 141–149. 2

[VCR*17] VADATHYA A. K., CHOLLETI S., RAMAJAYAM G., KANCHANA V., MITRA K.: Learning light field reconstruction from a single coded image. 3

[VRA*07] VEERARAGHAVAN A., RASKAR R., AGRAWAL A., MOHAN A., TUMBLIN J.: Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocusing. *ACM Transactions on Graphics 26*, 3 (Jul 2007). doi:10.1145/1276377.1276463. 2

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (Apr. 2004), 600–612. doi:10.1109/TIP.2003.819861. 8

[WGM09] WANG A., GILL P., MOLNAR A.: Light Field Image Sensors Based on the Talbot Effect. *Applied optics 48*, 31 (Nov 2009), 5897–5905. doi:10.1364/AO.48.005897. 2

[WGM11] WANG A., GILL P. R., MOLNAR A.: An Angle-sensitive CMOS Imager for Single-sensor 3D Photography. In *2011 IEEE International Solid-State Circuits Conference* (Feb 2011), pp. 412–414. doi:10.1109/ISSCC.2011.5746375. 2

[WJV*05] WILBURN B., JOSHI N., VAISH V., TALVALA E.-V., ANTUNEZ E., BARTH A., ADAMS A., HOROWITZ M., LEVOY M.: High performance imaging using large camera arrays. In *ACM SIGGRAPH 2005 Papers* (New York, NY, USA, 2005), SIGGRAPH '05, ACM, pp. 765–776. doi:10.1145/1186822.1073259. 1, 2

[WLD*06] WAKIN M. B., LASKA J. N., DUARTE M. F., BARON D., SARVOTHAM S., TAKHAR D., KELLY K. F., BARANIUK R. G.: Compressive imaging for video representation and coding. 3

[WLHR12] WETZSTEIN G., LANMAN D., HIRSCH M., RASKAR R.: Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4 (2012), 1–11. 10

[WNF09] WRIGHT S. J., NOWAK R. D., FIGUEIREDO M. A. T.: Sparse Reconstruction by Separable Approximation. *IEEE Transactions on Signal Processing 57*, 7 (July 2009), 2479–2493. doi:10.1109/TSP.2009.2016892. 7

[WZK*17] WANG T.-C., ZHU J.-Y., KALANTARI N. K., EFROS A. A., RAMAMOORTHI R.: Light field video capture using a learning-based hybrid imaging system. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017) 36*, 4 (2017). doi:10.1145/3072959.3073614. 3

[XL12] XU Z., LAM E. Y.: A High-resolution Lightfield Camera with Dual-mask Design. vol. 8500, pp. 1–11. doi:10.1117/12.940766. 3

[YZ11] YANG J., ZHANG Y.: Alternating Direction Algorithms for ℓ_1-Problems in Compressive Sensing. *SIAM Journal on Scientific Computing 33*, 1 (Feb 2011), 250–278. doi:10.1137/090777761. 3

© 2020 The Author(s)
Computer Graphics Forum © 2020 The Eurographics Association and John Wiley & Sons Ltd.