Optical imaging using binary sensors

Aurélien Bourquard,[1,*] **François Aguet,**[2] **and Michael Unser**[1]

[1]*Laboratoire d'imagerie biomédicale, École polytechnique fédérale de Lausanne, Switzerland*
[2]*Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA*
**aurelien.bourquard@epfl.ch*

Abstract: This paper addresses the problem of reconstructing an image from 1-bit-quantized measurements, considering a simple but non-conventional optical acquisition model. Following a compressed-sensing design, a known pseudo-random phase-shifting mask is introduced at the aperture of the optical system. The associated reconstruction algorithm is tailored to this mask. Our results demonstrate the feasibility of the whole approach for reconstructing grayscale images.

OCIS codes: (070.4560) Data processing by optical means; (100.2000) Digital image processing; (100.3010) Image reconstruction techniques; (100.3190) Inverse problems; (110.1758) Computational imaging; (110.4850) Optical transfer functions.

References and links

1. A. Stern, Y. Rivenson, and B. Javidi, "Optically compressed image sensing using random aperture coding," in "Proceedings of the SPIE - The International Society for Optical Engineering," (2008), pp. 69750D–1–10.
2. J. Romberg, "Sensing by random convolution," in "2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing," (2007), pp. 137–140.
3. M. F. Duarte, M. A. Davenport, D. Takbar, J. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling: Building simpler, smaller, and less-expensive digital cameras," IEEE Signal Process. Mag. **25**, 83–91 (2008).
4. R. F. Marcia and R. M. Willett, "Compressive coded aperture superresolution image reconstruction," in "IEEE International Conference on Acoustic, Speech and Signal Processes," (2008), pp. 833–836.
5. F. Sebert, Y. M. Zou, and L. Ying, "Toeplitz block matrices in compressed sensing and their applications in imaging," in "Proceedings of the 5th International Conference on Information Technology and Application in Biomedicine," (2008), pp. 47–50.
6. P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in "42nd Annual Conference on Information Sciences and Systems," (2008), pp. 16–21.
7. A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," SIAM Rev. **51**, 34–81 (2009).
8. L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D **60**, 259–268 (1992).
9. W. U. Bajwa, J. D. Haupt, G. M. Raz, S. J. Wright, and R. D. Nowak, "Toeplitz-structured compressed sensing matrices," in "IEEE Workshop on Statistical Signal Processing Proceedings," (2007), pp. 294–298.
10. H. Rauhut, "Circulant and toeplitz matrices in compressed sensing," in "Proceedings of SPARS'09," (2009).
11. W. Yin, S. Morgan, J. Yang, and Y. Zhang, "Practical compressive sensing with toeplitz and circulant matrices," Tech. rep., CAAM, Rice University (2010).
12. J. W. Goodman, *Introduction to Fourier Optics* (McGraw Hill Higher Education, 1996), 2nd ed.
13. M. Born and E. Wolf, *Principles of Optics* (Cambridge University Press, 1959), 7th ed.
14. M. Unser, "Splines: A perfect fit for signal and image processing," IEEE Signal Process. Mag. **16**, 22–38 (1999).
15. R. T. Rockafellar, *Convex Analysis (Princeton Mathematical Series)* (Princeton University Press, 1970).
16. S. P. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory **28**, 129–137 (1982).
17. J. Max, "Quantizing for minimum distortion," IRE Trans. Inf. Theory **IT-6**, 7–12 (1960).
18. L. Sbaiz, F. Yang, E. Charbon, S. Susstrunk, and M. Vetterli, "The gigavision camera," in "2009 IEEE International Conference on Acoustics, Speech and Signal Processing," (2009), pp. 1093–1096.

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

1. Introduction

The compression of information by physical means before its capture is called *compressed sensing* to suggest that the amount of data to be acquired is substantially reduced [1]. When the acquisition of data is expensive in terms of time or hardware, such an approach can be highly beneficial. This method is non-conventional, in the sense that a numerical reconstruction step has to be performed after acquisition, but recent work on compressed sensing [1–5] demonstrates the possibility of reconstructing conventional images from relatively few linear optical measurements. In compressed sensing, the measurement system (be it optical or purely theoretical) is assumed to be known, and is typically modeled in a linear-algebra framework where each observable quantity is taken as an appropriate linear combination of unknowns. The measurements correspond to real-valued samples, while the unknowns are to be recovered numerically.

Few-sample-reconstruction capabilities are attractive for usual imaging applications only when sensors are expensive. Conversely, inexpensive sensors can be plentiful but often introduce quantization aspects which, in our opinion, have a much greater practical relevance. For instance, Boufounos *et al.* [6] have investigated the extreme case of 1-bit compressed sensing, and demonstrated the advantages of quantization-specific reconstruction methods. Their work proposes a promising approach when dealing with quantization issues. However, the measurements they consider have not been explicitly associated with an optical device. To the best of our knowledge, no optical model has been specifically devised or merely evaluated for this 1-bit compressed-sensing problem.

In this paper, our first motivation is to investigate a binary-quantization paradigm with a simple optical model that leads to physically-realistic measurements in a diffraction-limited setting. Binary quantization is particularly appealing in hardware implementations where each sensor takes the form of a comparator to zero [6]. This corresponds to inexpensive and fast devices that can be made robust to saturation. Thus, the proposed acquisition method is potentially relevant in economic terms. It may also constitute a technological advantage in applications where the sensor response time is critical. From a conceptual point of view, our approach is to some extent the counterpart to Romberg's random-convolution framework [2] for 1-bit quantization. Indeed, we are pursuing the same overall goal, which is to reconstruct images based on a lesser amount of data. In our approach, we investigate the case where we strongly reduce the number of bits by considering 1-bit quantization, while compensating for an excessive loss of information by taking spatially denser samples into account. Our second interest is to devise an efficient reconstruction algorithm that exploits the simplicity of our forward model, and that is able to produce visual results on standard grayscale images.

Our general acquisition and reconstruction strategy is introduced in Sect. 2. In Sect. 3, we propose a binary optical acquisition device that uses a random-phase mask; we then derive a rigorous discretization of the model using B-spline basis functions. We specify the reconstruction problem in Sect. 4, and expose our optimized reconstruction algorithm in Sect. 5. We show visual results in Sect. 6.

2. Optical compressed-sensing approach

2.1. Theoretical aspects

Any linear measurement system corresponds to some matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, where M, N are the numbers of measurements and unknowns, respectively. Denoting $\mathbf{c}$ as the unknowns, the measurements $\mathbf{g}$ can be written as $\mathbf{g} = \mathbf{A}\mathbf{c}$. In this paper, the vectors (in bold lowercase) refer to lexicographically-ordered sequences, and each element of the M-vector $\mathbf{g}$ corresponds to one distinct measurement.

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

The theory of compressed sensing guarantees that $\mathbf{c} \in \mathbb{R}^N$ can be recovered from a small set of measurements if it is sufficiently *sparse* in some appropriate linear basis $\mathbf{\Phi}$. By sparsity, it is meant that $\mathbf{c}$ must consist of enough negligible entries once it has been represented in the basis $\mathbf{\Phi}$. As it turns out, natural images are often sparse in some transformed domains (e.g., wavelets). The importance of $\mathbf{\Phi}$, in the general theory, is its mere *existence*; its specific layout reflects the considered class of signals. The measurement matrix $\mathbf{A}$ bears no direct relation with $\mathbf{\Phi}$, except that, in order to be suitable, it must be *incoherent*—in the statistical sense—with that basis, which is almost surely the case if $\mathbf{A}$ contains independent and identically distributed (i.i.d.) random entries [7].

Once the measurements are obtained from the forward model, the reconstruction problem $\mathscr{P}$ is to find the sparsest solution $\mathbf{c}$ leading to these same measurements (up to some imprecision ε), given the system matrix $\mathbf{A}$. When $\mathbf{\Phi}$ is orthonormal, this can be expressed as

$$\mathscr{P} : \min_{\mathbf{c}} \left\| \mathbf{\Phi}^T \mathbf{c} \right\|_{\ell_0} \quad \text{subject to} \quad \|\mathbf{g} - \mathbf{A}\mathbf{c}\|_{\ell_2} \leq \varepsilon. \tag{1}$$

The left term maximizes the solution sparsity, while the right one ensures fidelity to the available measurements. In the general case, this problem is NP-hard (i.e., it cannot be solved in polynomial time). However, it has been shown that, when $\mathbf{c}$ is sparse, the sparsity measured in the ℓ_0-norm can be relaxed by using an ℓ_1-norm. In that case, one can equivalently solve the convex-optimization problem

$$\mathscr{P}' : \min_{\mathbf{c}} \|\mathbf{g} - \mathbf{A}\mathbf{c}\|_{\ell_2}^2 + \lambda \left\| \mathbf{\Phi}^T \mathbf{c} \right\|_{\ell_1}, \tag{2}$$

where $\lambda \in \mathbb{R}_+^*$ is a constant. This cost minimization is then tractable, and can be performed using standard optimization techniques.

Besides the aforementioned properties, it has been shown that compressed-sensing measurements are robust to quantization as well [6]. The corresponding problem can thus be treated as a variation of the classical one in which the measurements are quantized, and typically more numerous. In this work, we are going to deviate from the traditional compressed-sensing framework by first considering such quantized measurements, which requires the use of a modified data term in (2), and second by using a non-unitary regularization matrix $\mathbf{\Phi}$ that corresponds to total variation (TV), and which promotes piecewise-smooth solutions [8].

2.2. Overall strategy

In the literature, amenable optical implementations of compressed sensing and associated reconstruction approaches have been specifically devised and evaluated for problems involving few non-quantized measurements. In this paper, we propose another compressed-sensing-based imaging concept that uses 1-bit-quantized measurements. In terms of operations, our acquisition system can be split into two parts. The first part performs random linear measurements optically, and corresponds to a specific measurement matrix discussed below. The second part acquires a 1-bit-quantized version of these measurements using a binary-sensor array. Since the global acquisition system is strongly nonlinear due to its coarse-quantization stage, our reconstruction approach is adapted accordingly.

Measurement matrices consisting of i.i.d. random entries are incoherent with most orthonormal bases, which is a desirable property. Therefore, the corresponding optical models must be designed accordingly. Indeed, such matrices correspond to sequential-acquisition solutions, such as the single-pixel camera [3]. Meanwhile, it has been shown theoretically that some structured matrices that arise naturally in specific application areas also have the required incoherence property [2, 4, 5, 9–11]. In particular, the *random-convolution* matrices $\mathbf{A}_\chi$ that are associated to unit-amplitude and random-phase transfer functions are especially relevant in

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

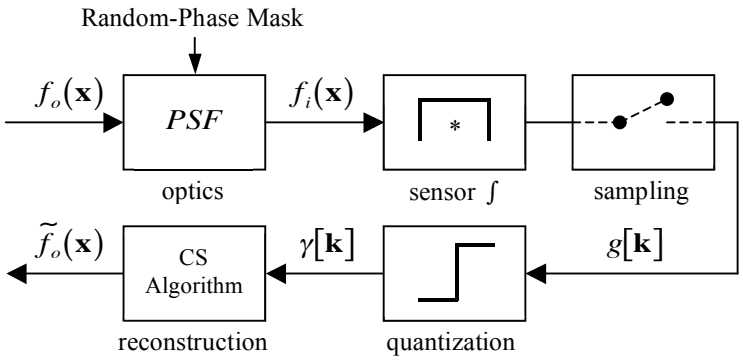

Fig. 1. Given an object $f_o(\mathbf{x})$ and a pseudo-random-phase mask, the optical system produces an (intermediate) image $f_i(\mathbf{x})$. A 1-bit sensor array samples and binarizes the latter to the measurements $\gamma[\mathbf{k}]$. The sampling process is uniform but non-ideal, meaning that each sensing area has some pre-integration effect. Note that the problem non-linearity (and hence complexity) is due to quantization. The binary and discrete data $\gamma[\mathbf{k}]$ are submitted to our compressed-sensing reconstruction algorithm, which is the last stage of our hybrid system. Assuming that the PSF is known, this algorithm reconstructs an image of the object. The result $\tilde{f}_o(\mathbf{x})$ can then be directly observed or evaluated.

the context of optical imaging [2]. This corresponds to simple and one-shot coherent-light optical-acquisition setups which are associated to pseudo-random and space-invariant point-spread functions (PSFs).

We choose to design the linear part of our acquisition model in the same spirit as in [2], with the important distinction that we are proposing a physically-realistic system that performs incoherent-light imaging. Since this approach is based on random convolution, it corresponds to some structured measurement matrix. Our reconstruction algorithm further exploits this specific convolution-type structure for fast large-scale reconstruction capabilities. Since operations that involve convolution matrices can be performed in the Fourier domain, they are fast and memory-efficient [9]. Thus, our sensing matrix not only suits the forward model, but also has a critical role to play in the whole reconstruction process. The general scheme is summarized in Fig. 1. Note that the object $f_o(\mathbf{x})$, the image $f_i(\mathbf{x})$, and the reconstruction $\tilde{f}_o(\mathbf{x})$ are continuously-defined, while $g[\mathbf{k}]$ and $\gamma[\mathbf{k}]$ are real-positive and binary sequences, respectively.

3. Forward model

3.1. Physical system

Our optical acquisition setup (Fig. 2) is centered (i.e., it is aligned with the optical axis) and follows a standard diffraction-limited model [13]. The object is planar and corresponds to the transmittance profile $f_o(\mathbf{x})$ with incoherent, parallel, and monochromatic illumination of intensity I and vacuum wavelength λ_0. In order to delocalize the PSF and to satisfy the compressed-sensing conditions, we insert a pseudo-random-phase mask in the exit pupil of the system. The medium refractive index n and the numerical aperture (NA) of the system are assumed to be known. A uniform CCD-like array of sensors is then exposed to the light intensity $f_i(\mathbf{x})$, which is sampled and binarized to $g[\mathbf{k}]$ and $\gamma[\mathbf{k}]$, respectively. The sensor pre-integration effect is modeled by the convolution filter $\phi(\mathbf{x})$.

Under the assumptions that the object is planar and that the NA is suitably low, the overall imaging process is isoplanatic (Fraunhofer regime); the PSF is thus constant over the sensor array. This can be modeled as the convolution

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

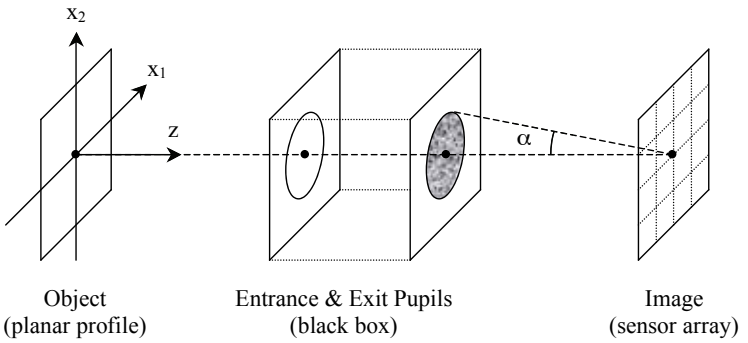

Fig. 2. The acquisition device is diffraction-limited; its circular exit pupil is phase-masked, which corresponds to a generalized pupil function. Geometric light propagation between the entrance and exit pupils [12] and unit magnification are assumed. Thus, the same coordinate system holds for both object $f_o(\mathbf{x})$ and image $f_i(\mathbf{x})$. All device components are centered and placed perpendicularly to the optical axis. The numerical aperture is then defined as $\mathrm{NA} = n\sin\alpha$.

$$f_i(\mathbf{x}) = (f_o * h)(\mathbf{x}), \tag{3}$$

where $*$ denotes a continuous-domain convolution, and where $h(\mathbf{x})$ is the corresponding space-invariant filter. Using normalized coordinates $\boldsymbol{\xi}$ for simplicity, the system aperture is defined as

$$\mathrm{circ}(\boldsymbol{\xi}) = \begin{cases} 1, & \|\boldsymbol{\xi}\| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The specificity of our system is that we introduce a piecewise-constant phase mask $p(\boldsymbol{\xi})$ at the aperture, with

$$p(\boldsymbol{\xi}) = \sum_{\mathbf{k}\in\mathbb{Z}^2} v[\mathbf{k}]\mathrm{rect}\left(\frac{\rho}{2}\boldsymbol{\xi} - \mathbf{k}\right). \tag{5}$$

There, $\mathrm{rect}(\cdot)$ denotes the two-dimensional rectangle function, and the values $v[\mathbf{k}]$ are independent and uniformly-distributed random variables from the pair $\{0, \pi\}$. The number of mask elements along the exit-pupil diameter is ρ. Given (4) and (5), we can define a *generalized pupil function* $q(\boldsymbol{\xi})$ that includes phase-distortion effects, including defocus [12, 13]. It is given by

$$q(\boldsymbol{\xi}) = \mathrm{circ}(\boldsymbol{\xi})\exp\left[-\mathrm{j}\left(k_0 W_{20}\|\boldsymbol{\xi}\|^2 + p(\boldsymbol{\xi})\right)\right], \tag{6}$$

where k_0 is the vacuum angular wavenumber. When the image is in focus, one can neglect the quadratic-phase coefficient W_{20}. In incoherent imaging, the intensity impulse response $h(\mathbf{x})$ corresponds to the inverse Fourier transform of the autocorrelation of the (generalized) pupil function [12]. It is expressed as

$$h(\mathbf{x}) = \mathscr{K}_h \left| \mathscr{F}\{q(\boldsymbol{\xi})\}\left(\frac{\mathrm{NA}}{\lambda_0}\mathbf{x}\right)\right|^2, \tag{7}$$

where $\mathscr{K}_h$ is a constant that ensures the energy of the light is conserved, and where $\mathscr{F}$ denotes the continuous Fourier transform that is defined as

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

$$\mathscr{F}\{f\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} f(\mathbf{x}) \exp\left\{-\mathrm{j}2\pi\boldsymbol{\omega}^T\mathbf{x}\right\} \mathrm{d}\mathbf{x}. \tag{8}$$

In order to produce a convolution kernel that is larger than the object size S with the given phase mask, the system is configured such that the frequency support of $h(\mathbf{x})$ is equal to $S^{-1}\rho$. The pre-filter $\phi(\mathbf{x})$, which acts before sampling, is given by

$$\phi(\mathbf{x}) = s\left(\frac{\mathbf{x}}{\Delta_s}\right), \tag{9}$$

where $s(\mathbf{x})$ is the sensor-scaled integration function, and where Δ_s is the corresponding sensor size. Finally, the non-quantized measurements are obtained by convolving the image $f_i(\mathbf{x})$ with the pre-filter $\phi(\mathbf{x})$ and sampling

$$g[\mathbf{k}] = (f_i * \phi)(\mathbf{x})|_{\mathbf{x}=\mathbf{k}\Delta_s}. \tag{10}$$

The measured sequence $g[\mathbf{k}]$ is finally binarized at the sensor level to the signs

$$\gamma[\mathbf{k}] = \mathscr{B}(g[\mathbf{k}], \tau) = \begin{cases} +1, & g[\mathbf{k}] \geq \tau \\ -1, & g[\mathbf{k}] < \tau, \end{cases} \tag{11}$$

where τ is an appropriate hardware-threshold value.

3.2. Exact discretization using B-splines

Besides measurements, which are necessarily discrete, the compressed-sensing formalism considers discretely-defined unknowns as well. In order to discretize them while keeping an underlying continuous-domain representation, we consider a B-spline expansion [14] for the object

$$f_o(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^2} c[\mathbf{k}]\varphi\left(\frac{\mathbf{x}}{\Delta_c} - \mathbf{k}\right), \tag{12}$$

where $c[\mathbf{k}]$ are real coefficients, where $\varphi(\mathbf{x}) = \varphi(x_1)\varphi(x_2)$ is the two-dimensional separable B-spline of degree m, and where Δ_c is the regular grid spacing.

We assume that the field of view includes N object coefficients and M sensors. The number of elements are the same along each dimension, so that N and M are square positive integers related by $N\Delta_c^2 = M\Delta_s^2$.

We now need to find the linear relation between the object coefficients $c[\mathbf{k}]$ and the non-quantized measurements $g[\mathbf{k}]$, according to the compressed-sensing formulation of Sect. 2. In vector notation, we find that

$$\mathbf{g} = \mathbf{D}_{(\mathscr{N})}\mathbf{A}_\chi\mathbf{U}_{(\mathscr{M})}\mathbf{c} = \mathbf{A}\mathbf{c}, \tag{13}$$

where $\mathbf{D}_{(\mathscr{N})}$ and $\mathbf{U}_{(\mathscr{M})}$ are downsampling-by-$\mathscr{N}$ and upsampling-by-$\mathscr{M}$ matrices. The integers $\mathscr{M}$ and $\mathscr{N}$ are such that the right-hand side of the equality $M/N = \mathscr{M}^2/\mathscr{N}^2$ is in reduced form. The matrix $\mathbf{A}_\chi$ is the convolution matrix that is associated with the filter $\chi[\mathbf{k}]$, with

$$\chi[\mathbf{k}] = \frac{I\Delta_s^2}{\mathscr{N}^2}\left((h*\phi)\left(\frac{\Delta_s\cdot}{\mathscr{N}}\right) * \varphi\left(\frac{\cdot}{\mathscr{M}}\right)\right)(\mathbf{x})|_{\mathbf{x}=\mathbf{k}}. \tag{14}$$

Although the underlying system matrix is not part of Romberg's family of random convolution matrices, the measurement matrix $\mathbf{A}$ nevertheless inherits from the randomness of the phase mask. Indeed, while the amplitude response of $\mathbf{A}_\chi$ is not ideal in the sense of [2], we have verified that its phase response is still essentially random and adequate for the purpose of our

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

imaging system. The important points are that the measurement matrix exactly represents our physical setup, and that it yields the correct discretization of the problem. The sequence $\chi[\mathbf{k}]$ can indeed be pre-determined with arbitrary precision.

4. Reconstruction problem

Given $\mathbf{A}$ and $\gamma[\mathbf{k}]$, we wish to reconstruct an image $\tilde{f}_o(\mathbf{x})$ that best approximates the object $f_o(\mathbf{x})$, up to scale and shift because information on the light intensity is lost after quantization. Similar to [6], we formulate our reconstruction problem in a variational framework. The specificity of our approach is the choice of the cost function to minimize, which is related to the design of our optimization algorithm.

The object that we consider is continuous, but the reconstruction problem can be formulated exactly using its discrete coefficients $c[\mathbf{k}]$. The solution $\tilde{c}[\mathbf{k}]$ is expressed as

$$\tilde{c} = \arg\min_{c} \underbrace{\mathscr{D}(c) + \lambda \mathscr{R}(c)}_{\mathscr{C}(c)}. \tag{15}$$

In this framework, the first term $\mathscr{D}(c)$ imposes solution fidelity to the binary measurements. The solution being always under-determined due to data quantization, the second term $\mathscr{R}(c)$ weighted by λ regularizes it according to the object sparse representation. The total cost value is denoted by $\mathscr{C}(c)$.

4.1. *Data fidelity*

The constraint of data fidelity demands that the reintroduction of $\tilde{f}_o(\mathbf{x})$ in place of $f_o(\mathbf{x})$ into the system of Fig. 1 results in the same set of binary samples $\gamma[\mathbf{k}]$. To impose this constraint in a gradual fashion, we introduce a new convex functional for sign consistency. Using the potential $\psi(t)$ as a penalty function, we write $\mathscr{D}(c)$ as

$$\mathscr{D}(c) = \sum_{\mathbf{k}} \psi\left(g[\mathbf{k}]\gamma[\mathbf{k}]\right), \tag{16}$$

where $\psi(t)$ is defined as

$$\psi(t) = \frac{\pi}{2M} - \begin{cases} t, & t < 0 \\ M^{-1}\arctan(Mt), & \text{otherwise.} \end{cases} \tag{17}$$

Given the form of $\gamma[\mathbf{k}]$, negative arguments of $\psi(t)$ correspond to sign inconsistencies, and vice versa. The first key feature of our penalty function $\psi(t)$ is its *linearity* for sign inconsistencies (Fig. 3). We found this property to be favorable for reducing error concentration at sign transitions, which greatly improves reconstruction sharpness. In addition, a small arctan-type penalty is active when the sign is correct. This feature ensures that the solution norm is nonzero and finite without abandoning convexity; we propose it as an alternative to non-convex normalization constraints. In our case, the convexity of $\mathscr{D}(c)$ can be easily deduced [15] given the continuity of $\psi(t)$.

4.2. *Regularization*

In order to regularize the solution, we propose to use a TV functional. Indeed, by minimizing the L_1-norm of the gradient, TV regularization is known to yield sharp edges in reconstructions [8], which is appropriate for visual data. Along with the above data term, this choice implies that the cost is convex. The functional $\mathrm{TV}(f_o)$ can be approximated by the sum

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

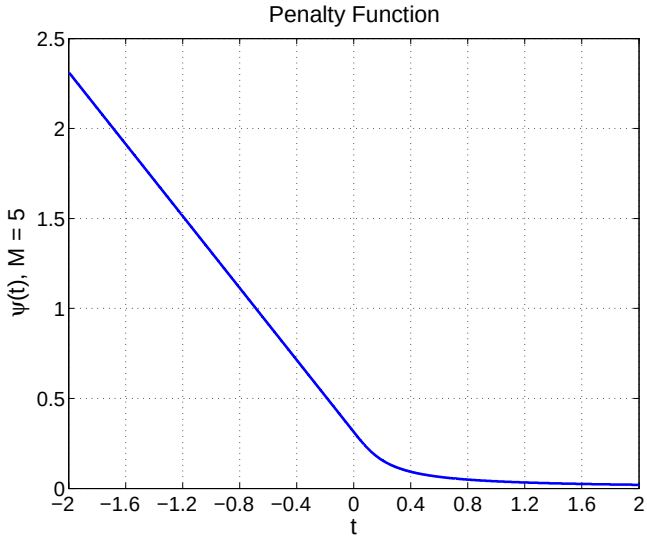

Fig. 3. Shape of our penalty function. The transition between the linear and arctan regimes of $\psi(t)$ is C^2-continuous, and takes place at $t = 0$. When t goes to infinity, the applied penalty tends to zero. The convexity of $\psi(t)$ clearly appears in this graph.

$$\mathscr{R}(c) = \sum_{\mathbf{k}} \theta_c[\mathbf{k}], \tag{18}$$

where $\theta_c[\mathbf{k}]$ is a sequence of gradient norms that depend on the coefficients $c[\mathbf{k}]$. Using $\star$ to denote a discrete-domain convolution, $\theta_c[\mathbf{k}]$ is determined from $c[\mathbf{k}]$ as

$$\theta_c[\mathbf{k}] = \left((c \star \varphi_{x_1})[\mathbf{k}]^2 + (c \star \varphi_{x_2})[\mathbf{k}]^2 + \upsilon^2 \right)^{\frac{1}{2}}, \tag{19}$$

where $\varphi_{x_{1,2}}[\mathbf{k}]$ are directional B-spline filters with staggered (i.e., half-shifted) first derivatives, and where υ is a small constant which circumvents TV non-differentiability at the origin. We have

$$\varphi_{x_1}[\mathbf{k}] = \varphi'(k_1 + 1/2)\varphi(k_2), \tag{20}$$
$$\varphi_{x_2}[\mathbf{k}] = \varphi'(k_2 + 1/2)\varphi(k_1), \tag{21}$$

where $\varphi'(x)$ is the B-spline first derivative whose symbolic expression can be found in [14].

5. Reconstruction algorithm

We have developed a preconditioned gradient-descent algorithm for the iterative optimization of (15). It is guaranteed to converge since the cost functional is convex. As we are going to show in Sect. 6, the role of the preconditioning operator $\mathbf{P}$ is essential, for it allows the solution to be approached in reasonable time.

Gradient descents are parameterized with an initial step size Ω and a relaxation parameter $\mu < 1$; the parameter $\mathscr{I}$ specifies the total number of iterations. The role of μ is to ensure that Ω is suitably small (i.e., that it decreases the cost function at each iteration). Starting from an initial guess $\tilde{c}^{(0)}[\mathbf{k}]$, and denoting the preconditioned cost gradient by $\nabla_{\mathbf{P}}\mathscr{C}$, the solution is found according to the five-step scheme expressed in vector notation below:

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

1. Default initialization: $\tilde{\mathbf{c}}^{(0)} \leftarrow \left\|\mathbf{P}\mathbf{A}^T\boldsymbol{\gamma}\right\|^{-1}\mathbf{P}\mathbf{A}^T\boldsymbol{\gamma}$ and $i \leftarrow 0$
2. Counter increase: $i \leftarrow i+1$
3. If $i > 1$ and $\mathscr{C}(\tilde{\mathbf{c}}^{(i-1)}) > \mathscr{C}(\tilde{\mathbf{c}}^{(i-2)})$, $\Omega \leftarrow \mu\Omega$
4. Gradient descent: $\tilde{\mathbf{c}}^{(i)} \leftarrow \tilde{\mathbf{c}}^{(i-1)} - \Omega\nabla_{\mathbf{P}}\mathscr{C}(\tilde{\mathbf{c}}^{(i-1)})$
5. If $i < \mathscr{I}$, return to step 2; otherwise terminate.

The preconditioned gradient of the cost function is given by the matrix-form expression

$$\nabla_{\mathbf{P}}\mathscr{C}(\cdot) = \mathbf{P}\left(\mathbf{A}^T\boldsymbol{\Gamma}\boldsymbol{\psi}'(\boldsymbol{\Gamma}\mathbf{A}\cdot) + \lambda(\mathbf{R}_1^T\boldsymbol{\theta}\mathbf{R}_1 + \mathbf{R}_2^T\boldsymbol{\theta}\mathbf{R}_2)\cdot\right), \tag{22}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix whose terms correspond to the signs $\gamma[\mathbf{k}]$, where $\mathbf{R}_{1,2}$ are convolution matrices corresponding to discrete convolution with the filters $\varphi_{x_{1,2}}[\mathbf{k}]$, respectively, and where $\boldsymbol{\theta}$ is a diagonal matrix containing the terms $\theta_c[\mathbf{k}]^{-1}$. The vector function $\boldsymbol{\psi}'(\mathbf{t})$ is defined as

$$\psi_i'(\mathbf{t}) = -\begin{cases} 1, & t_i < 0 \\ \left(1+M^2t_i^2\right)^{-1}, & \text{otherwise.} \end{cases} \tag{23}$$

Note that the above matrices all correspond to basic operations such as filtering or point-wise multiplication. Exploiting our problem structure, we specify the preconditioning operator $\mathbf{P}$ as the positive-definite convolution matrix given by

$$\mathbf{P} = \mathscr{K}_P\left(\mathbf{D}_{(\mathscr{M})}\mathbf{A}_\chi^T\mathbf{A}_\chi\mathbf{U}_{(\mathscr{M})} + \sigma\mathbf{R}\right)^{-1}, \tag{24}$$

which lends itself to an implementation in the Fourier domain. The matrix $\mathbf{R} = \mathscr{M}^2(\mathbf{R}_1^T\mathbf{R}_1 + \mathbf{R}_2^T\mathbf{R}_2)$ regularizes the inverse of (24), and is weighted by $\sigma \in \mathbb{R}_+^*$. The preconditioning operator is scaled by the constant $\mathscr{K}_P$. The role of $\mathbf{P}$ in the gradient expression of (22) is to compensate the amplitude-filtering effects of $\mathbf{A}$; its essence is to yield a pre-inversion of the forward operator.

6. Results and discussion

In this section, we report, evaluate, and discuss some experiments on grayscale images. The reconstruction algorithms were implemented in MATLAB. The object was assumed to be periodic, which is consistent with the use of an FFT-based algorithm to implement convolution, in particular, the preconditioning with the matrix $\mathbf{P}$ in (24).

All experiments are parameterized with $m = 1$ (linear B-splines), $\Omega = 0.05$, $\mu = 0.99$, $\lambda = 2\cdot10^{-4}$, $\sigma = 10^{-5}$, and $\upsilon = 10^{-5}$. The sensor-scaled integration function $s(\mathbf{x})$ is defined as a two-dimensional and separable rectangular window. The threshold τ is set such that the binary values $\gamma[\mathbf{k}]$ are equiprobable. The reference images as well as the reconstructions are defined on a Cartesian grid with $N = 256^2$ pixels. In order to provide a meaningful quality assessment, we matched the mean and variance of the solution coefficients to the reference signal.

We first want to compare the performance of our algorithm with its non-preconditioned counterpart. To do so, we consider the reconstruction of the *Bird* image. This image, as well as the other used in our experiments, are part of a standard set of test images. The parameters are $M = 256^2$, $\rho = 256$, $\mathscr{I} = 5\cdot10^3$, and the solution is initialized to zero. The evolution of the reconstruction quality (in terms of SNR) as a function of the number of iterations is shown in Fig. 4. Without preconditioning, the convergence turns out to be very slow; the optimum is actually reached after more than 10^6 iterations. In the preconditioned case, the algorithm converges

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

a thousand times faster, in about 10^3 iterations. From these results, we conclude that our preconditioning approach accelerates the reconstruction process by several orders of magnitude. It is thus a key element of our system that ensures convergence to the solution in reasonable time.

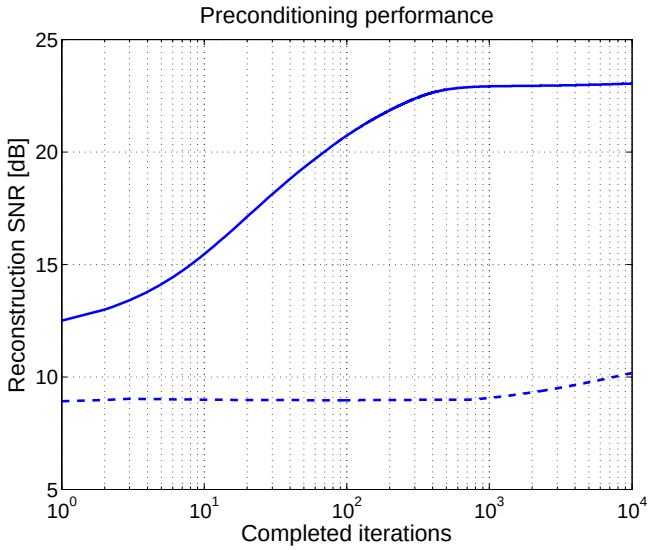

Fig. 4. Comparison between standard and preconditioned gradient descent for our algorithm. The reconstruction SNR is shown as a function of the number of iterations (logarithmic scale). In the preconditioned case, the corresponding SNR (solid line) reaches a plateau after relatively few iterations. Comparatively, the same result without preconditioning (dashed line) shows an extremely slow SNR progression.

Next, we provide some visual illustrations of the method. The parameters for the first experiment (*House* image) are $M = 256^2$, $\rho = 256$, and $\mathscr{I} = 10^3$. For the second experiment (*Peppers* image), we choose $M = 512^2$, $\rho = 512$, and $\mathscr{I} = 10^3$. Our approach is compared with conventional imaging in ideal conditions (i.e., ideal sampling and no phase mask), where optical acquisition alone is performed. Both modalities extract the same number of bits—in form of binary measurements—from the data. In the conventional case, the binary threshold is optimized with respect to the mean-squared quantization error. The solution is provided by the Lloyd-Max (LM) algorithm [16, 17].

Results of these experiments are shown in Figs. 5 and 6. Unlike the standard approach, our results recover substantial grayscale information; this also corroborates the reported SNR improvements. Our reconstructions illustrate the robustness of compressed-sensing measurements in the case of 1-bit quantization.

Comparing the two reconstructions, we notice that the quality (in terms of SNR) is highest in the first one, although fewer measurements are performed in that case. This striking difference can be explained by the strong piecewise-constant character of the acquired *House* image. Indeed, this object best suits our TV prior model (i.e., its gradient is negligible in most spatial locations, which is related to sparsity in some sense), which yields higher-quality reconstructions in similar acquisition conditions.

In Table 1, we have reported the reconstruction quality of different images, using several resolutions (i.e., measurement densities on the same array) for the acquisition. The fixed parameters are $\rho = 256$ and $\mathscr{I} = 500$. The results show that our approach remains advantageous at lower resolutions. When increasing M beyond the number of unknowns, the reconstruction

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

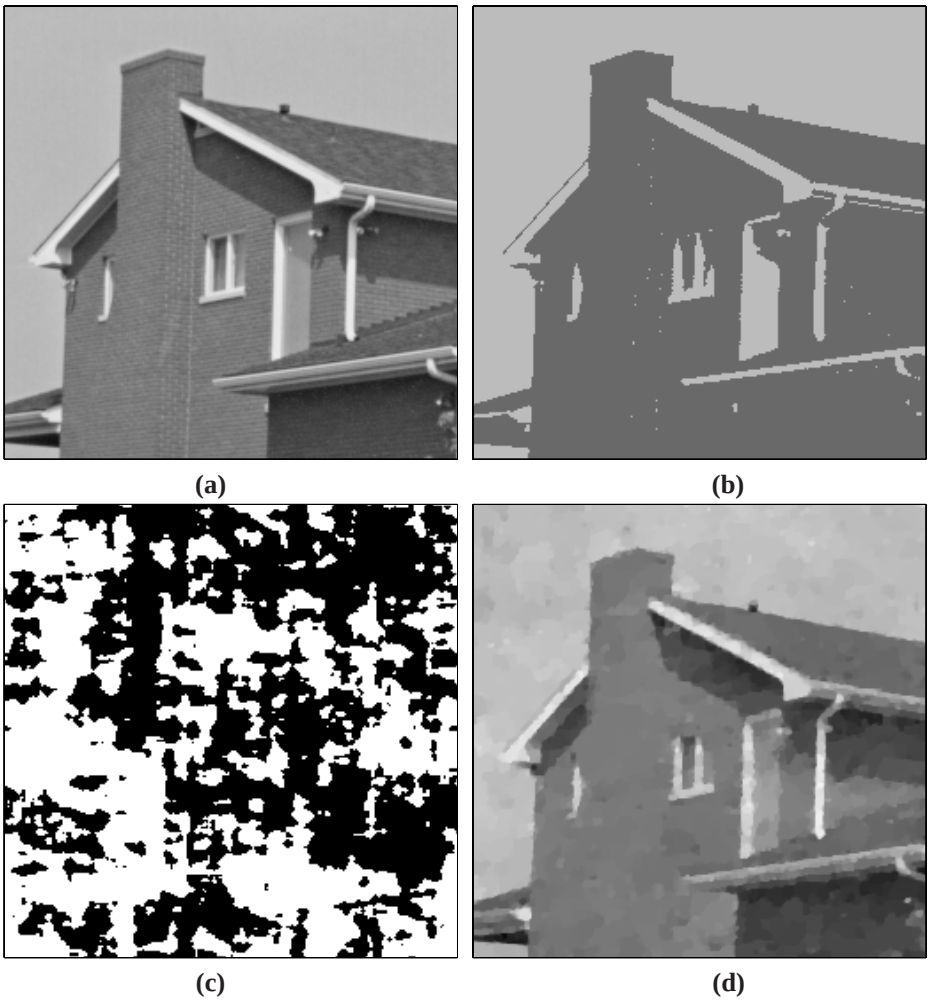

Fig. 5. *House* image. (a) Original object (b) Conventional solution with minimum-error threshold (Lloyd-Max): SNR = 16.59 dB (c) Binary acquisition using our method (d) Reconstruction using our method: SNR = 23.21 dB

quality further improves in all cases. These numerical experiments also suggest that the reconstruction performance is dependent upon the type of data. Indeed, the images that are best reconstructed have lower TV norms (e.g., *Bird* image). These findings are in line with the theoretical predictions.

Visually, the compressed-sensing acquisitions that are shown have substantially less spatial redundancy (i.e., more zero-crossing patterns) than their counterparts obtained through Lloyd-Max. Given the one-to-one mapping between bits and measurements, we thus infer that our system is able to take advantage of most of the information content that can be recovered from the acquisitions, which potentially leads to higher-quality reconstructions. This further confirms that our system is a compressed-sensing-based device. Note that, in this limit binary-acquisition scenario, the conceptual links between information content, redundancy, and compressed sensing become more intuitive.

These results validate the 1-bit imaging concept that we propose. However, aspects that are

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

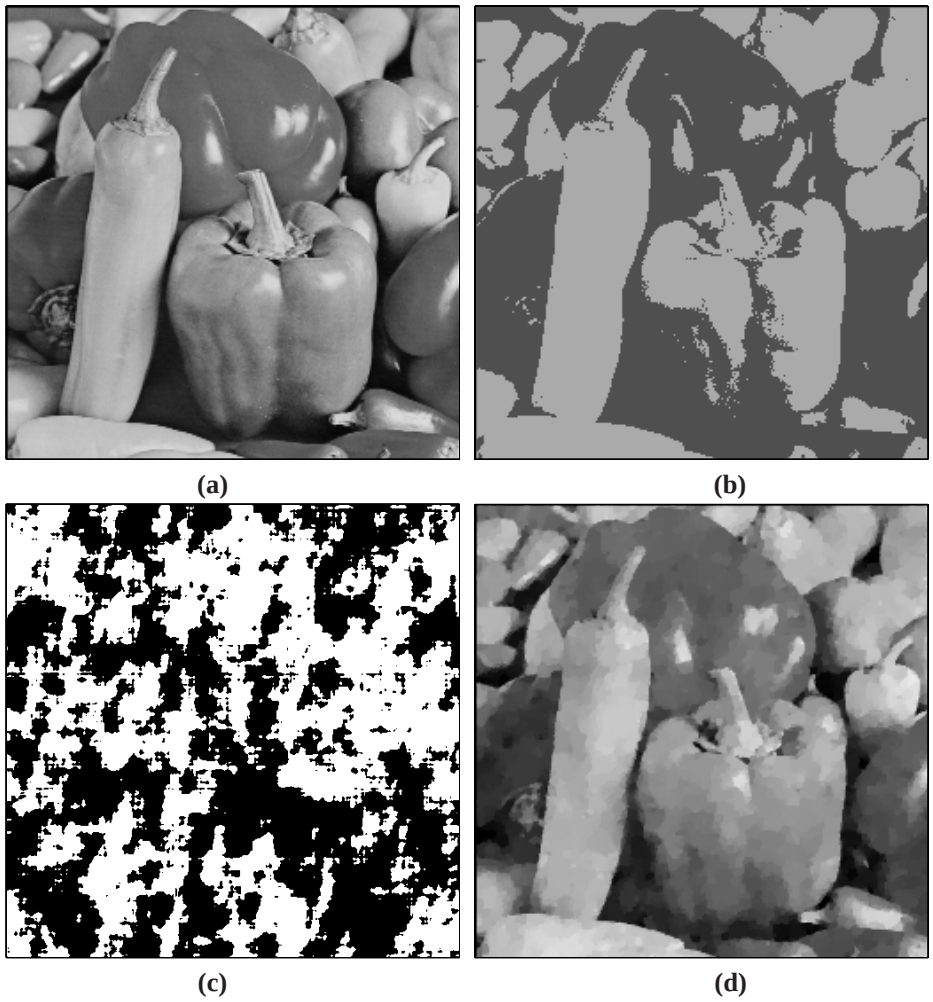

Fig. 6. *Peppers* image. (a) Original object (b) Conventional solution with minimum-error threshold (Lloyd-Max): SNR = 13.72 dB (c) Binary acquisition using our method (d) Reconstruction using our method: SNR = 19.10 dB

related to practical realization (e.g., precise estimation of the system PSF) will have to be addressed in further research. Let us mention that, according to the example of the gigapixel camera [18], the binary-sensor array of our system could be produced using standard memory-chip technology. Given the speed of 1-bit sensors, a variant of this acquisition system could be specifically designed (using several random masks) for high-speed video imaging. In that case, temporal redundancy may further improve the reconstruction quality and robustness.

7. Conclusions

Using optical phase masks, we have modeled a binary and incoherent optical device. In order to fit into the compressed-sensing formalism which is intrinsically discrete, our continuous model has been handled via B-spline expansions. Besides being physically meaningful, our optical model corresponds to a measurement matrix whose properties are suitable for numerical reconstruction. In particular, the convolutive structure of this matrix has allowed us to optimize

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010

Table 1. Reconstruction quality [dB] for different images. The results of our method with several sensor resolutions are reported on the left, while the standard-acquisition performance is shown in the middle. The TV norm of each reference image is given on the right.

Image / Resolution	64^2	128^2	256^2	512^2	Standard (256^2)	TV norm
Bird	18.35	21.27	22.95	23.66	15.78	$3.9 \cdot 10^5$
Bridge	11.87	13.36	14.46	14.66	12.54	$1.5 \cdot 10^6$
Cameraman	13.86	16.34	18.14	19.34	14.78	$9.2 \cdot 10^5$
Modified Shepp-Logan	7.28	12.57	17.33	22.47	7.56	$3.7 \cdot 10^5$

the reconstruction algorithm using fast preconditioning. This illustrates the deep link that can exist between forward-model structure and algorithmic performance. We have confirmed the feasibility of our approach by providing numerical simulations and concrete examples of image reconstructions. The latter also show that TV regularization is a suitable choice for the problem at hand.

Acknowledgments

We wish to thank the anonymous reviewers for their valuable suggestions and criticisms to improve the manuscript.

#121126 - $15.00 USD Received 7 Dec 2009; revised 5 Feb 2010; accepted 18 Feb 2010; published 24 Feb 2010