Distributed Representation of Geometrically Correlated Images With Compressed Linear Measurements

Vijayaraghavan Thirumalai, *Student Member, IEEE*, and Pascal Frossard, *Senior Member, IEEE*

***Abstract*—This paper addresses the problem of distributed coding of images whose correlation is driven by the motion of objects or the camera positioning. It concentrates on the problem where images are encoded with compressed linear measurements. We propose a geometry-based correlation model that describes the common information in pairs of images. We assume that the constitutive components of natural images can be captured by visual features that undergo local transformations (e.g., translation) in different images. We first identify prominent visual features by computing a sparse approximation of a reference image with a dictionary of geometric basis functions. We then pose a regularized optimization problem in order to estimate the corresponding features in correlated images that are given by quantized linear measurements. The correlation model is thus given by the relative geometric transformations between corresponding features. We then propose an efficient joint decoding algorithm that reconstructs the compressed images such that they are consistent with both the quantized measurements and the correlation model. Experimental results show that the proposed algorithm effectively estimates the correlation between images in multiview data sets. In addition, the proposed algorithm provides effective decoding performance that advantageously compares to independent coding solutions and state-of-the-art distributed coding schemes based on disparity learning.**

***Index Terms*—Correlation estimation, geometric transformations, quantization, random projections, sparse approximations.**

I. Introduction

In RECENT years, vision sensor networks have been gaining an ever-increasing popularity enforced by the availability of cheap semiconductor components. These networks typically produce highly redundant information so that an efficient estimation of the correlation between images becomes primordial for effective coding, transmission, and storage applications. The distributed coding paradigm becomes particularly attractive in such settings; it permits to efficiently exploit the correlation between images with low encoding complexity and minimal inter-sensor communication, which translates into power savings in sensor networks. One of the most important challenging tasks, however, resides in the proper modeling and estimation of the correlation between images.

In this paper, we consider the problem of finding an efficient distributed representation of correlated images where the common objects are displaced due to viewpoint change or motion of scene objects. In particular, we are interested in a scenario where the images are given under the form of few quantized linear measurements computed by very simple sensors. Even with such a simple acquisition stage, the images can be reconstructed under the condition that they have a sparse representation in a particular basis (e.g., discrete cosine transform (DCT) and wavelet) that is sufficiently different from the sensing matrices [3], [4]. Rather than independent image reconstruction, we are however interested in the joint reconstruction of the images. In particular, we focus here on estimating the underlying correlation between images from the compressed measurements. In contrary to most distributed compressive schemes in the literature, we want to estimate the correlation prior to image reconstruction for improved robustness at low coding rates.

We propose to model the correlation between images as geometric transformations of visual features, which provides a more efficient representation than block-based translational models that are commonly used in state-of-the-art coding solutions. We first compute the most prominent visual features in a reference image through a sparse approximation with geometric functions drawn from a parametric dictionary. Then, we formulate a regularized optimization problem whose objective is to identify the features in the compressed images that correspond to the prominent components in the reference images. Correspondences then define relative transformations between images that form the geometric correlation model. A regularization constraint ensures that the estimated correlation is consistent and corresponds to the actual motion of visual objects. We then use the estimated correlation in a new joint decoding algorithm that approximates multiple images. The joint decoding is cast as an optimization problem that warps the reference image according to the transformation described in the correlation information while enforcing the decoded images to be consistent with the

Manuscript received June 17, 2011; revised November 16, 2011; accepted January 20, 2012. Date of publication February 14, 2012; date of current version June 13, 2012. This work was supported in part by the Swiss National Science Foundation under Grant 200021-118230. This work was presented in part at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, TX, Mar. 2010 [1] and in part at the European Signal Processing Conference (EUSIPCO), Aalborg, Denmark, Aug. 2010 [2]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anthony Vetro.

The authors are with the Signal Processing Laboratory (LTS4), Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: vijayaraghavan.thirumalai@epfl.ch; pascal.frossard@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2188035

quantized measurements. We finally propose an extension of our algorithm to the joint decoding of multiview images.

While our novel framework could find applications in several problems such as distributed video coding or multiview imaging, we focus on the latter for illustrating the joint decoding performance. We show by experiments that the proposed algorithm computes a good estimation of the correlation between multiview images. In particular, the results confirm that the dictionaries based on geometric basis functions permit to capture the correlation more efficiently than a dictionary built on patches or blocks from the reference image [5]. In addition, we show that the estimated correlation model can be used to decode the compressed images by disparity compensation. Such a decoding strategy permits to outperform independent coding solutions based on JPEG 2000 and state-of-the-art distributed coding schemes based on disparity learning [6], [7] in terms of rate–distortion (RD) performance due to accurate correlation estimation. Finally, the experiments outline that the consistent prediction term proves to be effective in increasing the decoding quality of the images given by quantized linear measurements.

The rest of this paper is organized as follows. Section II briefly overviews the related work with emphasis on reconstruction from random projections and distributed coding algorithms. The geometric correlation model used in our framework is presented in Section III. Section IV describes the proposed regularized energy model for an image pair and the optimization algorithm. The consistent image prediction algorithm is described in Section V. Section VI describes the extension of our scheme to multiview images. Finally, experimental results are presented in Section VII, and Section VIII concludes this paper.

II. Related Work

We present in this section a brief overview of the related works in distributed image coding where we mostly focus on simple sensing solutions based on linear measurements. In recent years, signal acquisition based on random projections has actually received significant attention in many applications such as medical imaging, compressive imaging, or sensor networks. Donoho [3] and Candes *et al.* [4], [8] have shown that a small number of linear measurements contain enough information to reconstruct a signal, as long as it has a sparse representation in a basis that is incoherent with the sensing matrix. Rauhut *et al.* [9] extend the concept of signal reconstruction from linear measurements using redundant dictionaries. Signal reconstruction from linear measurements has been applied to different applications such as image acquisition [10]–[12] and video representation [13]–[15].

At the same time, the key in effective distributed representation certainly lies in the definition of good correlation models. Duarte *et al.* [16], [17] have proposed different correlation models for the distributed compression of correlated signals from linear measurements. In particular, they introduce three joint sparsity models (JSMs) in order to exploit the inter-signal correlation in the joint reconstruction. These three sparse models are respectively described by: 1) JSM-1, where the signals share a common sparse support plus a sparse innovation part specific to each signal; 2) JSM-2, where the signals share a common sparse support with different coefficients; and 3) JSM-3 with a non-sparse common signal with individual sparse innovation in each signal. These correlation models permit a joint reconstruction with a reduced sampling rate or equivalently a smaller number of measurements compared to the independent reconstruction for the same decoding quality. The sparsity models developed in [16] have been then applied to distributed video coding [18], [19] with random projections. The scheme in [18] used a modified gradient projection sparse algorithm [20] for the joint signal reconstruction. The authors in [19] have proposed a distributed compressive video coding scheme based on the sparse recovery with decoder side information. In particular, the prediction error between the original and side information frames is assumed to be sparse in a particular orthonormal basis (e.g., wavelet). Another distributed video coding scheme has been proposed in [5], which relies on an inter-frame sparsity model. A block of pixels in a frame is assumed to be sparsely represented by linear combinations of the neighboring blocks from the decoded key frames. In particular, an adaptive block-based dictionary is constructed from the previously decoded key frames and eventually used for signal reconstruction. Finally, iterative projection methods are used in [21] and [22] in order to ensure that a joint reconstruction of correlated images that are sparse in a dual-tree wavelet transform basis is consistent with the linear measurements in multiview settings. In general, state-of-the-art distributed compressive schemes [18]–[22] estimate the correlation model from two reconstructed reference images where the reference frames are reconstructed from the respective linear measurements by solving an l_2-TV or l_2-l_1 optimization problem. Unfortunately, reconstructing the reference images based on solving an l_2-l_1 or l_2-TV optimization problem is computationally expensive [3], [4]. In addition, the correlation model estimated from highly compressed reference images usually fails to capture the actual geometrical relationship between images. Motivated by these issues, we estimate in this paper a robust correlation model directly in the compressed domain without explicitly reconstructing the compressed images.

In multiview imaging or distributed video coding, the correlation is explained by the motion of objects or the change of viewpoint. Block-based translation models that are commonly used for correlation estimation fail to efficiently capture the geometry of objects; this results in a poor correlation estimation particularly when computed from the highly compressed images. Furthermore, most of the aforementioned schemes (except [5]) assume that the signal is sparse in a particular orthonormal basis (e.g., DCT or wavelet). This is also the case of the JSMs described above, which cannot be used to relate the scene objects by means of a local transform and unfortunately fail to provide an efficient joint representation of correlated images at low bit rates. It is more generic to assume the signals to be sparse in a redundant dictionary, which allows more flexibility in the design of the representation vectors. The most prominent geometric components in the images can be efficiently captured by dictionary functions. The correlation can be then estimated by comparing the most prominent

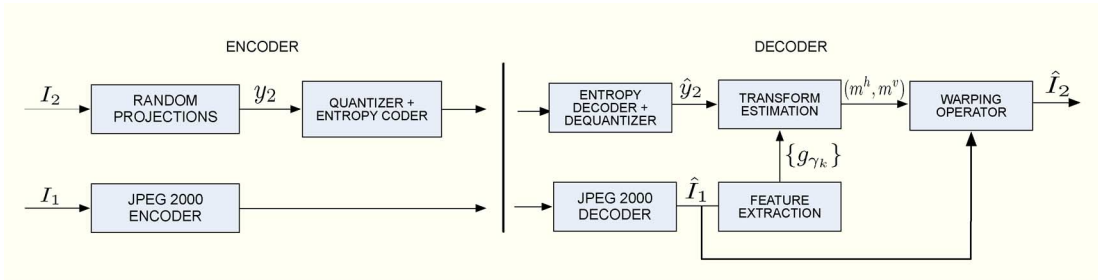

Fig. 1. Schematic of the proposed scheme. Images I_1 and I_2 are correlated through displacement of scene objects due to viewpoint change.

features in different images. Few works have been reported in the literature for the estimation of a correlation model using redundant structured dictionaries in multiview [23] or video applications [24]. However, these frameworks do not construct the correlation model from the linear measurements. In general, most of the schemes in classical disparity and motion estimation focus on estimating correlation model from original images [25], [26] and not from compressed images. We rather focus here on estimating the correlation from compressed images where the image is given with random linear measurements. The correlation model is built using the geometric transformations captured by a structured dictionary, which leads to an effective estimation of the geometric correlation between images.

Finally, the distributed schemes in the literature that are based on compressed measurements usually fail to estimate the actual number of bits for the image sequence representation (except [5]) and hence cannot be directly applied in practical coding applications. Quantization and entropy coding of the measurements is actually an open research problem due to the following two reasons: 1) the reconstructed signal from quantized measurements does not necessarily satisfy the consistent reconstruction property [27]; and 2) the entropy of the measurements is usually large, which leads to unsatisfactory coding performance in imaging applications [28]. Hence, it is essential to adapt the quantization techniques and reconstruction algorithms in order to reduce distortion in the reconstructed signal, such as [29] and [30]. The authors in [31] and [32] have also studied the asymptotic reconstruction performance of the signal under uniform and non-uniform quantization schemes. They have shown that a non-uniform quantization scheme usually gives smaller distortion in the reconstructed signal compared with a uniform quantization scheme. Recently, an optimal quantization strategy for the random measurements has been designed based on distributed functional scalar quantizers [33]. In this paper, we use a simple quantization strategy for realistic compression along with consistent prediction constraints in the joint decoding of correlated images in order to illustrate the potential of low-complexity sensing solutions in practical multiview imaging applications.

III. Framework

We first describe our framework for a pair of images, and then the extension to more images is presented in Section VI. We consider a pair of images I_1 and I_2 (with resolution $N = N_1 \times N_2$) that represent a scene taken from different viewpoints; these images are correlated through displacement of visual objects. The captured images are independently encoded and are transmitted to a joint decoder. The joint decoder estimates the relative transformations between the received signals and jointly decodes the images. The framework is illustrated in Fig. 1.

We focus on the particular problem where one of the images serves as a reference for the correlation estimation and the decoding of the second image. While the reference image I_1 could be encoded with any compression algorithm (e.g., JPEG, compressed sensing framework [12]), we choose here to encode the reference image I_1 with JPEG 2000 coding solutions. Next, we concentrate on the independent coding and joint decoding of the second image where the first image $\hat{I}_1$ serves as side information. The second image I_2 is projected on a random matrix Φ to generate the measurements $y_2 = \Phi I_2$. The measurements y_2 are quantized with a uniform quantization algorithm, and the quantized linear measurements are finally compressed with an entropy coder.

At the decoder, we first estimate the prominent visual features that carry the geometry information of the objects in the scene. In particular, the decoder computes a sparse approximation of the image $\hat{I}_1$ using a parametric dictionary of geometric functions. Such an approximation captures the most prominent geometrical features in the image $\hat{I}_1$. We then estimate the corresponding features in the second image I_2 directly from the quantized linear measurements $\hat{y}_2$ without implementing explicit image reconstruction steps. In particular, the corresponding features between images are related using a geometry-based correlation model where the correspondences describe local geometric transformations between images. The correlation information is further used to decode the compressed image $\hat{I}_2$ from the reference image $\hat{I}_1$. We finally ensure a consistent prediction of $\hat{I}_2$ by explicitly considering the quantized measurements $\hat{y}_2$ during the warping process. Before getting into the details of the correlation estimation algorithm, we describe the sparse approximation algorithm and the geometry-based correlation model built on a parametric dictionary.

We now describe the geometric correlation model that is based on matching the sparse geometric features in different images. We first compute a sparse approximation of the reference image $\hat{I}_1$ using geometric basis functions in a structured dictionary $\mathcal{D} = \{g_\gamma\}$, where g_γ is called an *atom*. The dictionary $\mathcal{D}$ is typically constructed by applying geometric transformations

Fig. 2. Sample Gaussian atoms with mother function $g(x,y) = (1/\sqrt{\pi})\exp(-(x^2+y^2))$ that undergo different sets of transformations.

Fig. 3. Illustration of the atom transform F in the approximation of the correlated images: (a) original correlated synthetic images and (b) sparse approximation of the images using atoms in the dictionary. The *rectangle* and *square* objects are related with transformations F^1 and F^2, respectively.

(given by unitary operator $U(\gamma)$) to a generating function g to form the atom g_γ. A geometric transformation indexed by γ consists of a combination of operators for anisotropic scales s_x and s_y, rotation θ, and translations t_x and t_y. For example, when g is a Gaussian function $g(x,y) = (1/\sqrt{\pi})\exp(-(x^2+y^2))$, the transformation g_γ is expressed as

$$g_\gamma(x,y) = \frac{1}{\sqrt{\pi}} \exp\left(-\left(g_1^2 + g_2^2\right)\right) \tag{1}$$

with

$$g_1 = \frac{\cos(\theta)(x-t_x) + \sin(\theta)(y-t_y)}{s_x}$$
$$g_2 = \frac{\cos(\theta)(y-t_y) - \sin(\theta)(x-t_x)}{s_y}.$$

In Fig. 2, we illustrate Gaussian atoms for different translation, rotation, and anisotropic scaling parameters. Now, we can write the linear approximation of the reference image $\hat{I}_1$ with functions in $\mathcal{D}$ as

$$\hat{I}_1 \approx \sum_{k=1}^{K} c_k g_{\gamma_k} \tag{2}$$

where $\{c_k\}$ represents the coefficient vector. The K number of atoms used in the approximation of $\hat{I}_1$ is usually much smaller than the dimensions of image $\hat{I}_1$. We use here a suboptimal solution based on matching pursuit [34], [35] in order to estimate the set of K atoms.

The correlation between images can be now described by the geometric deformation of atoms in different images [23], [24]. Once the reference image $\hat{I}_1$ is approximated as given in (2), the second image I_2 could be approximated with transformed versions of the atoms used in the approximation of $\hat{I}_1$. We can thus approximate I_2 as

$$I_2 \approx \sum_{k=1}^{K} c_k F^k(g_{\gamma_k}) = \sum_{k=1}^{K} c_k g_{\gamma'_k} \tag{3}$$

where $F^k(g_{\gamma_k})$ represents a local geometrical transformation of the atom g_{γ_k}. Due to the parametric form of the dictionary, it is interesting to note that the transformation F^k on g_{γ_k} boils down to a transformation $\delta\gamma$ of the atom parameters, i.e.,

$$F^k(g_{\gamma_k}) = U(\delta\gamma)g_{\gamma_k} = U(\delta\gamma \circ \gamma_k)g = g_{\delta\gamma\circ\gamma_k} = g_{\gamma'_k}. \tag{4}$$

For clarity, we show in Fig. 3 a sample synthetic correlated image pair and their sparse approximations using atoms in the dictionary. We see that the sparse approximations of images can be described with the transforms $\{F^k\}$ of atom parameters.

The true transformations $\{F^k\}$, however, are unknown in practical distributed coding applications. Therefore, the main challenge in our framework consists in estimating the local geometrical transformations $\{F^k\}$ when the second image I_2 is available in the form of quantized linear measurements $\hat{y}_2$.

IV. Correlation Estimation From Compressed Linear Measurements

A. Regularized Optimization Problem

We now describe our optimization framework for estimating the correlation between images. Given the set of K atoms $\{g_{\gamma_k}\}$ that approximates the first image $\hat{I}_1$, the correlation estimation problem consists in finding the corresponding visual patterns in the second image I_2 that is given only by compressed random measurements $\hat{y}_2$. This is equivalent to finding the correlation between images I_1 and I_2 with the JSM based on local geometrical transformations, as described in Section III.

In more details, we are looking for a set of K atoms in I_2 that corresponds to the K visual features $\{g_{\gamma_k}\}$ selected in the first image. We denote their parameters by Λ, where $\Lambda = (\gamma'_1, \gamma'_2, \ldots, \gamma'_K)$ for some γ'_k, $\forall k, 1 \le k \le K$. We propose to select this set of atoms $\{g_{\gamma'_k}\}$ in a regularized energy minimization framework as a tradeoff between efficient approximation of I_2 and smoothness or consistency of the local transformations between images. The energy model E proposed in our scheme is expressed as

$$E(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda) \tag{OPT-1}$$

where E_d and E_s represent the data and smoothness terms, respectively, and α_1 is the regularization parameter that balances the importance of the data and smoothness terms. The solution to our correlation estimation is given by the set of K atom parameters Λ^* that minimizes energy E, i.e.,

$$\Lambda^* = \arg\min_{\Lambda \in S} E(\Lambda). \tag{5}$$

Parameter S represents the search space given by

$$S = \{(\gamma'_1, \gamma'_2, \ldots, \gamma'_K) \mid \gamma'_k = \delta\gamma \circ \gamma_k, 1 \le k \le K, \delta\gamma \in \mathcal{L}\}. \tag{6}$$

The multidimensional search window $\mathcal{L} \subset \mathbb{R}^5$ is defined as $\mathcal{L} = [-\delta t_x, \delta t_x] \times [-\delta t_y, \delta t_y] \times [-\delta\theta, \delta\theta] \times [-\delta s_x, \delta s_x] \times [-\delta s_y, \delta s_y]$, where δt_x, δt_y, $\delta\theta$, δs_x, and δs_y determine the window size for each of the atom parameters (i.e., translations t_x and t_y, rotation θ, and scales s_x and s_y). Even if our formulation is able to handle complex transformations, they generally take the form of motion vectors or disparity information

in video coding or stereo imaging applications. The label sets and the search space S are drastically reduced in this case. The terms used in OPT-1 are described in the next paragraphs.

B. Data Cost Function

The data cost function computes (in the compressed domain) the accuracy of the sparse approximation of the second image with geometric atoms linked to the reference image. The decoder receives the measurements $\hat{y}_2$ that are computed by the quantized projections of I_2 onto a sensing matrix Φ. For each set of K atom parameters $\Lambda = \{\gamma'_k\}$, the data term E_d reports the error between measurements $\hat{y}_2$ and orthogonal projection of $\hat{y}_2$ onto Ψ_Λ that is formed by the compressed versions of the atoms, i.e., $\Psi_\Lambda = \Phi[g_{\gamma'_1}|g_{\gamma'_2}|\dots|g_{\gamma'_K}]$. It turns out that the orthogonal projection of $\hat{y}_2$ onto the subspace spanned by (column) vectors in Ψ_Λ is given as $\Psi_\Lambda \Psi_\Lambda^\dagger \hat{y}_2$, where $\dagger$ represents the pseudoinverse operator. More formally, the data cost is computed using the following relation:

$$E_d(\Lambda) = \left\| \hat{y}_2 - \Psi_\Lambda \Psi_\Lambda^\dagger \hat{y}_2 \right\|_2^2 = \|\hat{y}_2 - \Psi_\Lambda c\|_2^2. \tag{7}$$

The data cost function given in (7) first calculates the coefficients $c = \Psi_\Lambda^\dagger \hat{y}_2$ and then measures the distance between the observations $\hat{y}_2$ and $\Psi_\Lambda c$. In other words, data cost function E_d accounts for the intensity variations between images by estimating the coefficients c of the warped atoms.

However, when the measurements are quantized, the coefficient vector c fails to properly account for the error introduced by quantization. The quantized measurements only provide the index of the quantization interval containing the actual measurement value, and the actual measurement value could be any point in the quantization interval. Let $y_2(i)$ be the ith coordinate of the original measurement and $\hat{y}_2(i)$ be the corresponding quantized value. It can be noted that the joint decoder has only access to the quantized value $\hat{y}_2(i)$ and not the original value $y_2(i)$. Henceforth, the joint decoder knows that the quantized measurement lies within the quantization interval, i.e., $\hat{y}_2(i) \in \mathcal{R}_{\hat{y}(i)} = (r_i \;\; r_{i+1}]$, where r_i and r_{i+1} define the lower and upper bounds of the quantizer bin $\mathcal{Q}_i$. We therefore propose to refine the data term in the presence of quantization by computing a coefficient vector $\tilde{c}$ as the most consistent coefficient vector when considering all the possible measurement vectors that can result in the quantized measurements vector $\hat{y}_2$. In more details, the quantized measurements $\hat{y}_2$ can be produced by all the observation vectors $\tilde{y}_2 \in \mathcal{R}_{\hat{y}}$, where $\mathcal{R}_{\hat{y}}$ is the Cartesian product of all the quantized regions $\mathcal{R}_{\hat{y}(i)}$, i.e., $\mathcal{R}_{\hat{y}} = \prod_i \mathcal{R}_{\hat{y}(i)}$. The data cost term given in (7) can be thus modified as

$$\tilde{E}_d(\Lambda) = \min_{\tilde{c}, \tilde{y}_2} \|\tilde{y}_2 - \Psi_\Lambda \tilde{c}\|_2^2, \text{ s.t. } \tilde{y}_2 \in \mathcal{R}_{\hat{y}}. \tag{8}$$

Therefore, robust data term $\tilde{E}_d(\Lambda)$ first jointly estimates coefficients $\tilde{c}$ and measurements $\tilde{y}_2$, and then it computes the distance between $\tilde{y}_2$ and $\Psi_\Lambda \tilde{c}$. It can be shown that the Hessian of the objective function $h(\tilde{c}, \tilde{y}_2) = \|\tilde{y}_2 - \Psi_\Lambda \tilde{c}\|_2^2$ in (8) is positive semi-definite, i.e., $\nabla^2 h \succeq 0$, and hence, the objective function h is convex. In addition, region $\mathcal{R}_{\hat{y}}$ forms a closed convex set as each region $\mathcal{R}_{\hat{y}_i} = (r_i \;\; r_{i+1}], \forall i$ forms a convex set. Henceforth, the optimization problem given in (8) is convex, which leads to effective solutions.

C. Smoothness Cost Function

The goal of the smoothness term E_s in OPT-1 is to regularize the atom transformations such that the transformations are coherent for neighbor atoms. In other words, the atoms in a spatial neighborhood are likely to undergo similar transformations when the correlation between images is due to object or camera motion. Instead of directly penalizing the transformation $\{F^k\}$ to be coherent for neighbor atoms, we propose to generate a dense disparity (or motion) field from the atom transformations and to penalize the disparity (or motion) field such that it is coherent for adjacent pixels. This regularization is easier to handle than a regular set of transformations $\{F^k\}$ and directly corresponds to the physical constraints that explain the formation of correlated images.

In more details, for a given transformation value

$$\delta\gamma = (t'_x - t_x, t'_y - t_y, \theta' - \theta, s_x/s'_x, s_y/s'_y)$$

at pixel $\mathbf{z}$, we compute the horizontal component $\mathbf{m}^h$ and vertical component $\mathbf{m}^v$ of the motion field as

$$\begin{bmatrix} \mathbf{m}^h(\mathbf{z}) \\ \mathbf{m}^v(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} m(\mathbf{z}) - t_x \\ n(\mathbf{z}) - t_y \end{bmatrix} - SRT \tag{9}$$

where $(m(\mathbf{z}), n(\mathbf{z}))$ represent the Euclidean coordinates. Matrices S, R, and T represent the grid transformations due to scale, rotation, and translation changes, respectively. They are defined as

$$S = \begin{bmatrix} s_x/s'_x & 0 \\ 0 & s_y/s'_y \end{bmatrix}, \quad R = \begin{bmatrix} \cos(\theta' - \theta) & \sin(\theta' - \theta) \\ -\sin(\theta' - \theta) & \cos(\theta' - \theta) \end{bmatrix}$$
$$T = \begin{bmatrix} m(\mathbf{z}) - t_x - (t'_x - t_x) \\ n(\mathbf{z}) - t_y - (t'_x - t_x) \end{bmatrix}.$$

Finally, the smoothness cost E_s in OPT-1 is given as

$$E_s(\Lambda) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{N}} V_{\mathbf{z}, \mathbf{z}'} \tag{10}$$

where $\mathbf{z}$ and $\mathbf{z}'$ are the adjacent pixel locations and $\mathcal{N}$ is the usual 4-pixel neighborhood. The term $V_{\mathbf{z}, \mathbf{z}'}$ in (10) captures the distance between local transformations in neighboring pixels. It is defined as

$$V_{\mathbf{z}, \mathbf{z}'} = \min\left(\left| \mathbf{m}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z}') \right| + \left| \mathbf{m}^v(\mathbf{z}) - \mathbf{m}^v(\mathbf{z}') \right|, \tau \right). \tag{11}$$

The parameter τ in (11) sets a maximum limit to the penalty; it helps to preserve the discontinuities in the transformation field that exist at the boundaries of visual objects [36].

D. Optimization Algorithm

We now describe the optimization methodology that is used to solve OPT-1. Recall that our objective is to assign a transformation F to each atom g_{γ_k} in the reference image in order to build a set of smooth local transformations that is consistent with quantized measurements $\hat{y}_2$. The candidate transformations are chosen from a finite set of labels $\mathcal{L} = \mathcal{L}_x \times \mathcal{L}_y \times \mathcal{L}_\theta \times \mathcal{L}_a \times \mathcal{L}_b$, where $\mathcal{L}_x$, $\mathcal{L}_y$, $\mathcal{L}_\theta$, $\mathcal{L}_a$, and $\mathcal{L}_b$ refer to the label sets corresponding to translation along x- and y-directions, rotations, and

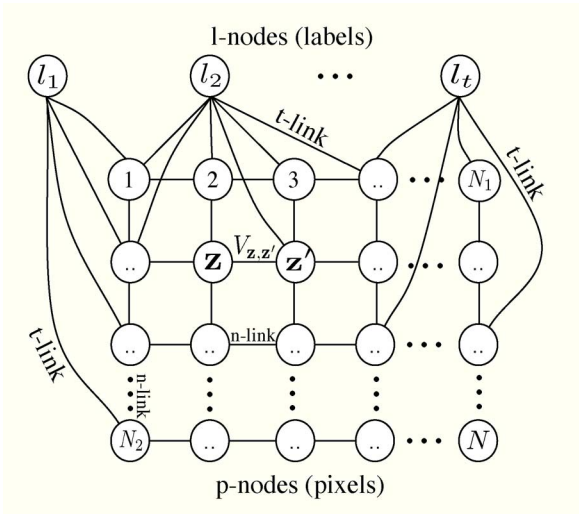

Fig. 4. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed using the set of vertices $\mathcal{V} = \mathcal{Z} \cup \mathcal{L}$, where the pixel nodes $\mathcal{Z} = \{1, 2, \ldots, N\}$ and label nodes $\mathcal{L} = \{l_1, l_2, \ldots, l_t\}$. Each pixel $\mathbf{z}$ is connected to the l-node with a t-link. Some t-links are omitted for the sake of clarity. The pixels $\mathbf{z}, \mathbf{z}' \in \mathcal{N}$ are connected with an n-link. The correlation solution is given a multiway cut that leaves each p-node connected with only one t-link [36].

anisotropic scales, respectively [see (6)]. One could use an exhaustive search on the entire label $\mathcal{L}$ to solve OPT-1. However, the cost for such a solution is high as the size of the label set $\mathcal{L}$ exponentially grows with the size of the search windows δt_x, δt_y, $\delta\theta$, δs_x, and δs_y. Rather than doing an exhaustive search, we use graph-based minimization techniques that converge to strong local minima or global minima in a polynomial time with tractable computational complexity [36], [37].

Usually, in graph cut algorithms, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed using sets of vertices $\mathcal{V}$ and edges $\mathcal{E}$. The sets of vertices are given as $\mathcal{V} = \mathcal{Z} \cup \mathcal{L}$, where $\mathcal{Z}$ defines nodes corresponding to the pixels in the images (p-nodes) and $\mathcal{L}$ defines the label nodes (l-nodes), as shown in Fig. 4. The p-nodes that are in the neighborhood $\mathcal{N}$ are connected by an edge called n-link. The cost of n-link usually corresponds to the penalty of assigning different labels to the adjacent pixels, as given by $V_{\mathbf{z},\mathbf{z}'}$. In addition, each p-vertex $\mathbf{z}$ is connected to the l-node by an edge called t-link. The cost of a t-link connecting a pixel and a label corresponds to the penalty of assigning the corresponding label to that pixel; this cost is normally derived from the data term. The final solution is given by a multiway cut that leaves each p-vertex connected with exactly one t-link. For more details, we refer the reader to [36].

In order to solve our OPT-1 problem, we first need to map our cost functions on the graph in order to assign weights to the n- and t-links. For a given pair of transformation labels at pixels $\mathbf{z}$ and $\mathbf{z}'$, it is straightforward to calculate the weights of the n-links using (11). It should be noted that the motion field for a given label is computed using (9). We now describe how to calculate the cost of the t-links based on data cost $E_d(\Lambda)$. Let $\mathcal{Z}_k$ be the set of pixels in the support of the atom g_{γ_k} that is given as

$$\mathcal{Z}_k = \{\mathbf{z} = (x, y) | g_{\gamma_k}(x, y) > \epsilon\} \tag{12}$$

where $\epsilon > 0$ is a constant. Using this definition, we calculate the t-link penalty cost of connecting a label node $l_k \in \mathcal{L}$ to all the pixel nodes $\mathbf{z}$ in the support of the atom g_{γ_k} as $E_d(\Lambda)$ given in (7), where $\Lambda = (\gamma_1, \gamma_2, \ldots, l_k \circ \gamma_k, \ldots, \gamma_K)$. That is, the t-link cost computed between the label l_k and pixels $\mathbf{z}, \forall \mathbf{z} \in \mathcal{Z}_k$ is $E_d(\Lambda)$ with $\Lambda = (\gamma_1, \gamma_2, \ldots, l_k \circ \gamma_k, \ldots, \gamma_K)$. However, due to atom overlapping, the pixels in the overlapping region could be assigned more than one label. In such cases, we compute the cost corresponding to the index k' of the atom that has the maximum atom response. Index k' is computed as

$$k' = \arg \max_{k=1,2,\ldots,K} w_{\mathbf{z}}^{(k)} \tag{13}$$

where $w_{\mathbf{z}}^{(k)}$ is the response of the kth atom at location $\mathbf{z}$, i.e., $w_{\mathbf{z}}^{(k)} = g_{\gamma_k}(\mathbf{z}) = g_{\gamma_k}(x, y)$. After mapping the cost functions on the graph, we calculate the correlation solution using a maximum-flow/minimum-cut algorithm [36]. Finally, the data term E_d in OPT-1 can be replaced with the robust data term $\tilde{E}_d$ given in (8) in order to provide robustness to quantization errors. The resulting optimization problem can be efficiently solved using graph cut algorithms as described above.

E. Complexity Considerations

We now briefly discuss the computational complexity of our correlation estimation algorithm which can be basically divided into two stages. The first stage finds the most prominent features in the reference image using sparse approximations in a structured dictionary. The second stage estimates the transformation for all the features in the reference image by solving the OPT-1 regularized optimization problem.

Overall, our framework offers a very simple encoding stage with image acquisition based on random linear projections. The computational burden is shifted to the joint decoder, which can still tradeoff complexity and performance. Even if the decoder is able to handle computationally complex tasks in our framework, the complexity of our system stays reasonable due to the efficiency of graph cut algorithms whose complexity is bounded by a low-order polynomial [36], [37]. Complexity can be further reduced in both stages compared to the generic implementation proposed above. For example, the complexity of the sparse approximations of the reference image can be significantly reduced using a tree-structured dictionary without significant loss in the approximation performance [38]. In addition, a block-based dictionary can be used in order to reduce the complexity of the transformation estimation problem with block-based computations. However, experiments show that this comes at a price of a performance penalty in the correlation estimation accuracy. Overall, it is clear that the decoding scheme proposed above offers high flexibility with an interesting tradeoff between the complexity and the performance. For example, one might decide to use the simple data cost E_d even when the measurements are quantized; it leads to a simpler scheme but to a reduced correlation estimation accuracy.

V. Consistent Image Prediction by Warping

After correlation estimation, one can simply reconstruct an approximate version of the second image $\hat{I}_2$ by warping the reference image $\hat{I}_1$ using a set of local transformations that forms

the warping operator $\mathcal{W}_\Lambda$ (see Fig. 1). The resulting approximation is, however, not necessarily consistent with quantized measurements $\hat{y}_2$; the measurements corresponding to the projection of the image $\hat{I}_2$ on the sensing matrix Φ are not necessarily equal to $\hat{y}_2$. The consistency error might be quite significant because the atoms used to compute the correlation and the warping operator do not optimally handle the texture information.

We therefore propose to add a consistency term E_t in the OPT-1 energy model and to form a new optimization problem for improved image prediction. The consistency term forces the image predicted through the warping operator to be consistent with the quantized measurements. We define this additional term E_t as the square of the l_2 norm difference between the quantized measurements generated from the reconstructed image $\hat{I}_2 = \mathcal{W}_\Lambda(\hat{I}_1)$ and measurements $\hat{y}_2$. The consistency term E_t is written as

$$E_t(\Lambda) = \left\| \hat{y}_2 - \mathcal{Q}[\Phi \hat{I}_2] \right\|_2^2 = \left\| \hat{y}_2 - \mathcal{Q}\left[\Phi \mathcal{W}_\Lambda(\hat{I}_1)\right] \right\|_2^2 \quad (14)$$

where $\mathcal{Q}$ is the quantization operator. In the absence of quantization, the consistency term simply reads as

$$E_t(\Lambda) = \left\| y_2 - \Phi \mathcal{W}_\Lambda(\hat{I}_1) \right\|_2^2. \quad (15)$$

We then merge the three cost functions E_d, E_s, and E_t with regularization constants α_1 and α_2 in order to form a new energy model E_R for consistent image prediction. It is given as

$$E_R(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda) + \alpha_2 E_t(\Lambda). \quad \text{(OPT-2)}$$

We now highlight the differences between the terms E_d and E_t used in OPT-2. Data cost E_d adapts the coefficient vector to consider the intensity variations between images, but it fails to properly handle the texture information. On the other hand, consistency term E_t warps the atoms by considering the texture information in the reconstructed image $\hat{I}_1$, but it fails to carefully deal with the intensity variations between images. These two terms therefore impose different constraints on the atom selection that effectively reduce the search space. We have experimentally observed that the quality of the predicted image $\hat{I}_2$ is maximized when all three terms are activated in the OPT-2 optimization problem.

We propose to use the optimization method based on graph cuts described in Section IV.D in order to solve OPT-2. In particular, we map the consistency cost E_t into the graph (see Fig. 4) in addition to the data cost E_d and smoothness cost E_s. For a given $\Lambda = (\gamma_1, \gamma_2, \ldots, l_k \circ \gamma_k, \ldots, \gamma_K)$, we propose to compute the t-link cost of connecting the label $l_k \in \mathcal{L}$ to the pixels $\mathbf{z}, \forall \mathbf{z} \in \mathcal{Z}_k$ as a cumulative sum of $E_d(\Lambda) + \alpha_2 E_t(\Lambda)$. In the overlapping regions, as described earlier, we take the value corresponding to the atom index k' that has maximum response as given in (13). Then, the n-link weights for the adjacent pixels $\mathbf{z}$ and $\mathbf{z}'$ are computed based on (11). After mapping the cost functions on the graph, the correlation solution is finally estimated using maximum-flow/minimum-cut algorithms [36]. Finally, the data cost E_d in OPT-2 can be again replaced by the robust data term $\tilde{E}_d$ given in (8). We show later that the performance of our scheme improves by using the robust data term $\tilde{E}_d$ in the presence of quantization. At last, the complexity of estimating the correlation model with the OPT-2 problem is tractable due to the efficiency of graph cut algorithms [36], [37].

VI. Correlation Estimation of Multiple Image Sets

So far, we have focused on the distributed representation of image pairs. In this section, we describe the extension of our framework to the data sets with J correlated images denoted as $I_1, I_2, \ldots, I_J$. Similar to the stereo setup, we consider I_1 as the reference image. This image is given in a compressed form $\hat{I}_1$, and its prominent features are extracted at a decoder with a sparse approximation over the dictionary $\mathcal{D}$ (see Section III). The images $I_2, \ldots, I_J$ are independently sensed using the measurement matrix Φ, and their respective measurements $y_2, \ldots, y_J$ are quantized and entropy coded. Our framework can be applied to image sequences or multiview imaging. For the sake of clarity, we focus on a multiview imaging framework where the multiple images are captured from different viewpoints.

We are interested in estimating a depth map Z that captures the correlation among J images by assuming that the camera parameters are given *a priori*. The depth map is constructed using the set of K features $\{g_{\gamma_k}\}$ in the reference image and the quantized measurements $\hat{y}_2, \ldots, \hat{y}_J$. We assume that the depth values Z are discretized such that the inverse depth $1/Z$ is uniformly sampled in the range $[1/Z_{\max}, 1/Z_{\min}]$, where $Z_{\min}$ and $Z_{\max}$ are the minimal and maximal depths in the scene, respectively [39]. The problem is equivalent to finding a set of labels $l \in \mathcal{L}$ that effectively captures the depth information for each atom g_{γ_k} or pixel $\mathbf{z}$ in the reference image, where $\mathcal{L}$ is a discrete set of labels corresponding to different depths. We propose to estimate the depth information with an energy minimization problem OPT-3 which includes three cost functions as follows:

$$H(\Lambda) = H_d(\Lambda) + \alpha_1 H_s(\Lambda) + \alpha_2 H_t(\Lambda) \quad \text{(OPT-3)}$$

where H_d, H_s, and H_t represent the data, smoothness, and consistency terms, respectively. These three terms are balanced with regularization constants α_1 and α_2.

The data term H_d assigns a set of labels $l_1, l_2, \ldots, l_K$ respectively to the K atoms $g_{\gamma_1}, g_{\gamma_2}, \ldots, g_{\gamma_K}$ while respecting consistency with the quantized measurements. It reads as

$$H_d(\{l_k\}) = \sum_{j=2}^{J} \left\| \hat{y}_j - \Psi_\Lambda^j \Psi_\Lambda^{j\dagger} \hat{y}_j \right\|_2^2 \quad (16)$$

where

$$\Psi_\Lambda^j = \Phi[\mathcal{P}_j(g_{\gamma_1}, l_1), \mathcal{P}_j(g_{\gamma_2}, l_2), \ldots, \mathcal{P}_j(g_{\gamma_k}, l_k), \ldots, \mathcal{P}_j(g_{\gamma_K}, l_K)].$$

Operator $\mathcal{P}_j(g_{\gamma_k}, l)$ represents the projection of the atom g_{γ_k} to the jth view when the local transformation is given by depth label l (see Fig. 5). It can be noted that the data term in (16) is similar to the data term described earlier for image pairs [see (7)] except that the sum is computed for all the views. Depending on the relative position of the jth camera with respect to the reference camera, the projection $\mathcal{P}_j(g_{\gamma_k}, l)$ can involve changes in the translation, rotation, or scaling parameter or combinations of them. Therefore, the projection $\mathcal{P}_j(g_{\gamma_k}, l)$ of the atom g_{γ_k} to the jth view approximately corresponds to another atom in the

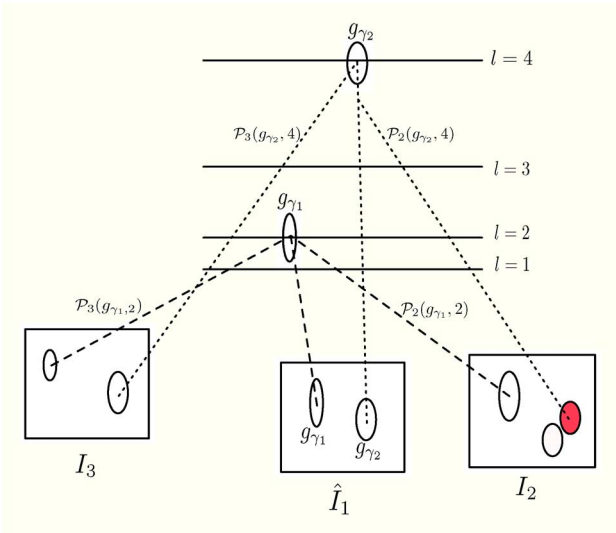

Fig. 5. Illustration of the atom interactions in the multiview imaging scenario. The original position of the features in all the images is marked in black color. The projection of the first feature g_{γ_1} at $l=2$ in views I_2 and I_3 corresponds to the actual position of the feature in the respective views and thus forms a valid 3-D region at $l=2$. Meanwhile, the projection of the second feature g_{γ_2} at $l=4$ corresponds to the actual position only in view I_3 but not in view I_2 (highlighted in red color). Hence, the second feature does not intersect at $l=4$, which results in suboptimal solution at $l=4$.

dictionary $\mathcal{D}$. It is interesting to note that the data cost is minimal if the projection of the atom g_{γ_k} onto another view corresponds to its actual position in this view.[1] This happens when the depth label l corresponds to the true distance to the visual object represented by the atom g_{γ_k}. For example, the projection of the feature g_{γ_1} in Fig. 5 corresponds to the actual position of the features in views I_2 and I_3. Therefore, the data cost for this feature g_{γ_1} at label $l=2$ is minimal. On the other hand, the projection of the feature g_{γ_2} is far from the actual position of the corresponding feature in view I_2. The corresponding data cost $\|y_2 - \Psi_\Lambda^2 \Psi_\Lambda^{2\dagger} y_2\|_2^2$ is high in this case, which indicates a suboptimal estimation of the depth label $l=4$.

The smoothness cost H_s enforces consistency in the depth label for the adjacent pixels $\mathbf{z}$ and $\mathbf{z}'$. It is given as

$$H_s = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} \min\left(|Z(\mathbf{z}) - Z(\mathbf{z}')|, \tau\right) \tag{17}$$

where τ is a constant and $\mathcal{N}$ represents the usual 4-pixel neighborhood. Finally, the consistency term H_t favors depth labels that lead to image predictions that are consistent with the quantized measurements. We compute the consistency for all the views as the cumulative sum of terms E_t given in (14). More formally, the consistency term H_t in the multiview scenario is computed as

$$\begin{aligned} H_t\left(\{l_k\}\right) &= \sum_{j=2}^{J} \left\| \hat{y}_j - \mathcal{Q}[\Phi \hat{I}_j] \right\|_2^2 \\ &= \sum_{j=2}^{J} \left\| \hat{y}_j - \mathcal{Q}\left[\Phi \mathcal{W}^j\left(\hat{I}_1, \{l_k\}\right)\right] \right\|_2^2 \end{aligned} \tag{18}$$

[1]We assume that there is no occlusion.

where $\mathcal{W}^j(\hat{I}_1, \{l_k\})$ predicts the jth view using the set of labels $\{l_k\}$ and the set of K atoms $\{g_{\gamma_k}\}$. Finally, the OPT-3 optimization problem can be solved in polynomial time using the graph-based optimization methodologies described in Section IV-D. In this case, the weights to the t-links connecting between the label l_k and the pixels $\mathbf{z}, \forall \mathbf{z} \in \mathcal{Z}_k$ are assigned as $H_d + \alpha_2 H_t$. The n-link cost for the neighboring pixels $\mathbf{z}, \mathbf{z}' \in \mathcal{N}$ is assigned as $\min(|Z(\mathbf{z}) - Z(\mathbf{z}')|, \tau)$.

VII. Experimental Results

A. Setup

In this section, we report the performance of the correlation estimation algorithms in stereo and multiview imaging frameworks. In order to compute a sparse approximation of the reference image at a decoder, we use a dictionary $\mathcal{D}$ that is constructed using two generating functions, as explained in [35]. The first one consists of 2-D Gaussian functions in order to capture the low-frequency components (see Fig. 2). The second function represents a Gaussian function in one direction and the second derivative of a Gaussian in the orthogonal direction in order to capture the edges. The discrete parameters of the functions in the dictionary are chosen as follows. Translation parameters t_x and t_y take any positive value and cover the full height N_1 and width N_2 of the image. Ten rotation parameters are used between 0 and π with increments of $\pi/18$. Five scaling parameters are equidistributed in the logarithmic scale from 1 to $N_1/8$ vertically and 1 to $N_2/9.77$ horizontally. Image I_2 is captured by random linear projections using a scrambled block Hadamard transform with a block size of 8 [12]. Measurements y_2 are quantized using a uniform quantizer. The bit rate is computed by encoding the quantized measurements using an arithmetic coder. Unless stated differently, the parameters α_1 and α_2 in the optimization problems are selected based on trial-and-error experiments such that the estimated transformation field maximizes the quality of the predicted image $\hat{I}_2$.

B. Generic Transformation

We first study the performance of our scheme with a pair of synthetic images that contains three objects. Original images I_1 and I_2 are given in Fig. 6(a) and (b), respectively. It is clear that the common objects in the images have different positions and scales. The absolute error between the original images is given in Fig. 6(c), where the peak signal-to-noise ratio (PSNR) between I_1 and I_2 is found to be 15.6 dB.

We encode the reference image I_1 to a quality of 35 dB, and the number of features used for the approximation of $\hat{I}_1$ is set to $K=15$. The transformation field is estimated with $\delta t_x = \delta t_y = 3$ pixels, $\delta s_x = \delta s_y = 2$ samples, and $\delta\theta = 0$. We first estimate the transformation field with the OPT-1 problem by setting $\alpha_1 = 0$, i.e., smoothness term E_s is not activated. The resulting motion field is shown in Fig. 6(d). In Fig. 6(d), we observe that the proposed scheme gives a good estimation of the transformation field even with a 5% measurement rate that is quantized with 2 bits. We further see that the image $\hat{I}_2$ predicted with the help of the estimated correlation information is closer to the original image I_2 than to I_1 [see Fig. 6(e)]. We then include the consistency term in addition to the data

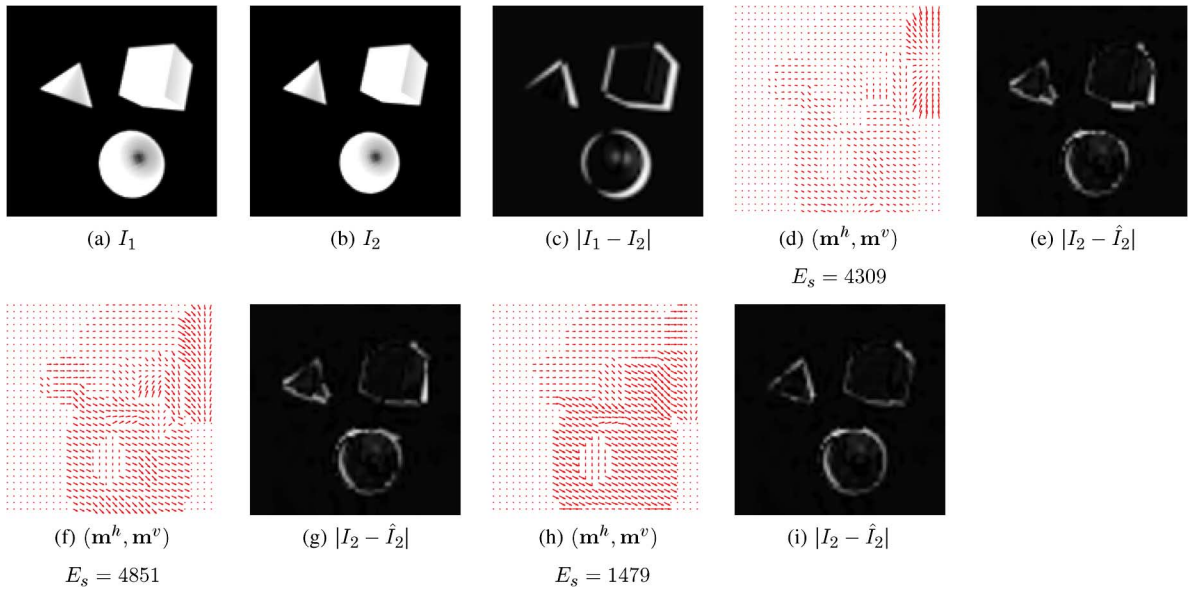

Fig. 6. Comparison of the estimated motion fields and the predicted images with the OPT-1 and OPT-2 problems in the synthetic scene. The motion field is estimated using a measurement rate of 5% with a 2-bit quantization. (a) Original image I_1. (b) Original image I_2. (c) Absolute error between I_1 and I_2. (d) Motion field estimated with OPT-1 without activating E_s, i.e., $\alpha_1 = 0$. (e) Prediction error with OPT-1 when the motion field in (d) is used for image prediction. (f) Motion field estimated with OPT-2 without activating E_s. (g) Prediction error with OPT-2 when the motion field in (f) is used for image prediction. (h) Motion field estimated with OPT-2. (i) Prediction error with OPT-2 when the motion field in (h) is used for image prediction. The smoothness energy E_s values of the motion fields are (d) 4309, (f) 4851, and (h) 1479. The PSNRs of the predicted images $\hat{I}_2$ in (e), (g), and (i) with respect to I_2 are 20, 20.4, and 21.53 dB, respectively.

cost, and we solve the OPT-2 problem without activating the smoothness term, i.e., $\alpha_1 = 0$. The estimated transformation field and the prediction error are shown in Fig. 6(f) and (g), respectively. We observe that the consistency term improves the quality of the motion field and the prediction quality. Finally, we highlight the benefit of enforcing smoothness constraint in our OPT-2 problem. The estimated transformation field with the OPT-2 problem, including the smoothness term, is shown in Fig. 6(h). By comparing the motion fields in Fig. 6(d) and (f), we see that the motion field in Fig. 6(h) is smoother and more coherent; this confirms the benefit of the smoothness term. Quantitatively, the smoothness energy E_s of the motion field shown in Fig. 6(h) is 1479, which is clearly smaller compared with the solutions given in Fig. 6(d) and (f) (i.e., 4309 and 4851, respectively). In addition, the smoothness term effectively improves the quality of the predicted image $\hat{I}_2$ since it gets closer to the original image I_2, as shown in Fig. 6(i).

C. Stereo Image Coding

We now study the performance of our distributed image representation algorithms in stereo imaging frameworks. We use two data sets, namely, *Plastic* and *Sawtooth*.[2] The images are downsampled to a resolution of $N_1 = 144$ and $N_2 = 176$ (original resolution of the data sets are 370×423 and 434×380, respectively). We carry out experiments using views 1 and 3 for the Plastic data set and views 1 and 5 for the Sawtooth data set. These data sets have been captured by a camera array where different viewpoints are uniformly arranged on a line. As this corresponds to translating the camera along one of the image coordinate axes, the disparity estimation problem becomes a 1-D search problem and the smoothness term in (10) is accordingly simplified. Viewpoint 1 is selected as the reference image I_1, and it is encoded such that the quality of $\hat{I}_1$ is approximately 33 dB. Matching pursuit is then performed on $\hat{I}_1$ with $K = 30$ atoms and $K = 60$ atoms for the Plastic and Sawtooth data sets, respectively. The measurements on the second image are generally quantized using a 2-bit quantizer. At the decoder, the search for the geometric transformations $\{F^k\}$ is carried out along the translational component t_x with window size $\delta t_x = 4$ pixels and no search is considered along the vertical direction, i.e., $\delta t_y = 0$. Unless explicitly stated, we use the data cost E_d given in (7) in the OPT-1 and OPT-2 problems.

We first study the accuracy of the estimated disparity information. In Fig. 7, we show the estimated disparity field $\mathbf{m}^h$ from 8870 quantized measurements (i.e., a measurement rate of 35%) for the Plastic data set. The ground truth $\mathbf{M}^h$ is given in Fig. 7(a). The transformation is estimated by solving OPT-1, and the resulting dense disparity field is illustrated in Fig. 7(b). In this particular experiment, parameter α_1 is selected such that the error in the disparity map is minimized. The disparity error (DE) is computed between the estimated disparity field $\mathbf{m}^h$ and the ground truth $\mathbf{M}^h$ as $\text{DE} = (1/N_1 \times N_2) \sum_{\mathbf{z}=(x,y)} \{|\mathbf{M}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z})| \geq 1\}$, where $N_1 \times N_2$ represents the pixel resolution of the image [25]. In Fig. 7(b), we observe that OPT-1 gives a good estimation of the disparity map; in particular, the disparity value is correctly estimated in the regions with texture or depth discontinuities. We could also observe that the estimation of the disparity field is, however, less precise in smooth regions as expected from feature-based methods. Fortunately, the wrong estimation of the disparity value corresponding to the smooth region in the images does not significantly affect the warped or predicted image quality [25]. Fig. 7(c) confirms such a distribution of the disparity estimation error where the white pixels denote an

[2]These image sets are available at http://vision.middlebury.edu/stereo/data/.

(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Fig. 7. Comparison of the estimated disparity fields with OPT-1 and OPT-2 for the Plastic data set. (a) Ground truth disparity field $\mathbf{M}^h$ between views 1 and 2. (b) Estimated disparity field with OPT-1. (c) Error in the disparity map with OPT-1 (DE $= 10.8\%$). (d) Estimated disparity field with OPT-2. (e) Error in the disparity map with OPT-2 (DE $= 4.1\%$). The disparity field is estimated using a measurement rate of 35% with a 2-bit quantization.

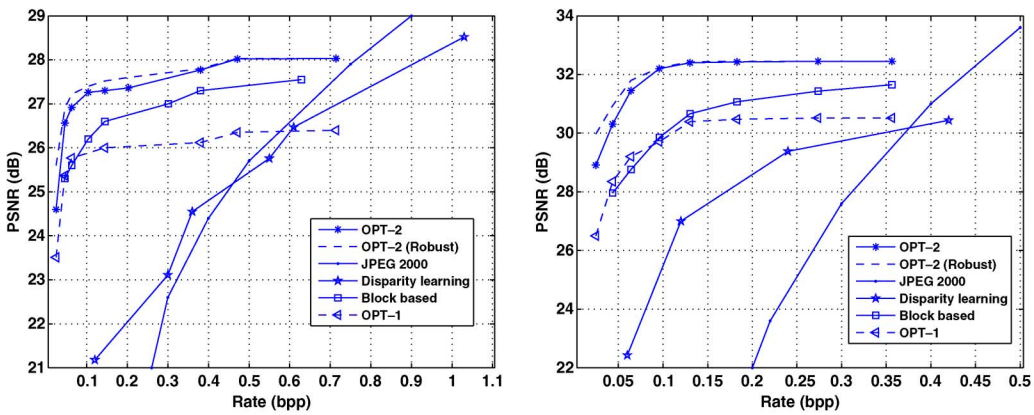

Fig. 8. Comparison of the RD performances between the proposed scheme, DSC scheme [6], block-based scheme [5], and independent coding solutions based on JPEG 2000 for (a) Sawtooth data set and (b) Plastic data set.

estimation error larger than one. We can see that the error in the disparity field is highly concentrated along the edges since crisp discontinuities cannot be accurately captured due to the scale and smoothness of the atoms in the chosen dictionary. The disparity information estimated by OPT-2 is presented in Fig. 7(d), and the corresponding error is shown in Fig. 7(e). In this case, the regularization constants α_1 and α_2 in the OPT-2 problem are selected such that the DE is minimized. We see that the addition of the consistency term E_t in the correlation estimation algorithm improves the performance.

We then study the RD performance of the proposed algorithms in the prediction of the image $\hat{I}_2$ in Fig. 8. We show the performance of the reconstruction by warping the reference image according to the correlation computed by OPT-1 and OPT-2. We then highlight the benefit of using the robust data term $\tilde{E}_d$ in the OPT-1 problem (denoted as *OPT-2 (Robust)*). We use the optimization toolbox based on CVX [40] in order to solve the optimization problem given in (8). We then compare the RD performance to a distributed coding solution (DSC) based on the low-density parity-check encoding of DCT coefficients, where the disparity field is estimated at the decoder using expected maximization (EM) principles [6] (denoted as *Disparity learning*). Then, in order to demonstrate the benefit of geometric dictionaries, we propose a scheme denoted as *block-based* that adaptively constructs the dictionary using blocks or patches in the reference image [5]. We construct a dictionary in the joint decoder from the reference image $\hat{I}_1$ segmented into 8×8 blocks. We then use the optimization scheme described in OPT-2 to select the best block from the adaptive dictionary. In order to have a fair comparison, we encode the reference image I_1 similarly for both schemes (i.e., *Disparity learning* and *block-based*) with a quality of 33 dB (see Section III) and the search window size is fixed to $\delta t_x = 4$ pixels along the horizontal direction. Finally, we also provide the performance of a standard JPEG 2000 independent encoding of image I_2. In Fig. 8, we first see that measurement consistency term E_t significantly improves the decoding quality as OPT-2 gives better performance than OPT-1. We further see that the OPT-2 problem with robust data cost improves the quality of the reconstructed image $\hat{I}_2$ by 0.5–1 dB at low bit rates. Then, the results confirm that the proposed algorithms unsurprisingly outperform independent coding based on JPEG 2000; this outlines the benefits of the use of correlation in the decoding of compressed correlated images. At high rate, the performance of the proposed algorithms, however, tends to saturate as our model mostly handles the geometry and the correlation between images, but it is not able to efficiently handle the fine details or texture in the scene due to the image decoding $\hat{I}_2$ based on warping. In Fig. 8, it is then clear that the reconstruction of image $\hat{I}_2$ based on OPT-1 and OPT-2 outperforms the DSC coding scheme based on EM principles due to the accurate correlation estimation. It is worth mentioning that the state-of-the-art DSC scheme based on disparity learning compensates also for the prediction error in addition to correlation estimation. Although this is the case, our scheme outperforms the DSC scheme due to an accurate disparity field estimation. Finally, the experimental results also

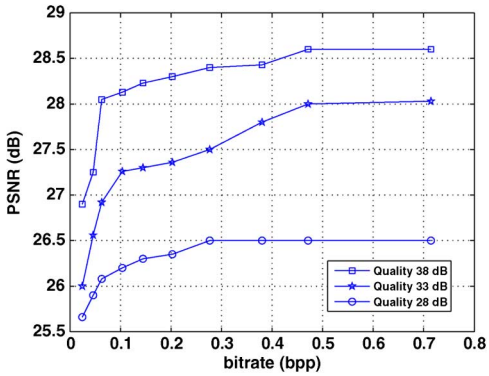

Fig. 9. RD performance with OPT-2 for decoding $\hat{I}_2$ (view 5) as a function of the quality of the reference image $\hat{I}_1$ (respectively 28, 33, and 38 dB) in the Sawtooth data set.

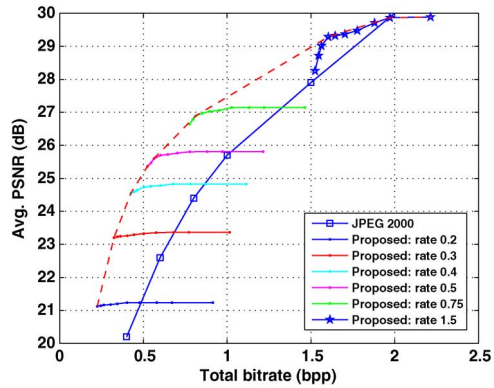

Fig. 10. Cumulative RD performance of views 1 and 5 for the Sawtooth data set. OPT-2 is used to predict the image $\hat{I}_2$ (view 5) using the image $\hat{I}_1$ (view 1) as the reference image. The image at view 5 is predicted with varying reference image bit rates 0.2, 0.3, 0.4, 0.5, 0.75, and 1.5 bpp.

show that our schemes outperform the scheme based on the block-based dictionary mainly because of the richer representation of the geometry and local transformations with the structured dictionaries.

We then study the influence of the quality of reference image $\hat{I}_1$ on the decoding performance of the second image $\hat{I}_2$. We use OPT-2 to decode $\hat{I}_2$ (viewpoint 5) by warping when the reference image is encoded at different qualities (i.e., different bit rates). Fig. 9 shows that the predicted image $\hat{I}_2$ quality improves with the quality of the reference image $\hat{I}_1$ as expected. While we have observed that the error in the disparity estimation is not dramatically reduced by improved reference quality, the warping stage permits to provide more details in the representation of $\hat{I}_2$ when the reference is of better quality. Now, we study the cumulative RD performance of views 1 and 5 for the Sawtooth data set, i.e., we include the bit rate and quality of the reference image I_1 (viewpoint 1) in addition to the rate and quality of image I_2 (viewpoint 5). Fig. 10 shows the joint RD performance at reference image bit rates 0.2, 0.3, 0.4, 0.5, 0.75, and 1.5 bpp. In our experiments, for a given reference image quality, we estimate the correlation model using OPT-2 (with 2-bit quantized measurements), and we compute the joint RD performance at that specific reference image bit rate. As shown before, the RD performance improves with increasing reference image quality. When we take the convex hull of the RD performances (which corresponds to implementing a proper rate allocation strategy), we outperform independent coding solutions based on JPEG 2000.

We now study the influence of the quantization bit rate on the RD performance of $\hat{I}_2$ with the OPT-2 optimization scheme. We compress the measurements y_2 using 2-, 4-, and 6-bit uniform quantizers. As expected, the quality of the correlation estimation degrades when the number of bits reduces, as shown in Fig. 11(a). However, it is largely compensated by the reduction in bit rate in the RD performance as confirmed by Fig. 11(b). This means that the proposed correlation estimation is relatively robust to quantization so that it is possible to attain good RD performance by drastic quantization of the measurements. Finally, we study the improvement offered by the robust data term $\tilde{E}_d$ [see (8)] in OPT-2 when the measurements have been compressed with a 2-bit uniform quantizer. In Fig. 11(a), it is clear that the proposed robust data term improves the performance due to efficient handling of noise in the quantized measurements.

D. Multiview Image Representation

We finally evaluate the performance of our multiview correlation estimation algorithms using five images from the *Tsukuba* data set (center, left, right, bottom, and top views) and five frames (frames 3–7) from the *Flower Garden* sequence [39]. These data sets are downsampled by a factor 2, and the resolution used in our experiments is 144 $\times$ 192 pixels and 120 $\times$ 180 pixels, respectively. In both data sets, the reference image I_1 (center view and frame 5, respectively) is encoded with a quality of approximately 33 dB. The measurements $y_j, \forall j \in \{2,3,4,5\}$ computed from the remaining four images are quantized using a 2-bit quantizer. We first compare our results to a stereo setup where the disparity information is estimated with the OPT-2 problem between the center and left images in Tsukuba data set. Fig. 12 compares the inverse depth error (sum of the labels with an error larger than one with respect to ground truth) between the multiview and stereo scenarios. In this particular experiment, parameters α_1 and α_2 are selected such that they minimize the error in the depth image with respect to the ground truth. It is clear from the plot that the depth error is small for a given measurement rate when all the views are available. It should be noted that the x-axis in Fig. 12 represents the measurement rate per view. Hence, the total number of measurements used in the multiview scenario is higher when compared with that in the stereo case. However, these experiments show that the proposed multiview scheme gives a better depth image when more images are available. Similar experimental findings have been observed for the Flower Garden sequence.

We then study the RD performance of the proposed multiview scheme in the decoding of four images (top, left, right, and bottom images in the Tsukuba data set and frames 3, 4, 6, and 7 in the Flower Garden sequence). The images are decoded by warping the reference image $\hat{I}_1$ using the estimated depth image. Fig. 13 compares the joint RD performance (for four images) of

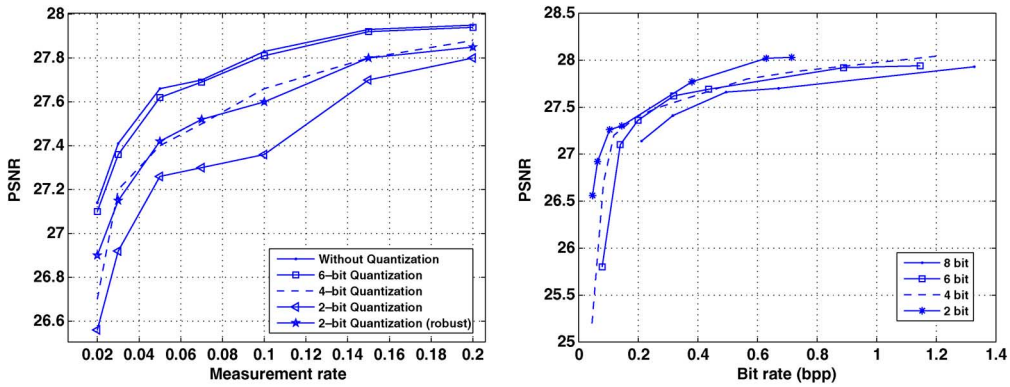

Fig. 11. Effect of measurement quantization on the quality of the image $\hat{I}_2$ decoded with OPT-2 scheme in the Sawtooth data set. The quality of the predicted image $\hat{I}_2$ is given in terms of (a) measurement rate and (b) bit rate. The benefit of using robust data cost is illustrated using a 2-bit uniform quantizer.

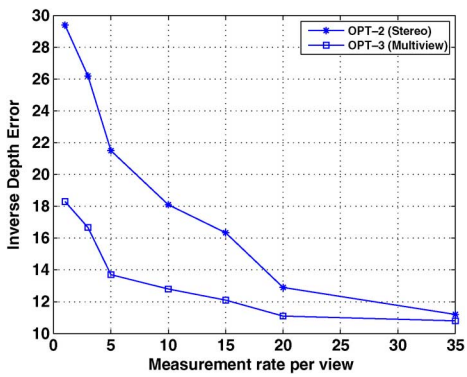

Fig. 12. Inverse depth error at various measurement rates of the Tsukuba multiview data set. OPT-2 and OPT-3 problems are used to estimate the depth in stereo and multiview scenarios, respectively. The measurements are quantized using a 2-bit quantizer.

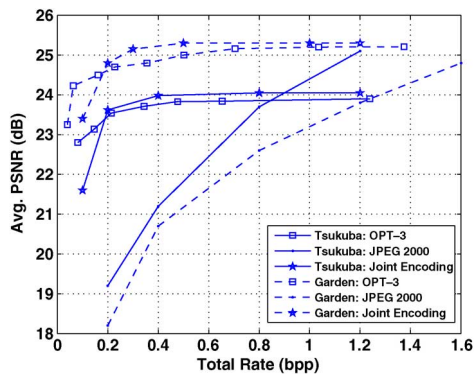

Fig. 13. Comparison of the joint RD performances between the proposed OPT-3 scheme, joint encoding scheme, and independent coding scheme based on JPEG 2000. The bit rate of the reference image I_1 is not included in the total bit budget.

our multiview scheme with respect to independent coding performance based on JPEG 2000. As expected, the proposed multiview scheme outperforms independent coding solutions based on JPEG 2000 as it benefits from the correlation between images. Furthermore, as observed in distributed stereo coding, the proposed multiview coding scheme saturates at high rates as the warping operator captures only the geometry and correlation between images but not the texture information.

Finally, we compare our results with a joint encoding approach where the depth image is estimated from the original images and transmitted to the joint decoder. At the decoder, the views are predicted from the reconstructed reference image $\hat{I}_1$ and the compressed depth image with the help of view prediction. The results are presented in Fig. 13 (denoted as *Joint Encoding*), where the bit rate is computed only on the depth image encoded using a JPEG 2000 coding solution. The main difference between the proposed and joint encoding frameworks is that the quantized linear measurements are transmitted for a depth estimation in the former scheme, whereas the depth information is directly transmitted in the latter scheme. Therefore, by comparing these two approaches, we can judge the accuracy of the estimated correlation model or equivalently the quality of the predicted view at a given bit rate. In Fig. 13, we see that at low bit rate <0.2, the proposed scheme estimates better structural information compared with the joint encoding scheme due to the geometry-based correlation representation. However, at rates above 0.2, we see that our scheme competes with joint coding solutions. This leads to the conclusion that the proposed scheme effectively estimates the depth information from the highly compressed quantized measurements. It should be noted that in the joint encoding framework, the depth images are estimated at a central encoder. In contrary to this, we estimate the depth images at the central decoder from the independently compressed visual information; this advantageously reduces the complexity at the encoder, which makes it attractive for distributed processing applications.

VIII. Conclusion

In this paper, we have presented a novel framework for the distributed representation of correlated images with quantized linear measurements, along with joint decoding algorithms that exploit the geometrical correlation among multiple images. We have proposed a regularized optimization problem in order to identify the geometrical transformations between compressed images, which results in smooth disparity or depth fields between a reference and one or more predicted image(s). We have proposed a low-complexity algorithm for the correlation estimation problem which offers an effective tradeoff between the

complexity and accuracy of the solution. In addition, we have proposed a new consistency criterion such that transformations are consistent with the compressed measurements in the predicted image. Experimental results demonstrate that the proposed methodology provides a good estimation of dense disparity/depth fields in different multiview image data sets. We also show that our geometry-based correlation model is more efficient than block-based correlation models. Finally, the consistent constraints prove to offer effective decoding quality such that the proposed algorithm outperforms JPEG 2000 and DSC schemes in terms of RD performance, even if the images are reconstructed by warping. This clearly positions our scheme as an effective solution for distributed image processing with low encoding complexity.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editor for their careful reviews and valuable suggestions that undoubtedly helped in improving the quality of this paper.

REFERENCES

[1] V. Thirumalai and P. Frossard, "Motion estimation from compressed linear measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 714–717.

[2] V. Thirumalai and P. Frossard, "Joint reconstruction of correlated images from compressed linear measurements," in *Proc. Eur. Signal Process. Conf.*, Aalborg, Denmark, 2010.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[5] J. P. Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," in *Proc. Picture Coding Symp.*, 2009, pp. 1–4.

[6] D. Varodayan, Y. C. Lin, A. Mavlankar, M. Flierl, and B. Girod, "Wyner–Ziv coding of stereo images with unsupervised learning of disparity," in *Proc. Picture Coding Symp.*, Lisbon, Portugal, 2007.

[7] D. Varodayan, D. Chen, M. Flierl, and B. Girod, "Wyner–Ziv coding of video with unsupervised motion vector learning," *EURASIP Signal Process., Image Commun.*, vol. 23, no. 5, pp. 369–378, Jun. 2008.

[8] E. J. Candes and J. Romberg, "Practical signal recovery from random projections," in *Proc. SPIE Comput. Imag.*, 2005, pp. 76–86.

[9] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.

[10] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[11] S. Mun and J. Fowler, "Block compressed sensing of images using directional transforms," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 3021–3024.

[12] L. Gan, T. T. Do, and T. D. Tran, "Fast compressive imaging using scrambled Hadamard ensemble," in *Proc. Eur. Signal Image Process. Conf.*, Lausanne, Switzerland, 2008.

[13] V. Stankovic, L. Stankovic, and S. Cheng, "Compressive video sampling," in *Proc. Eur. Signal Image Process. Conf.*, Lausanne, Switzerland, 2008.

[14] J. Y. Park and M. B. Wakin, "A multiscale framework for compressive sensing of video," in *Proc. Picture Coding Symp.*, 2009, pp. 1–4.

[15] N. Vaswani, "Kalman filtered compressed sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 893–896.

[16] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Asilomar Conf. Signal Syst. Comput.*, 2005, pp. 1537–1541.

[17] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Universal distributed sensing via random projections," in *Proc. Inf. Process. Sens. Netw.*, 2006, pp. 177–185.

[18] L. W. Kang and C. S. Lu, "Distributed compressive video sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 1169–1172.

[19] T. T. Do, Y. Chen, D. T. Nguyen, N. Nguyen, L. Gan, and T. D. Tran, "Distributed compressed video sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 1393–1396.

[20] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[21] M. Trocan, T. Maugey, J. E. Fowler, and B. Pesquet-Popescu, "Disparity-compensated compressed-sensing reconstruction for multiview images," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2010, pp. 1225–1229.

[22] M. Trocan, T. Maugey, E. W. Tramel, J. E. Fowler, and B. Pesquet-Popescu, "Multistage compressed-sensing reconstruction of multiview images," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2010, pp. 111–115.

[23] I. Tosic and P. Frossard, "Geometry based distributed scene representation with omnidirectional vision sensors," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1033–1046, Jul. 2008.

[24] O. D. Escoda, G. Monaci, R. M. Figueras, P. Vandergheynst, and M. Bierlaire, "Geometric video approximation using weighted matching pursuit," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1703–1716, Aug. 2009.

[25] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense stereo," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr.–Jun. 2002.

[26] S. Baker, S. Roth, D. Scharstein, M. Black, J. Lewis, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[27] P. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. Int. Conf. Inf. Sci. Syst.*, 2008, pp. 16–21.

[28] A. Schulz, L. Velho, and E. A. B. da Silva, "On the empirical rate–distortion performance of compressive sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2009, pp. 3049–3052.

[29] L. Jacques, D. K. Hammond, and M. J. Fadili, "Dequantizing compressed sensing: When oversampling and non-Gaussian constraints combine," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 559–571, Jan. 2011.

[30] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 149–152, Feb. 2010.

[31] A. Fletcher, S. Rangan, and V. Goyal, "On the rate–distortion performance of compressed sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. III-885–III-888.

[32] W. Dai, H. V. Pham, and O. Milenkovic, Distortion-Rate Functions for Quantized Compressive Sensing [Online]. Available: http://arxiv.org/abs/0901.0749 2009

[33] J. Sun and V. Goyal, "Optimal quantization of random measurements in compressed sensing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 6–10.

[34] G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[35] R. M. Figueras, P. Vandergheynst, and P. Frossard, "Low-rate and flexible image coding with redundant representations," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 726–739, Mar. 2006.

[36] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Jan. 2002.

[37] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

[38] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4685–4697, Dec. 2006.

[39] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. ECCV*, 2002, pp. 8–40.

[40] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, Version 1.21 [Online]. Available: http://cvxr.com/cvx Apr. 2011

Vijayaraghavan Thirumalai (S'10) received the M.Tech. degree in electronics and instrumentation engineering from the Indian Institute of Science, Bangalore, India, in 2005 and the M.S. and Ph.D. degrees in communication systems and electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2007 and 2012, respectively.

He is currently with the Signal Processing Laboratory (LTS4), EPFL. His research interests include image and video compression, distributed source coding, sparse approximations, 3-D and multiview coding, and compressed sensing.

Pascal Frossard (S'96–M'01–SM'04) received the M.S. and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1997 and 2000, respectively.

Between 2001 and 2003, he was a member of the Research Staff in the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he worked on media coding and streaming technologies. Since 2003, he has been a Faculty Member at EPFL, where he heads the Signal Processing Laboratory (LTS4). His research interests include image representation and coding, visual information analysis, distributed image processing and communications, and media streaming systems.

Dr. Frossard was the General Chair of IEEE ICME 2002 and Packet Video 2007. He was the Technical Program Chair of EUSIPCO 2008 and a member of the organizing or technical program committees of numerous conferences. He has been an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2004–present) and the IEEE TRANSACTIONS ON IMAGE PROCESSING (2010–present). He was also an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2006–2011. He is the Vice-Chair of the IEEE Image, Video and Multidimensional Signal Processing Technical Committee (2007–present). He is an Elected Member of the IEEE Visual Signal Processing and Communications Technical Committee (2006–present) and of the IEEE Multimedia Systems and Applications Technical Committee (2005–present). He has served as the Vice-Chair of the IEEE Multimedia Communications Technical Committee (2004–2006) and as a member of the IEEE Multimedia Signal Processing Technical Committee (2004–2007). He was the recipient of the Swiss NSF Professorship Award in 2003, the IBM Faculty Award in 2005, the IBM Exploratory Stream Analytics Innovation Award in 2008, and the IEEE Transactions on Multimedia Best Paper Award in 2011.