Learning Joint Intensity-Depth Sparse Representations

Ivana Tošić and Sarah Drewes

***Abstract*— This paper presents a method for learning overcomplete dictionaries of atoms composed of two modalities that describe a 3D scene: 1) image intensity and 2) scene depth. We propose a novel joint basis pursuit (JBP) algorithm that finds related sparse features in two modalities using conic programming and we integrate it into a two-step dictionary learning algorithm. The JBP differs from related convex algorithms because it finds joint sparsity models with different atoms and different coefficient values for intensity and depth. This is crucial for recovering generative models where the same sparse underlying causes (3D features) give rise to different signals (intensity and depth). We give a bound for recovery error of sparse coefficients obtained by JBP, and show numerically that JBP is superior to the group lasso algorithm. When applied to the Middlebury depth-intensity database, our learning algorithm converges to a set of related features, such as pairs of depth and intensity edges or image textures and depth slants. Finally, we show that JBP outperforms state of the art methods on depth inpainting for time-of-flight and Microsoft Kinect 3D data.**

***Index Terms*— Sparse approximations, dictionary learning, hybrid image-depth sensors.**

I. Introduction

HYBRID image-depth sensors have recently gained a lot of popularity in many vision applications. Time of flight cameras [1], [2] provide real-time depth maps at moderate spatial resolutions, aligned with the image data of the same scene. Microsoft Kinect [3] also provides real-time depth maps that can be registered with color data in order to provide 3D scene representation. Since captured images and depth data are caused by the presence of same objects in a 3D scene, they represent two modalities of the same phenomena and are thus correlated. This correlation can be advantageously used for denoising corrupted or inpainting missing information in captured depth maps. Such algorithms are of significant importance to technologies relying on image-depth sensors for 3D scene reconstruction or visualization [3], [4], where depth maps are usually noisy, unreliable or of poor spatial resolution.

Manuscript received May 5, 2013; revised October 26, 2013 and February 22, 2014; accepted March 3, 2014. Date of publication March 19, 2014; date of current version April 3, 2014. The work of I. Tošić was supported by the Swiss National Science Foundation through the Fellowship under Grant PA00P2-134159. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bulent Sankur.

I. Tošić was with the Helen Wills Neuroscience Institute, University of California at Berkeley, Berkeley, CA 94720 USA. She is now with Ricoh Innovations, Corporation, Menlo Park, CA 94025 USA (e-mail: ivana@ric.ricoh.com).

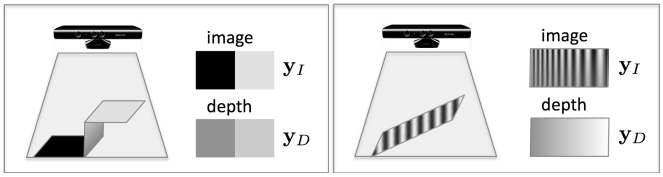
S. Drewes was with the Department of Industrial Engineering and Operations Research, University of California at Berkeley, Berkeley, CA 94720 USA. She is now with the MathWorks GmbH, Ismaning 85737, Germany (e-mail: sarah.drewes@mathworks.de).

Digital Object Identifier 10.1109/TIP.2014.2312645

Solving inverse problems such as denoising or inpainting usually involves using prior information about data. Sparse priors over coefficients in learned linear generative models have been recently applied to these problems with large success [5]–[7]. A similar approach has been proposed for learning sparse models of depth only, showing state-of-the-art performance in depth map denoising and offering a general tool for improving existing depth estimation algorithms [8]. However, learning sparse generative models for joint representation of depth and intensity images has not been addressed yet. Correlation between intensity and depth has been exploited for a long time in computer vision tasks such as depth from stereo [9]. Unlike most of these approaches that use hand designed priors, such as relation of depth and image smoothness, here we try to learn such features from the data. Learning such models from natural 3D data is of great importance for many applications involving 3D scene reconstruction, representation and compression.

This paper proposes a method for learning joint depth and intensity sparse generative models. Each of these two modalities is represented using overcomplete linear decompositions, resulting in two sets of coefficients. These two sets are coupled via a set of hidden variables, where each variable multiplies exactly one coefficient in each modality. Consequently, imposing a sparse prior on this set of coupling variables results in a common sparse support for intensity and depth. Each of these hidden variables can be interpreted as a presence of a depth-intensity feature pair arising from the same underlying cause in a 3D scene. To infer these hidden variables under a sparse prior, we propose a convex, second order cone program named Joint Basis Pursuit (JBP). Compared to Group Lasso (GL) [10], [11], which is commonly used for coupling sparse variables, JBP gives significantly smaller coefficient recovery error. In addition, we bound theoretically this error by exploiting the restricted isometry property (RIP) [12] of the model. Finally, we propose an intensity-depth dictionary learning algorithm based on the new model and JBP. We show its superiority to GL in model recovery experiments using synthetic data, as well as in inpainting experiments using real time-of-flight and Kinect 3D data.

We first explain in Section II why existing models are not sufficient for intensity-depth representation. Section III introduces the proposed intensity-depth generative model. Inference of its hidden variables is achieved via the new JBP algorithm presented in Section IV, while learning of model parameters is explained in Section V. Section VI gives relations of the



(a) (b)

Fig. 1. Examples of two typical image-depth features in 3D scenes. (a) Example 1: 3D edge, (b) Example 2: slanted texture.

proposed JBP to prior art. Experimental results are presented in Section VII.

II. Why Aren't Existing Models Enough?

To model the joint sparsity in intensity and depth, one might think that simple, existing models would suffice. For example, an intuitive approach would be to simply merge depth and image pixels into one array of pixels. If we denote the vectorized form of the intensity image as $\mathbf{y}^I$ and depth image as $\mathbf{y}^D$, both of length L, this "merged" model can be written as:

$$\begin{bmatrix} \mathbf{y}^I \\ \mathbf{y}^D \end{bmatrix}_{2L\times 1} = \begin{bmatrix} \mathbf{\Phi}^I \\ \mathbf{\Phi}^D \end{bmatrix}_{2L\times N} \cdot \mathbf{c}$$

where intensity and depth are assumed to be sparse in dictionaries $\mathbf{\Phi}^I$, resp. $\mathbf{\Phi}^D$ both of size $L \times N$. The sparse vector $\mathbf{c}$ of length N would then couple sparse patterns in intensity and depth, i.e., couple intensity and depth atoms in pairs. However, since the vector of coefficients $\mathbf{c}$ is common, intensity and depth atoms within a pair will be multiplied by the same value. Let us now look at two simple synthetic examples of 3D scenes whose intensity and depth images are shown in Fig. 1. The first example is a 3D edge and the second is a textured pattern on a slanted surface. These are two common intensity-depth features in real scenes. Since it has the flexibility of using different atoms for intensity and depth, the merged model will be able to represent both features. However, since intensity and depth coefficients have equal values, variability in magnitude between intensity and depth would have to be represented by different atom pairs, leading to a combinatorial explosion in dictionary size.

Another model that has been widely used in literature for representing correlated signals is the joint sparsity model, where signals share the same sparse support in $\mathbf{\Phi}$ of size $L\times N$, but with different coefficients:

$$\begin{bmatrix} \mathbf{y}^I & \mathbf{y}^D \end{bmatrix}_{L\times 2} = \mathbf{\Phi}_{L\times N} \cdot [\mathbf{a} \;\; \mathbf{b}]_{N\times 2}\,, \quad supp(\mathbf{a}) = supp(\mathbf{b}),$$

where *supp* denotes the sparse support. Therefore, the property of this model is that signals are represented using the same atoms multiplied by different coefficients. Obviously, the joint sparsity model would be able to represent the intensity-depth edge in Fig. 1 using a piecewise constant atom and different coefficients for intensity and depth. However, in the slanted texture example, because the depth image is linear and the intensity is a chirp, no atom can model both. The joint sparsity model would then have to decouple these two features in different atoms, which is suboptimal for representing slanted textures.

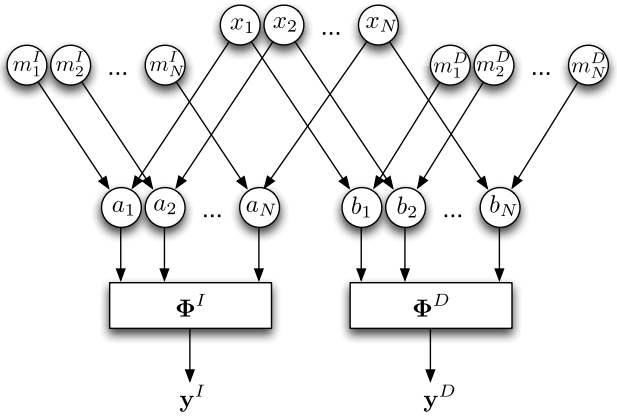


Fig. 2. Graphical representation of the proposed intensity-depth generative model.

It becomes clear that we need a model that allows joint representation with different atoms and different coefficients, but with a common sparse support (the pattern of non-zero coefficients needs to be the same). We introduce such a model in the next section.

III. Intensity-Depth Generative Model

Let us first set the notation rules. Throughout the rest of the paper, vectors are denoted with bold lower case letters, matrices with bold upper case letters. The ℓ_p-norm is denoted as $\|\cdot\|_p$ (for any $p \in \mathbb{R}_+$) and Frobenius norm as $\|\cdot\|_F$. Letters I, D in superscripts refer to intensity and depth, respectively. Sets are represented with calligraphic fonts and $|\cdot|$ denotes the cardinality of a set. Column-wise and row-wise concatenations of vectors $\mathbf{a}$ and $\mathbf{b}$ are denoted as $[\mathbf{a}\;\;\mathbf{b}]$ and $[\mathbf{a};\mathbf{b}]$, respectively.

Graphical representation of the proposed joint depth-intensity generative model is shown in Fig. 2. Intensity image $\mathbf{y}^I$ and depth image $\mathbf{y}^D$ (in vectorized forms) are assumed to be sparse in dictionaries $\mathbf{\Phi}^I$, resp. $\mathbf{\Phi}^D$, i.e., they are represented as linear combinations of dictionary atoms $\{\boldsymbol{\phi}_i^I\}_{i\in\mathcal{I}}$ and $\{\boldsymbol{\phi}_i^D\}_{i\in\mathcal{I}}$, resp. :

$$\begin{aligned} \mathbf{y}^I &= \mathbf{\Phi}^I\mathbf{a} + \boldsymbol{\eta}^I = \sum_{i\in\mathcal{I}_0} \boldsymbol{\phi}_i^I a_i + \boldsymbol{\eta}^I \\ \mathbf{y}^D &= \mathbf{\Phi}^D\mathbf{b} + \boldsymbol{\eta}^D = \sum_{i\in\mathcal{I}_0} \boldsymbol{\phi}_i^D b_i + \boldsymbol{\eta}^D, \end{aligned} \tag{1}$$

where vectors $\mathbf{a}$ and $\mathbf{b}$ have a small number of non-zero elements and $\boldsymbol{\eta}^I$ and $\boldsymbol{\eta}^D$ represent noise vectors. $\mathcal{I}_0$ is the set of indexes identifying the columns (i.e., atoms) of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ that participate in sparse representations of $\mathbf{y}^I$ and $\mathbf{y}^D$. Its cardinality is much smaller than the dictionary size, hence $|\mathcal{I}_0| \ll |\mathcal{I}|$, where $\mathcal{I} = \{1, 2, \ldots, N\}$ denotes the index set of all atoms. This means that each image can be represented as a combination of few, representative features described by atoms, modulated by their respective coefficients. Because depth and intensity features correspond to two modalities arising from the same 3D features, we model the coupling between coefficients a_i and b_i through latent variables x_i as:

$$a_i = m_i^I x_i; \qquad b_i = m_i^D x_i, \quad \forall i \in \mathcal{I}, \tag{2}$$

where the variables m_i^I, m_i^D represent the magnitudes of the sparse coefficients and x_i represent the activity of these coefficients. Ideally, these variables should be binary, 0 representing the absence and 1 representing the presence of a depth-intensity feature pair. In that case $\sum_i x_i$ counts the number of non-zero such pairs. However, inference of binary values represents a combinatorial optimization problem of high complexity that depends on dictionary properties and the permission of noise, cf. [13]. We relax the problem by allowing x_i to attain continuous values between 0 and 1, which has been proven to provide a very good approximation in a similar context, cf., e.g., [14] and [15]. An important thing to note here is that there is a one-to-one correspondence between intensity and depth atoms, i.e., $\boldsymbol{\phi}_i^I$ and $\boldsymbol{\phi}_i^D$ form pairs for all $i = 1, \ldots, N$. This correspondence is also visible on the graph in Fig. 2 where each x_i is connected to exactly two nodes: a_i and b_i.

By assuming that the vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)^\mathsf{T}$ is sparse, we assume that $\mathbf{y}^I$ and $\mathbf{y}^D$ are described by a small number of feature pairs $(\boldsymbol{\phi}_i^I, \boldsymbol{\phi}_i^D)$ that are either prominent in both modalities (both m_i^I and m_i^D are significant) or in only one modality (either m_i^I or m_i^D is significant). In these cases x_i is non-zero, which leads to non-zero values for either a_i or b_i, or both. If x_i is zero, both a_i and b_i are also zero. Hence, the sparsity assumption on $\mathbf{x}$ enforces a compact description of both modalities by using simultaneously active coefficients. In addition, when such pairs cannot approximate both images, the model also allows only one coefficient within a pair to be non-zero. Therefore, the model represents intensity and depth using a small set of joint features and a small set of independent features. The main challenge is to simultaneously infer the latent variables $\mathbf{x}$, $\mathbf{m}^I = (m_1^I, m_2^I, \ldots, m_N^I)^\mathsf{T}$ and $\mathbf{m}^D = (m_1^D, m_2^D, \ldots, m_N^D)^\mathsf{T}$ under the sparsity assumption on $\mathbf{x}$. In the next section we propose a convex algorithm that solves this problem.

IV. Joint Basis Pursuit

Let us re-write the intensity-depth generative model, including all unknown variables, in matrix notation as:

$$\begin{bmatrix} \mathbf{y}^I \\ \mathbf{y}^D \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}^I & 0 \\ 0 & \boldsymbol{\Phi}^D \end{bmatrix} \cdot \begin{bmatrix} \mathbf{M}^I \\ \mathbf{M}^D \end{bmatrix} \cdot \mathbf{x} + \begin{bmatrix} \boldsymbol{\eta}^I \\ \boldsymbol{\eta}^D \end{bmatrix},$$

where $\mathbf{M}^I = \mathrm{diag}(m_1^I, m_2^I, \ldots, m_N^I)$ and $\mathbf{M}^D = \mathrm{diag}(m_1^D, m_2^D, \ldots, m_N^D)$. Suppose first that we know dictionaries $\boldsymbol{\Phi}^I$ and $\boldsymbol{\Phi}^D$ and we want to find joint sparse representations of intensity and depth, i.e., to solve for variables $\mathbf{x}, \mathbf{m}^I, \mathbf{m}^D$, under a Gaussian noise assumption (i.e., assuming a quadratic representation error). To do this, we formulate the following optimization problem:

$$\begin{aligned} \text{OPT1}: \ & \text{solve for } \mathbf{x}, \mathbf{m}^I, \mathbf{m}^D && (3)\\ & \min \sum_i x_i, \quad \text{where } x_i \in [0,1],\ i = 1, \ldots, N \\ \text{subject to:} \ & \|\mathbf{y}^I - \boldsymbol{\Phi}^I \mathbf{M}^I \mathbf{x}\|_2^2 \le (\epsilon^I)^2 && (4)\\ & \|\mathbf{y}^D - \boldsymbol{\Phi}^D \mathbf{M}^D \mathbf{x}\|_2^2 \le (\epsilon^D)^2 && (5)\\ & |m_i^I| \le U^I && (6)\\ & |m_i^D| \le U^D && (7) \end{aligned}$$

where ϵ^I, ϵ^D are allowed approximation errors and U^I and U^D are upper bounds on the magnitudes $\mathbf{m}^I$ and $\mathbf{m}^D$. In practice, the values of these upper bounds can be chosen conservatively as high finite values. By minimizing the sum of coupling variables $x_i \in [0, 1]$, OPT1 minimizes the ℓ_1 norm of the vector $\mathbf{x}$ and thus imposes sparsity on $\mathbf{x}$. This means that OPT1 looks for a solution with a small number of non-zero coupling variables x_i that in turn activate a small number of coefficient pairs (a_i, b_i) through (2). This optimization problem is hard to solve using the above formulation, since the first two constraints are non-convex due to the terms $\mathbf{M}^I\mathbf{x}$ and $\mathbf{M}^D\mathbf{x}$ which are bilinear in the variables $\mathbf{x}$, $\mathbf{m}^I$ and $\mathbf{m}^D$. To overcome this issue, we transform it into an equivalent problem by introducing the change of variables given by Eqs. (2) deriving:

$$\begin{aligned} \text{OPT2}: \ & \text{solve for } \mathbf{x}, \mathbf{a}, \mathbf{b} && (8)\\ & \min \sum_i x_i, \quad \text{where } x_i \in [0,1],\ i = 1, \ldots, N \\ \text{subject to:} \ & \|\mathbf{y}^I - \boldsymbol{\Phi}^I \mathbf{a}\|_2^2 \le (\epsilon^I)^2 && (9)\\ & \|\mathbf{y}^D - \boldsymbol{\Phi}^D \mathbf{b}\|_2^2 \le (\epsilon^D)^2 && (10)\\ & |a_i| \le U^I x_i && (11)\\ & |b_i| \le U^D x_i, && (12) \end{aligned}$$

which is a convex optimization problem with linear and quadratic constraints that can be solved efficiently, i.e., in polynomial time, using log-barrier algorithms, cf. [16] and [17]. A variety of free and commercial software packages are available like IBM ILOG CPLEX [18], that we use in our experiments.

The problems (OPT1) and (OPT2) are equivalent using the variable transformation in Eqs. (2) in the following sense:

Lemma 1. *For any optimal solution* $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ *of (OPT2),* $\mathbf{x}^*$ *is also an optimal solution to (OPT1) with corresponding matrices* $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ *according to* (2)*. Also, any optimal solution* $(\mathbf{x}^*, (\mathbf{M^I})^*, (\mathbf{M^D})^*)$ *of (OPT1) defines an optimal solution* $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ *to (OPT2)* .

Proof: For any $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ and corresponding $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ that satisfy Eqs. (2), conditions (9) and (10) are equivalent to (4) and (5) by definition. Moreover, since $\mathbf{x}^*$ is nonnegative, conditions (11) and (12) are equivalent to (6) and (7). Hence, any $\mathbf{x}^*$ that is optimal for (OPT2) with corresponding $(\mathbf{a}^*, \mathbf{b}^*)$ is optimal for (OPT1) with corresponding $(\mathbf{M^I})^*$, $(\mathbf{M^D})^*$ and vice versa. ∎

Thus, Lemma 1 states that each optimal solution of (OPT1) induces an optimal solution of (OPT2) and vice versa. An immediate consequence of the form of the objective function and constraints in (OPT2) is that $\mathbf{x}^*$ is chosen such that (11) and (12) are both feasible and at least one of them is active. Formally, this is stated by the following lemma.

Lemma 2. *For any optimal solution* $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ *of (OPT2), at least one of the constraints* (11) *and* (12) *is active for each component i, hence we have*

$$x_i^* = \max\{\frac{|a_i^*|}{U^I}, \frac{|b_i^*|}{U^D}\}, \qquad \forall i = 1, \ldots, N. \tag{13}$$

Proof: Otherwise it would be a contradiction to the optimality of $\mathbf{x}^*$. ∎

In the following, we refer to the optimization problem (OPT2) as Joint Basis Pursuit (JBP), where $\mathbf{x}$ is the vector of joint (coupling) variables in the signal model. It is important to know the theoretical bounds on the norm of the difference between the solution $(\mathbf{a}^*, \mathbf{b}^*)$ found by JBP and the true coefficients $(\mathbf{a}, \mathbf{b})$ of the model (1).

Based on the non-coupled case that is treated in [13], we develop bounds on the difference of the optimal solution of (OPT2) and a sparse signal to be recovered. For this purpose, we assume that the matrix

$$\mathbf{A} := \begin{bmatrix} \mathbf{\Phi}^I & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}^D \end{bmatrix} \tag{14}$$

satisfies the S-restricted isometry property with a constant δ_S. This property of a linear system is defined as follows. Denote $\mathbf{A}_{\mathcal{T}}$, $\mathcal{T} \subset 1, \ldots, n$ as the $n \times |\mathcal{T}|$ submatrix obtained by extracting the columns of $\mathbf{A}$ corresponding to the indices in set $\mathcal{T}$. The S-restricted isometry constant δ_S is then defined as:

Definition 1. *[12] The S-restricted isometry constant δ_S of $\mathbf{A}$ is the smallest quantity such that*

$$(1 - \delta_S)\|\mathbf{s}\|_2^2 \leq \|\mathbf{A}_{\mathcal{T}}\mathbf{s}\|_2^2 \leq (1 + \delta_S)\|\mathbf{s}\|_2^2 \tag{15}$$

for all subsets $\mathcal{T}$ with $|\mathcal{T}| \leq S$ and coefficient sequences (s_j), $j \in \mathcal{T}$.

When $\delta_S << 1$, this property requires that every set of columns with cardinality less than S approximately behaves like an orthonormal system. It can thus be related to the maximal value of the inner product between any two columns in the matrix $\mathbf{A}$, usually called the coherence of the dictionary:

$$\mu = \max_{i,j \neq i} |\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle|, \tag{16}$$

where $\boldsymbol{\phi}_i$ and , $\boldsymbol{\phi}_j$ are two different unit-norm atoms in the dictionary (i.e., two columns of $\mathbf{A}$) and $\langle \cdot \rangle$ denotes the inner product. With this definition, it can be easily shown that $\delta_S = \mu(|\mathcal{T}| - 1)$ satisfies the RIP inequality (15).

Before we present the bound on the coefficient recovery error of JBP, let us first define some prerequisites. Assume we are given a pair of signals $(\mathbf{y}^I, \mathbf{y}^D)$ as in Eq. (1), with sparse coefficients $(\mathbf{a}^0, \mathbf{b}^0)$, which satisfy constraints (9) and (10). Let $\mathcal{T}_0$ be the support of $\mathbf{x}^0$ which is at the same time the support of at least $\mathbf{a}^0$ or $\mathbf{b}^0$ and contains the support of the other one or it coincides with the support of both. Without loss of generality, let us assume that

$$\|\mathbf{y}^I\|_2 = \|\mathbf{y}^D\|_2 =: f_0, \tag{17}$$

which can be easily obtained by normalization. Assume also that the components of $\mathbf{a}^0$ and $\mathbf{b}^0$ satisfy the bound constraints[1]

$$|a_i^0| \leq f_0, \quad |b_i^0| \leq f_0, \qquad \forall i = 1, \ldots, N, \tag{18}$$

i.e., in the remainder of the paper we assume the same bounds on a_i and b_i: $U^I = U^D = U = f_0$. It is also useful in practice to select the approximation error ϵ in terms of the fraction of the total signal energy, so we denote $\epsilon = \eta f_0$, where $0 \leq \eta < 1$[2].

Let further α_i denote the scale between the smaller and larger coefficient for each index i within the sparse support set $\mathcal{T}_0$, i.e.:

$$\alpha_i = \min\{\frac{|a_i^0|}{|b_i^0|}, \frac{|b_i^0|}{|a_i^0|}\}, \qquad \forall i \in \mathcal{T}_0, \tag{19}$$

and let γ denote:

$$\gamma = 1 - \min_{i \in \mathcal{T}_0} \alpha_i. \tag{20}$$

Parameter γ describes the level of similarity between sparse coefficients in the two signals, which is decreasing with higher similarity.[3] In the trivial case when $a_i^0 = b_i^0$, $\forall i \in \mathcal{T}_0$ we have that $\gamma = 0$. In all other cases $\gamma \leq 1$.

Let further $\mathbf{x}^0$ denote an auxiliary vector that satisfies

$$\max\{|a_i^0|, |b_i^0|\} = U x_i^0, \qquad \forall i \in \mathcal{T}_0$$

namely $(\mathbf{x}^0, \mathbf{a}^0, \mathbf{b}^0)$ is a feasible solution to (OPT2), where $\mathbf{x}^0$ is chosen such that (11) and (12) are both feasible and (at least) one of them is active.

Finally, let $(\mathbf{x}^*, \mathbf{a}^*, \mathbf{b}^*)$ be an optimal solution to (OPT2). Then we have the following worst case bound on the distance of these.

Theorem 1. *Let $(\mathbf{a}^0, \mathbf{b}^0)$ and $(\mathbf{a}^*, \mathbf{b}^*)$ as defined above and choose $U = f_0$ with f_0 from (17) and $\epsilon^I = \epsilon^D = \eta f_0$, where $0 \leq \eta < 1$. Then*

$$\|[\mathbf{a}^0; \mathbf{b}^0] - [\mathbf{a}^*; \mathbf{b}^*]\|_2^2 \leq \left[\frac{|\mathcal{T}_0|}{M}(C + \gamma\sqrt{|\mathcal{T}_0|})^2 + C^2\right] f_0^2 \tag{21}$$

holds for a constant C that depends on the signal model parameter γ, the sparse support size $|\mathcal{T}_0|$ and the approximation parameter η, and where the M-restricted isometry property is satisfied for the linear system, cf. Def. 1. In particular, we have:

$$C = \frac{4\eta\sqrt{M} + \gamma\,|\mathcal{T}_0|\sqrt{1 + \delta_M}}{\sqrt{M(1 - \delta_{M+|\mathcal{T}_0|})} - \sqrt{|\mathcal{T}_0|(1 + \delta_M)}}. \tag{22}$$

The proof of this Theorem is given in Appendix .

V. Intensity-Depth Dictionary Learning

In the previous section we have shown how to find sparse coefficients in the joint depth-intensity generative model, assuming that the model parameters, i.e., dictionaries $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ are given. Since in general we do not have those parameters, we propose to learn them from a large database of intensity-depth image examples. Dictionary learning for sparse approximation has been a topic of intensive research in the last couple of years. Almost all existing algorithms are based on Expectation-Maximization, i.e., they are iterative algorithms that consist of two steps: 1) inference of sparse coefficients for a large set of signal examples while keeping the dictionary

[1] Although the assumption in Eq. (18) does not hold in general, in practical applications using learned dictionaries we found that it is always satisfied. However, if one wants to use a bound that is surely satisfied, one should choose $U = f_0/\sigma_{min}$, where σ_{min} is the smallest of all singular values of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$.

[2] One can chose η to be different for image and depth in the case where image and noise statistics differ due to the properties of an acquisition device.

[3] Note that γ does not impose any further coupling between coefficients $\mathbf{a}$ and $\mathbf{b}$, it only quantifies the similarity of their magnitudes.

parameters fixed, and 2) dictionary optimization to minimize the reconstruction error while keeping the coefficients fixed. We follow the same approach here, using JBP in the first step, conjugate gradient in the second step and then iterating these two steps until convergence. Once JBP in iteration k finds the sparse coefficients $\mathbf{a}^{(k)}, \mathbf{b}^{(k)}$ and the coupling variables $\mathbf{x}^{(k)}$, optimization of $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ becomes decoupled. Therefore, in the learning step we use conjugate gradient to independently optimize the following objectives:

$$(\mathbf{\Phi}^I)^{(k)} = \min_{\mathbf{\Phi}^I} \|\mathbf{Y}^I - \mathbf{\Phi}^I \mathbf{P}^{(k)}\|_F^2 + \rho \|\mathbf{\Phi}^I\|_F \tag{23}$$

$$(\mathbf{\Phi}^D)^{(k)} = \min_{\mathbf{\Phi}^D} \|\mathbf{Y}^D - \mathbf{\Phi}^D \mathbf{Q}^{(k)}\|_F^2 + \rho \|\mathbf{\Phi}^D\|_F, \tag{24}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{Y}^I$, $\mathbf{Y}^D$, $\mathbf{P}^{(k)}$ and $\mathbf{Q}^{(k)}$ are matrices whose columns are $\mathbf{y}_j^I$, $\mathbf{y}_j^D$, $\mathbf{a}_j^{(k)}$ and $\mathbf{b}_j^{(k)}$ respectively, and $j = 1, \ldots, J$ indexes the signal examples from a given database. In addition to the reconstruction error, we have added a normalization constraint on the dictionaries, scaled by a small parameter ρ, in order to control the dictionary norms as usually done in dictionary learning. After the learning step in iteration k, we solve again JBP in iteration $k+1$ by using dictionaries $(\mathbf{\Phi}^I)^{(k)}$ and $(\mathbf{\Phi}^D)^{(k)}$ obtained in iteration k. Before showing the performance of the proposed learning algorithm, we review prior art that we will use for experimental comparisons in Section VII.

VI. Relation to Prior Art

To the best of our knowledge, there has not been any work that addresses the problem of learning joint intensity-depth sparse representations. Therefore, we overview prior work that focuses on sparse approximation algorithms that bear similarities to JBP, i.e., algorithms that find sparse approximations of two signals sharing a common sparse support. Such algorithms can be grouped into two categories with respect to the signal model they address: a) simultaneous sparse approximation algorithms, and b) group sparse approximation algorithms. We further discuss how these relate to JBP.

Simultaneous sparse approximation algorithms recover a set of jointly sparse signals modeled as $\mathbf{y}^s = \mathbf{\Phi}\mathbf{x}^s + \boldsymbol{\epsilon}^s = \sum_{i \in \mathcal{I}} \boldsymbol{\phi}_i x_i^s + \boldsymbol{\epsilon}^s$, $s = 1, \ldots, S$, where S is the number of signals $\mathbf{y}^s$, $\mathbf{\Phi}$ is the dictionary matrix and $\boldsymbol{\epsilon}^s$ is a noise vector for signal $\mathbf{y}^s$. Vectors of sparse coefficients $\mathbf{x}^s$ share the same sparsity support set $\mathcal{I}$, i.e., they have non-zero entries at the same positions. For the case of two signals, for example image intensity and depth, this model is a noisy version of the second model discussed in Sec. II. One of the earliest algorithms in this group is the Simultaneous Variable Selection (SVS) algorithm introduced by Turlach et. al. [19]. SVS selects a common subset of atoms for a set of signals by minimizing the representation error while constraining the ℓ_1-norm of the maximum absolute values of coefficients across signals. Formally, SVS solves the following problem:

$$\text{(SVS)}: \quad \min \frac{1}{2} \sum_{s=1}^{S} \|\mathbf{y}^s - \Phi \mathbf{x}^s\|_2^2, \tag{25}$$

$$\text{subject to:} \quad \sum_i \max\{|x_i^1|, \ldots, |x_i^S|\} \leq \tau, \tag{26}$$

where τ is given. Let $\mathbf{X}$ denote the matrix with $\mathbf{x}^s$, $s = 1, \ldots, S$ as columns. We can see that the left hand side of the constraint in SVS is obtained by applying the ℓ_∞-norm to rows (to find the largest coefficients for all explanatory variables), followed by applying the ℓ_1-norm to the obtained vector in order to promote sparsity of the support. We denote this norm as $\|\mathbf{X}\|_{\infty,1}$. Versions of the same problem for the unconstrained case and the error-constrained case have been studied by Tropp [20].

To see the relation of SVS to JBP, we use Lemma 2, which allows us to formulate the JBP for $U^I = U^D$ as:

$$\min: \quad t \tag{27}$$

$$\text{subject to:} \quad \|\mathbf{y}^D - \mathbf{\Phi}^D \mathbf{a}\|_2^2 \leq \epsilon^2 \tag{28}$$

$$\|\mathbf{y}^I - \mathbf{\Phi}^I \mathbf{b}\|_2^2 \leq \epsilon^2 \tag{29}$$

$$\sum_i \max\{|a_i|, |b_i|\} \leq t. \tag{30}$$

Therefore, JBP operates on the same $\ell_{\infty,1}$-norm of the coefficient matrix as SVS. However, in contrast to SVS, JBP minimizes the number of non-zero elements in both $\mathbf{a}$ and $\mathbf{b}$ by minimizing $\|[\mathbf{a}\ \mathbf{b}]\|_{\infty,1}$ and constraining the approximation error induced by the coefficients. A much more important difference of our work and [19] is that we allow for different sets of atoms for intensity and depth. Thus, in JBP, each signal can be represented using a different dictionary, but with coefficient vectors that share the same positions of non-zero entries. This makes JBP applicable to intensity-depth learning, in contrast to SVS. Finally, we remark here that choosing the objective function as we did allows for a smooth convex representation of the last constraint (30).

Group sparse approximation algorithms recover a signal modeled as $\mathbf{y} = \sum_i \mathbf{H}_i \mathbf{x}_i + \boldsymbol{\epsilon}$, where $\mathbf{H}_i$ is a submatrix of a big dictionary matrix $\mathbf{H}$. This model is useful for signals whose sparse support has a group structure, namely when groups of coefficients are either all non-zero or all zero. The first algorithm proposed for group sparse approximation was a generalization of Lasso, developed by Bakin [10], [11], and later studied by other authors (e.g. Yuan and Lin [21]). Group Lasso (GL) refers to the following optimization problem:

$$\text{(GL)}: \min \|\mathbf{y} - \sum_i \mathbf{H}_i \mathbf{x}_i\|_2^2 + \lambda \sum_i \|\mathbf{x}_i\|_p. \tag{31}$$

The most studied variant of GL is for $p = 2$, because it leads to a convex optimization problem with efficient implementations. The group sparsity model can be used to represent intensity-depth signals by considering pairs $(a_i, b_i), i = 1, \ldots, N$ as groups. In this case, GL with $p = 2$ becomes:

$$\text{(GL-ID)}: \quad \min(\|\mathbf{y}^I - \sum_i \boldsymbol{\phi}_i^I a_i\|_2^2 \tag{32}$$

$$+ \|\mathbf{y}^D - \sum_i \boldsymbol{\phi}_i^D b_i\|_2^2 + \lambda \sum_i \sqrt{a_i^2 + b_i^2}). \tag{33}$$

The drawback of GL with $p = 2$ is that the square norm averages the coefficients in two modalities and does not distinguish between pairs with a different balance of coefficients. In other words, it has rotational symmetry in the space of coefficient pairs. This means that if there is a particular structure in the

distribution of coefficient pairs (for example governed by the ratio between intensity and depth) GL might not be able to recover that structure. Choosing $p = \infty$ avoids this problem and allows selection of pairs that might have asymmetric joint distributions. In that case the regularizer penalizes the norm $\|[\mathbf{a}\ \mathbf{b}]\|_{\infty,1}$. Rather than solving the unconstrained problem of group lasso with $p = \infty$ and a non-smooth objective, JBP reaches a similar goal by solving a constrained convex optimization problem with smooth constraints. It also eliminates the need for tuning the Lagrange multiplier.

We should also mention here block sparse models, which are a generalization of group sparse models where a set of signals shares the same group structure and where that structure might not be known apriori [22]. Example application of these models is representation of face images. In the case of image-depth modeling, the group structure is known and given by the bimodal structure of the data, thus the modeling reduces to group sparsity. Nevertheless, one can envisage in the future application of block sparse models on top of proposed bimodal image-depth representation for modeling signals such as face image-depth data.

VII. Experimental Results

We have performed two sets of experiments in order to evaluate the proposed JBP and dictionary learning based on JBP. The first set of experiments uses simulated random data, with the goal to determine the model recovery performance of JBP when the ground truth signal models are given. In the second set, we apply JBP and dictionary learning on real depth-intensity data and show its performance on a depth inpainting task. In both cases, JBP has been compared to Group Lasso (GL). For the depth inpainting task, we also compare JBP to inpainting using total variation (TV) [23] and using learned depth dictionary [8].

A. Model Recovery

To evaluate the performance of JBP, we have generated a set of pairs of signals of size $L = 64$, denoted by $\{\mathbf{y}_j^I\}$ and $\{\mathbf{y}_j^D\}$, $j = 1, \ldots, 500$. Signals in each pair have a common sparsity support of size $|\mathcal{T}_0|$, and they are sparse in random, Gaussian iid dictionaries $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$ of size 64×128. Their coefficients, $\{\mathbf{a}_j\}$ and $\{\mathbf{b}_j\}$, $j = 1, \ldots, 500$ are random, uniformly distributed, and do not have the same values nor signs. However, their ratios α_i (as defined in Eq. 19) are bounded from below, which gives a certain value of γ (see Eq. 20). This results in a distribution of coefficients shown by the scatter plot in Fig. 3, left panel, for $\gamma = 0.25$ and $|\mathcal{T}_0| = 10$. For smaller γ this joint distribution would be even more directional (thinner side lobes), while for $\gamma = 0$ it would fill out the whole space. Right panel of Fig. 3 shows joint distributions of coefficients in two modalities estimated from signals corrupted with noise of signal-to-noise ratio (SNR) equal to 20dB. If we zoom in, we can see that for small coefficient values, JBP and GL differ significantly. Due to its rotational symmetry in the space of coefficient pairs, GL cannot distinguish between pairs with a different balance of coefficients, especially for small coefficients. Unlike GL, JBP is able to recover the particular structure (driven by γ) in the distribution of coefficient pairs.

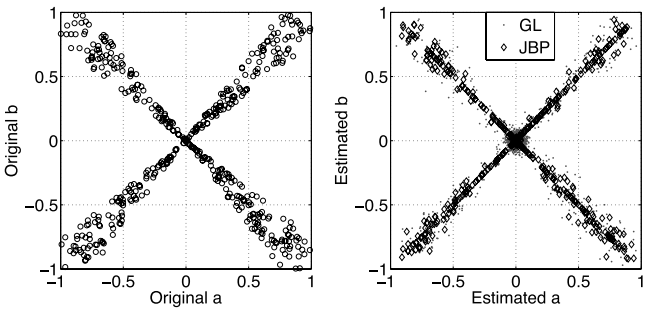


Fig. 3. Scatter plots illustrating joint distribution of coefficients for $\gamma = 0.25$ and $|\mathcal{T}_0| = 10$. Left: original coefficients (no noise). Right: estimated coefficients from signals corrupted with noise (SNR = 20dB) using JBP and GL.

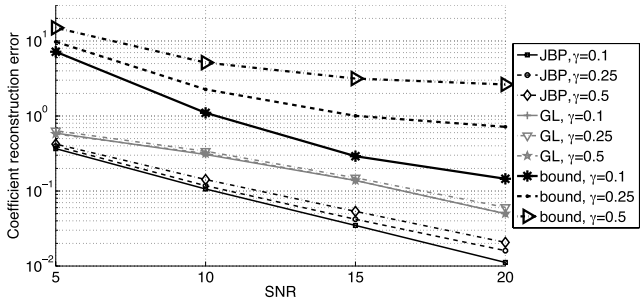
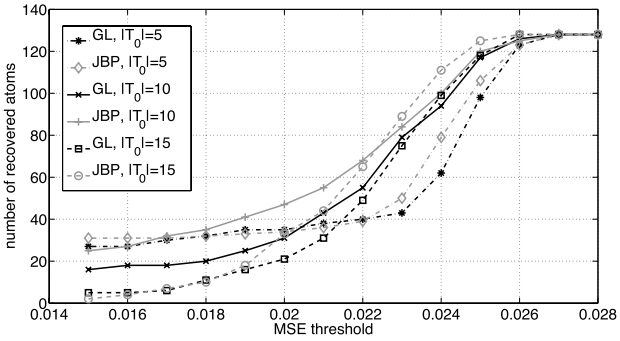


Fig. 4. JBP model recovery performance for random signals. Average coefficient reconstruction error is plotted for different signal-to-noise (SNR) ratios between sparse signals and Gaussian noise.

Fig. 4 shows the relative coefficient reconstruction error $\|\mathbf{a}^* - \mathbf{a}\|_2^2/\|\mathbf{a}\|_2^2 + \|\mathbf{b}^* - \mathbf{b}\|_2^2/\|\mathbf{b}\|_2^2$, where $(\mathbf{a}^*, \mathbf{b}^*)$ denote the reconstructions of original values $(\mathbf{a}, \mathbf{b})$. The error is averaged over 50 different signals and plotted versus the SNR between sparse signals and Gaussian noise. Coefficients have the sparsity parameter $|\mathcal{T}_0| = 10$, and the evaluation has been performed for three different values of γ: $0.1, 0.25, 0.5$. For each SNR value, we have chosen the η parameter in JBP to get the level of the reconstruction error the same as the noise level. For GL, we have performed recovery for a range of λ values and took the best results. We have compared JBP with GL and with the theoretical bound in Eq. (21), for $M = L = 64$ and for $\gamma = 0.1, 0.25, 0.5$. Instead of using the dictionary coherence value for δ, which would give the worst-case bounds, we use the mean of inner products between all atoms in order to calculate and plot the average case bounds. We can see that JBP outperforms GL for a large margin. Moreover, we can see that GL gives the same performance irrespective of γ, while JBP reaches better performance for smaller γ. This supports our claim that GL performs recovery without distinguishing between different structures in the data. Finally we see that the actual performance of JBP is much better than predicted by the theory, showing that the average derived bound is rather conservative.

Furthermore, we have used these randomly generated signals as training sets in our dictionary learning algorithm, in order to recover the original dictionary. All signals have



(b)

Fig. 5. Recovery performance of dictionary learning using JBP and GL for different sparsity $|\mathcal{T}_0|$: number of recovered atoms versus the MSE threshold under which the atom is considered recovered.

been corrupted with Gaussian noise of SNR=10 dB. For three sparsity values $|\mathcal{T}_0| = 5, 10, 15$, we have applied the proposed learning algorithm starting from a random initial dictionary. For comparison, we have replaced the JBP in the inference step with GL, while keeping the learning step exactly the same. We refer to this method as GL-based learning. For JBP we have chosen $\eta = 0.1$, which gives the reconstruction error equal to the noise level of 10 dB. Similarly, we have chosen $\lambda = 0.1$, to reach the similar reconstruction error level for GL. Fig. 5 shows the cumulative plots of the number of recovered atoms versus the threshold value above which the atom is considered recovered. The threshold is given in mean square error (MSE) between the original atoms and the recovered ones. We can see that learning based on JBP is superior to GL-based learning for most threshold values, or performs similarly for a small number of threshold values.

B. Intensity-Depth Dictionary Learning

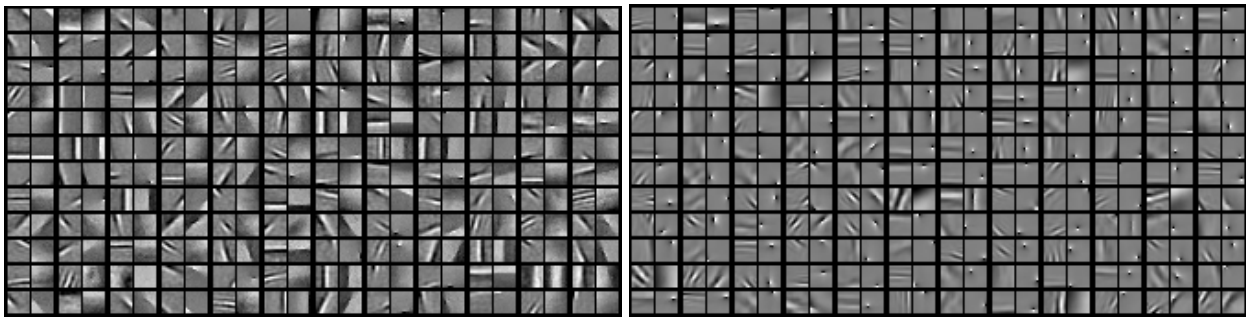
In our second set of experiments we have evaluated the performance of JBP and dictionary learning on real depth-intensity images. We have learned a depth-intensity overcomplete dictionary on the Middlebury 2006 benchmark depth-intensity data [24]. Each intensity image has been whitened (whole image), i.e., its frequency spectrum has been flattened, as initially proposed in [5]. Such pre-processing speeds up the learning. Depth data could not be whitened because it would introduce Gibbs artifacts around the missing regions at occlusions. We handle such missing pixels by masking. Learning has been performed in a patch-mode. Namely, in each iteration of the two-step learning process, a large number of depth-intensity pairs of 12×12 size patches have been randomly selected from data. Each depth and intensity patch within a pair coincide in a 3D scene. Patches have been normalized to have norm one, and η has been set to 0.1. We have chosen this value such that we get a good reconstruction of depth, without the quantization effects present in Middlebury depth maps (i.e., such that the quantization error is subsumed by the reconstruction error). We have learned dictionaries $\mathbf{\Phi}^I$ and $\mathbf{\Phi}^D$, each of size 144×288, i.e., twice overcomplete. For comparison, we have also learned depth-intensity dictionaries using GL-based learning, where $\lambda = 0.3$ has been chosen to obtain the same average reconstruction error as in JBP.[4]

Fig. 6(a) and (b) show parts of dictionaries learned by JBP and GL, respectively. The JBP-learned dictionary contains more meaningful features, such as coinciding depth-intensity edges, while GL-learned dictionary only has few of those. JBP dictionary atoms also exhibit correlation between orientations of the Gabor-like intensity atoms and the gradient angle of depth atoms. This is quite visible in the scatter plots of orientation angles of intensity atoms vs gradient angles of depth atoms, as shown in Fig. 7. We can see that for JBP there is significant clustering around the diagonal (corresponding to a 90° angle between intensity atom orientation and depth atom gradient). On the other hand, we cannot see this effect when using GL for learning. To the best of our knowledge, this is the first time that the correlation between depth gradient angles and texture orientations is found to emerge from natural scenes data (see [25] for some recent research in the area of 3D scene statistics).

We have also evaluated the statistics of parameters α and γ for a set of 250 patches, randomly selected from the training data. Values of α are evaluated for each atom within the sparse support (i.e., atoms with a sufficiently large coupling variable: $x > 0.1$) and then averaged over all atoms in all patches. The obtained mean and standard deviation values of α are 0.61 and 0.36, respectively. Since γ is defined via a bound on α, we have evaluated it per patch and then averaged, obtaining mean and standard deviation values of 0.81 and 0.27, respectively. We can see that the training data exhibits a large range of ratios between intensity and depth coefficients, which is efficiently captured by JBP.

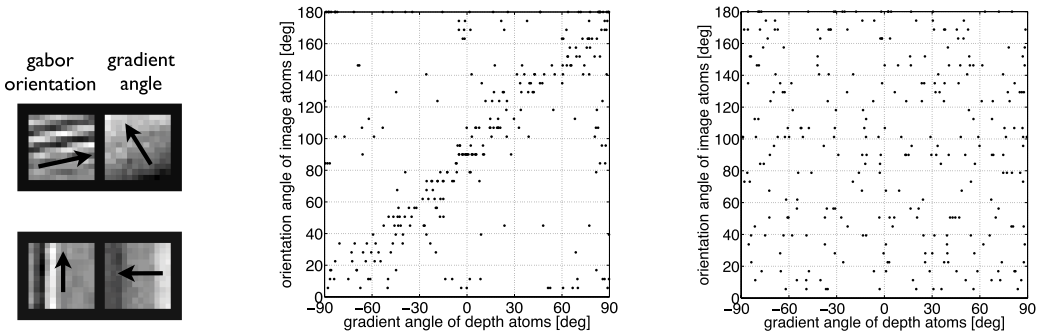
Finally, we have compared the performance of JBP and GL, and the corresponding learned dictionaries, on an inpainting task for data obtained with two different hybrid sensors: a time-of-flight (TOF) camera [26] and the Kinect sensor [3]. We have chosen the TOF and Kinect data to show that learned dictionaries of intensity-depth are not linked to particular depth sensors. The goal of inpainting is to fill out missing pixels from the depth map based on other available depth and intensity data. Since the image sizes are larger then our learned dictionary patches, we perform inpainting patch-wise, using a sliding window to reconstruct overlapping patches and then average them in depth pixel domain. Following previously introduced notation, we are given a vectorized intensity patch $\mathbf{y}^I$ and a depth patch as $\mathbf{y}^D$, both of size L. A set of pixels in $\mathbf{y}^D$ are labeled as missing using a mask vector $\mathbf{o}$ of the same size, which contains zeros on the position of missing pixels and ones elsewhere. Let us denote with $\mathbf{O}$ a diagonal matrix of size $L \times L$ that has the mask vector $\mathbf{o}$ on its diagonal. To solve the inpainting problem using JPB we need to solve the problem OPT2 for $\mathbf{x}, \mathbf{a}, \mathbf{b}$ with a modified constraint (10) as $\|\mathbf{O}\mathbf{y}^D - \mathbf{O}\mathbf{\Phi}^D\mathbf{b}\|_2^2 \leq (\epsilon^D)^2$. The obtained solution for $\mathbf{b}^*$ is then used to reconstruct the inpainted depth

[4]Note that λ value is a bit different here than in the synthetic data experiments, which is probably due to different noise statistics. Unlike in JBP, where η determines the reconstruction error, in unconstrained GL the error is harder to control and we can only try different λ until we reach the desired error SNR.



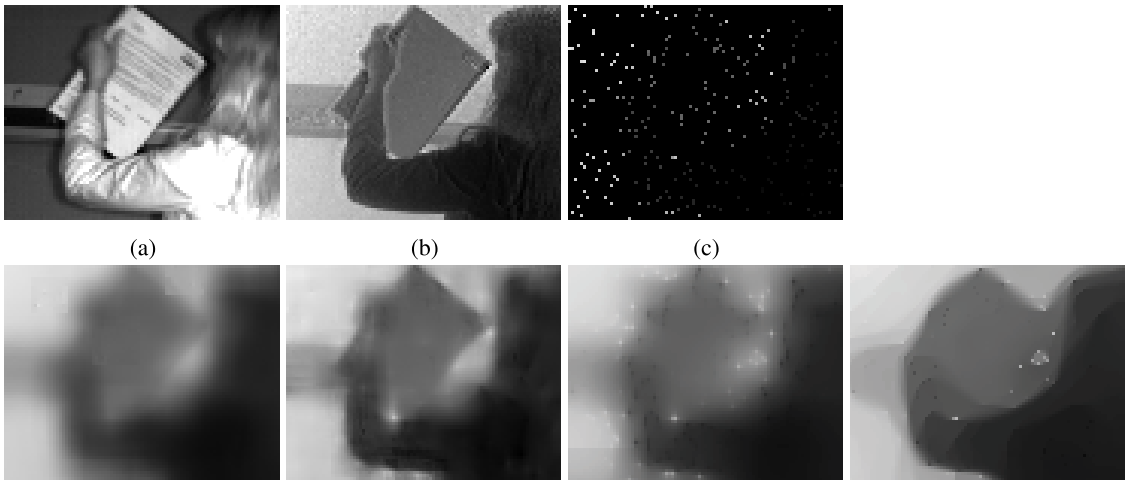
(a) (b)

Fig. 6. Learned intensity-depth dictionaries (only half of the atoms are displayed). Each column contains a set of atom pairs $(\boldsymbol{\phi}_1^I, \boldsymbol{\phi}^D)$, where the left part is an intensity atom and the right part is a depth atom. (a) JBP-learned dictionaries, (b) GL-learned dictionaries.



(a) (b) (c)

Fig. 7. Correlation between depth atom gradients and image intensity atom orientations. (a) Illustration of atom pairs that have 90 degrees angle between the orientation of the Gabor-like intensity part and the gradient angle of the depth part. Scatter plots of intensity orientation vs depth gradient angle for (b) JBP and (c) GL.



(a) (b) (c)

(d) (e) (f) (g)

Fig. 8. Inpainting results on time of flight data. (a) Original intensity image, (b) Original depth image, (c) 4% of kept depth pixels, (d) reconstructed depth with GL; MSE = 7.3e−3, MSSIM = 0.60; (e) reconstructed depth with JBP, MSE = 4.5e−3, MSSIM = 0.67; (f) reconstructed depth with learned depth dictionary, MSE = 8.8e−3, MSSIM = 0.60; (g) reconstructed depth with total variation inpainting, MSE = 7.9e−3, MSSIM = 0.61.

patch: $\hat{\mathbf{y}}^D = \boldsymbol{\Phi}^D \mathbf{b}^*$. Similarly, for the GL inpainting, we solve (33) where instead of the term $\|\mathbf{y}^D - \sum_i \boldsymbol{\phi}_i^D b_i\|_2^2$ we use $\|\mathbf{O}\mathbf{y}^D - \mathbf{O}\sum_i \boldsymbol{\phi}_i^D b_i\|_2^2$. The inpainted patch is reconstructed similarly as in JBP. Parameters η and λ were chosen the same as in learning.

For the TOF camera data, we have randomly removed 96% of depth pixels from an intensity-depth pair. Original intensity and depth images are shown in Fig. 8(a) and (b), respectively. From the original intensity image and 4% of depth pixels [shown in Fig. 8(c)], we have reconstructed the whole depth image, using GL with the GL-learned dictionary [Fig. 8(d)], and using JBP with the JBP-learned dictionary [Fig. 8(e)]. Since we had only 4% of pixels we have chosen a step size of 1 for sliding the window, which increased the probability that pixels are reconstructed in at least one patch. In addition, we have applied inpainting using the method proposed in [8]

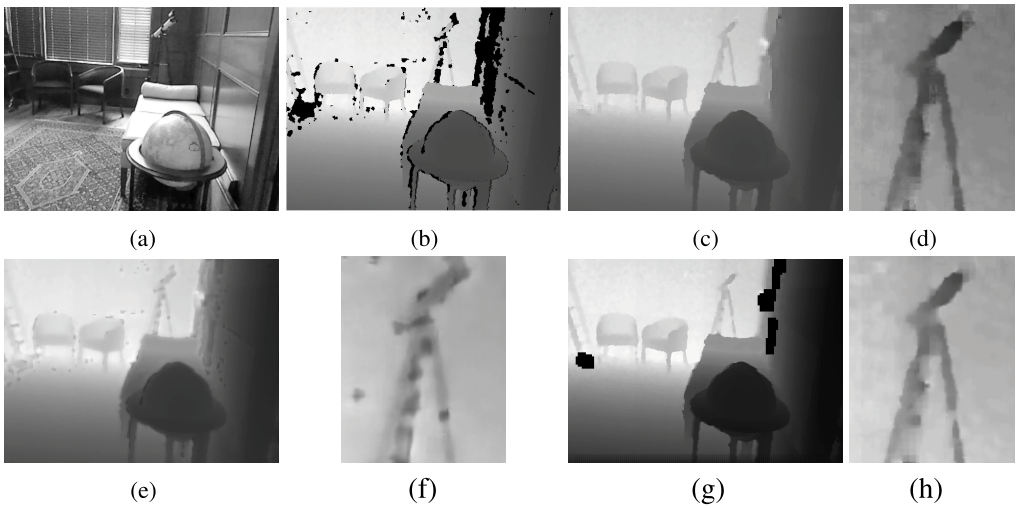


Fig. 9. Inpainting results on Kinect data from the NYU database. (a) Original intensity image, (b) Reprojected depth image with missing regions, (c) reconstructed depth with JBP, (d) zoom of (c); (e) reconstructed depth with GL; (f) zoom of (e); (g) reconstructed depth with learned depth dictionary; (h) zoom of (g).

based on a dictionary learned only from depth maps [Fig. 8(f)] and TV inpainting on depth masked image only [Fig. 8(g)]. We can see that JBP gives the best performance with the mean square error MSE $= 4.5\text{e}{-3}$ and mean structural similarity index [27] MSSIM $= 0.67$, followed by GL (MSE $=$ $7.3\text{e}{-3}$, MSSIM $= 0.60$), TV (MSE $= 7.9\text{e}{-3}$, MSSIM $=$ 0.61) and depth dictionary (DD) inpainting (MSE $= 8.8\text{e}{-3}$, MSSIM $= 0.60$). For the Kinect data, we have used one example image-depth pair from the "Homeoffice" scene in the NYU Kinect dataset [28], shown in Fig. 9. After the reprojection of the depth map to register it with the image data, many missing regions appear, as shown in Fig. 9(b). We apply previously explained inpainting procedure (with the same parameters and a window step of 4) to fill in these missing regions, and compare the results obtained with JBP, GL and DD. Since we do not have the ground truth in this case (missing data by the sensor), we can only make a visual comparison. From images in Figs. 9(c)–(h), we can see that JBP reconstructed depth map has the best quality and is even able to reconstruct the missing parts of the telescope tripod. Finally, we should mention that the quality improvement of JBP over GL comes at a higher computational cost. On a 3.1GHz 4-core Linux machine, JBP for one pair of 12×12 patches takes about 5–10 seconds, while GL for one value of λ takes 0.02 seconds. Both JBP and GL inpainting processes can be implemented in parallel. Therefore, for applications where time is critical GL is a better option, while for applications where quality of reconstruction is crucial JBP should be chosen.

VIII. Conclusion

We have presented an algorithm for learning joint overcomplete dictionaries of image intensity and depth. The proposed method, called JBP, is based on a novel second order cone program for recovering signals of joint sparse support in dictionaries with two modalities. We have derived a theoretical bound for the coefficient recovery error of JBP and shown its superiority to Group Lasso. Unlike GL, which does not distinguish between pairs with a different balance of coefficients, JBP can find a particular coefficient structure driven by the magnitude difference between intensity and depth signals. Moreover, since the performance of JBP increases with higher correlation between coefficients (smaller γ), dictionaries learned using JBP are expected to find intensity-depth structures with higher correlation between the two modalities. When applied to the Middlebury image-depth database, the proposed learning algorithm converges to a dictionary of intensity-depth features, such as coinciding edges and image grating-depth slant pairs. The learned features exhibit a significant correlation of depth gradient angles and texture orientations, which is an important result in 3D scene statistics research. Finally, we have shown that JBP with the learned dictionary can reconstruct meaningful depth maps from only 4% of depth pixels. These results outline the value of our method for 3D technologies based on hybrid image-depth sensors. In future work, we would like to modify JBP to take into account specific noise characteristics of hybrid image-depth sensors, for example by replacing the univariate Gaussian noise model with a more general multivariate Gaussian noise model used in [8].

Appendix

A. Proof of Theorem 1

Let us first prove the following lemma:

Lemma 3. *For* $\mathbf{h} := [\mathbf{a}^*; \mathbf{b}^*] - [\mathbf{a}^0; \mathbf{b}^0]$ *it holds true that* $\|\mathbf{h}_{\mathcal{T}_0^C}\|_1 \leq \|\mathbf{h}_{\mathcal{T}_0}\|_1 + \gamma\, U|\mathcal{T}_0|$, *where* $\mathcal{T}_0^C$ *denotes the complement set of* $\mathcal{T}_0$ *and* $\mathbf{h}_{\mathcal{T}}$ *denotes the subvector of* $\mathbf{h}$ *corresponding to* $\mathcal{T}$.

Proof: Define

$$\mathcal{I}_a^0 := \{i \in \mathcal{I} : |a_i^0| = U{x_i}^0\},$$
$$\mathcal{I}_b^0 := \{i \in \mathcal{I} \setminus \mathcal{I}_a^0 : |b_i^0| = U{x_i}^0\},$$
$$\mathcal{I}_a^* := \{i \in \mathcal{I} : |a_i^*| = U{x_i}^*\},$$
$$\mathcal{I}_b^* := \{i \in \mathcal{I} \setminus \mathcal{I}_a^* : |b_i^*| = U{x_i}^*\}.$$

Due to Lemma 2, we have that $\mathcal{I}_a^0 \cup \mathcal{I}_b^0 = \mathcal{I}$ and $\mathcal{I}_a^* \cup \mathcal{I}_b^* = \mathcal{I}$, and due to the definition above it holds that $\mathcal{I}_a^0 \cap \mathcal{I}_b^0 = \emptyset$ and $\mathcal{I}_a^* \cap \mathcal{I}_b^* = \emptyset$. Therefore, we have that:

$$\begin{aligned}\|[\mathbf{a}^*; \mathbf{b}^*]\|_1 &= \sum_{i\in\mathcal{I}_a^*} |a_i^*| + \sum_{i\in\mathcal{I}_b^*} |b_i^*| + \sum_{i\in\mathcal{I}_a^*} |b_i^*| + \sum_{i\in\mathcal{I}_b^*} |a_i^*| \\ &\leq U \sum_{i\in\mathcal{I}} |x_i^*| + U \sum_{i\in\mathcal{I}_a^*} |x_i^*| + U \sum_{i\in\mathcal{I}_b^*} |x_i^*| = 2U\|\mathbf{x}^*\|_1. \quad (34)\end{aligned}$$

Similarly, we have that:

$$\begin{aligned}\|[\mathbf{a}^0; \mathbf{b}^0]\|_1 &= \sum_{i\in\mathcal{I}_a^0} |a_i^0| + \sum_{i\in\mathcal{I}_b^0} |b_i^0| + \sum_{i\in\mathcal{I}_a^0} |b_i^0| + \sum_{i\in\mathcal{I}_b^0} |a_i^0| \\ &\geq U \sum_{i\in\mathcal{I}} |x_i^0| + \min_{i\in\mathcal{T}_0} \alpha_i \Big(\sum_{i\in\mathcal{I}_a^0} |a_i^0| + \sum_{i\in\mathcal{I}_b^0} |b_i^0|\Big) \\ &\geq^{(20)} 2U\|\mathbf{x}^0\|_1 - \gamma\, U|\mathcal{T}_0|. \quad (35)\end{aligned}$$

Due to optimality of $\mathbf{x}^*$, we have $\|\mathbf{x}^*\|_1 \leq \|\mathbf{x}^0\|_1$, which combined with (34) and (35) gives:

$$\|[\mathbf{a}^*; \mathbf{b}^*]\|_1 \leq 2U\|\mathbf{x}^0\|_1 \leq \|[\mathbf{a}^0; \mathbf{b}^0]\|_1 + \gamma\, U|\mathcal{T}_0|. \quad (36)$$

Due to $\mathbf{a}^0_{\mathcal{T}_0^C} = \mathbf{0}$ and $\mathbf{b}^0_{\mathcal{T}_0^C} = \mathbf{0}$, we can write

$$\begin{aligned}&\|[\mathbf{a}^0; \mathbf{b}^0] + \mathbf{h}\|_1 = \|[\mathbf{a}^0_{\mathcal{T}_0}; \mathbf{b}^0_{\mathcal{T}_0}; \mathbf{0}] + [\mathbf{h}_{\mathcal{T}_0}; \mathbf{h}_{\mathcal{T}_0^C}]\|_1 \\ &= \|[\mathbf{a}^0_{\mathcal{T}_0}; \mathbf{b}^0_{\mathcal{T}_0}] + \mathbf{h}_{\mathcal{T}_0}\|_1 + \|\mathbf{h}_{\mathcal{T}_0^C}\|_1. \quad (37)\end{aligned}$$

Thus, using the triangle inequality and the definition of $\mathbf{h}$ we derive:

$$\begin{aligned}&\|[\mathbf{a}^0; \mathbf{b}^0]\|_1 - \|\mathbf{h}_{\mathcal{T}_0}\|_1 + \|\mathbf{h}_{\mathcal{T}_0^C}\|_1 \leq \|[\mathbf{a}^0; \mathbf{b}^0] + \mathbf{h}\|_1 \\ &= \|[\mathbf{a}^*; \mathbf{b}^*]\|_1 \leq^{(36)} \|[\mathbf{a}^0; \mathbf{b}^0]\|_1 + \gamma\, U|\mathcal{T}_0|\end{aligned}$$

and thus $\|\mathbf{h}_{\mathcal{T}_0^C}\|_1 \leq \|\mathbf{h}_{\mathcal{T}_0}\|_1 + \gamma\, U|\mathcal{T}_0|$. ■

We are now ready to prove Theorem 1.

Proof: Let $\mathbf{A}$ be defined as in Eq. (14). Then we have from (9) and (10) that $\|\mathbf{A}\mathbf{h}\|_2 \leq 4\epsilon = 4\eta f_0$. Assume we have divided $\mathcal{T}_0^C$ into subsets of size M, more precisely, we have $\mathcal{T}_0^C = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_{n-|\mathcal{T}_0|}$, where $\mathcal{T}_i$ are sorted by decreasing order of $\mathbf{h}_{\mathcal{T}_0^C}$, and where $\mathcal{T}_{01} = \mathcal{T}_0 \cup \mathcal{T}_1$. Without alternations-cf. [13]-it holds true that $\|\mathbf{h}_{\mathcal{T}_{01}^C}\|_2^2 \leq \|\mathbf{h}_{\mathcal{T}_0^C}\|_1^2/M$. Using Lemma 3 yields

$$\begin{aligned}\|\mathbf{h}_{\mathcal{T}_{01}^C}\|_2^2 &\leq (\|\mathbf{h}_{\mathcal{T}_0}\|_1 + \gamma\, U|\mathcal{T}_0|)^2/M \\ &\leq (\sqrt{|\mathcal{T}_0|}\|\mathbf{h}_{\mathcal{T}_0}\|_2 + \gamma\, U|\mathcal{T}_0|)^2/M, \quad (38)\end{aligned}$$

where the second step follows from the norm inequality. Hence:

$$\begin{aligned}\|\mathbf{h}\|_2^2 &= \|\mathbf{h}_{\mathcal{T}_{01}}\|_2^2 + \|\mathbf{h}_{\mathcal{T}_{01}^C}\|_2^2 \leq (1 + \frac{|\mathcal{T}_0|}{M})\|\mathbf{h}_{\mathcal{T}_0}\|_2^2 \\ &\quad + \frac{2\gamma\, U|\mathcal{T}_0|^{3/2}}{M}\|\mathbf{h}_{\mathcal{T}_0}\|_2 + \frac{(\gamma\, U|\mathcal{T}_0|)^2}{M}. \quad (39)\end{aligned}$$

From the restricted isometry, cf. Def. 1, we get

$$\begin{aligned}\|\mathbf{A}\mathbf{h}\|_2 &= \|\mathbf{A}_{\mathcal{T}_{01}}\mathbf{h}_{\mathcal{T}_{01}} + \sum_{j\geq 2} \mathbf{A}_{\mathcal{T}_j}\mathbf{h}_{\mathcal{T}_j}\|_2 \\ &\geq \|\mathbf{A}_{\mathcal{T}_{01}}\mathbf{h}_{\mathcal{T}_{01}}\|_2 - \|\sum_{j\geq 2} \mathbf{A}_{\mathcal{T}_j}\mathbf{h}_{\mathcal{T}_j}\|_2 \\ &\geq \|\mathbf{A}_{\mathcal{T}_{01}}\mathbf{h}_{\mathcal{T}_{01}}\|_2 - \sum_{j\geq 2} \|\mathbf{A}_{\mathcal{T}_j}\mathbf{h}_{\mathcal{T}_j}\|_2 \\ &\geq \sqrt{1-\delta_{M+|\mathcal{T}_0|}}\|\mathbf{h}_{\mathcal{T}_{01}}\|_2 - \sqrt{1+\delta_M}\sum_{j\geq 2}\|\mathbf{h}_{\mathcal{T}_j}\|_2 \\ &\geq \sqrt{1-\delta_{M+|\mathcal{T}_0|}}\|\mathbf{h}_{\mathcal{T}_0}\|_2 - \sqrt{1+\delta_M}\sum_{j\geq 2}\|\mathbf{h}_{\mathcal{T}_j}\|_2 \quad (40)\end{aligned}$$

where δ_S is a constant chosen such that the inequalities hold, which follows from inequality (4) in [13]. Here, $\mathbf{A}_{\mathcal{T}}$ denotes the columns of $\mathbf{A}$ corresponding to the index set $\mathcal{T}$. In analogy to [13], due to the ordering of the sets $\mathcal{T}_j$ by decreasing order of coefficients, we have: $|\mathbf{h}_{\mathcal{T}_{j+1}(t)}| \leq \|\mathbf{h}_{\mathcal{T}_j}\|_1/M$ meaning each component in $\mathbf{h}_{\mathcal{T}_{j+1}}$ is smaller than the average of the components in $\mathbf{h}_{\mathcal{T}_j}$ (absolute value-wise). Thus, we get:

$$\begin{aligned}\|\mathbf{h}_{\mathcal{T}_{j+1}}\|_2^2 &= \sum_{t\in\mathcal{T}_{j+1}} \|\mathbf{h}_t\|_2^2 \leq \sum_{t\in\mathcal{T}_{j+1}} \|\mathbf{h}_{\mathcal{T}_j}\|_1^2/M^2 \\ &\leq M\|\mathbf{h}_{\mathcal{T}_j}\|_1^2/M^2 = \|\mathbf{h}_{\mathcal{T}_j}\|_1^2/M, \quad \text{and} \\ \sum_{j\geq 2}\|\mathbf{h}_{\mathcal{T}_j}\|_2 &\leq \sum_{j\geq 1}\|\mathbf{h}_{\mathcal{T}_j}\|_1/\sqrt{M} = \|\mathbf{h}_{\mathcal{T}_0^C}\|_1/\sqrt{M} \\ &\leq^{(\text{Lemma 3})} (\|\mathbf{h}_{\mathcal{T}_0}\|_1 + \gamma\, U|\mathcal{T}_0|)/\sqrt{M} \\ &\leq \sqrt{|\mathcal{T}_0|/M}\|\mathbf{h}_{\mathcal{T}_0}\|_2 + \gamma\, U|\mathcal{T}_0|/\sqrt{M} \quad (41)\end{aligned}$$

where the last step follows from the norm inequality. Combining Eq. (41) and Eq. (40), we get:

$$\begin{aligned}&\|A\mathbf{h}\|_2 \geq \sqrt{1-\delta_{M+|\mathcal{T}_0|}}\|\mathbf{h}_{\mathcal{T}_0}\|_2 \\ &-\sqrt{1+\delta_M}\sqrt{|\mathcal{T}_0|/M}\|\mathbf{h}_{\mathcal{T}_0}\|_2 - \gamma\, U|\mathcal{T}_0|\sqrt{1+\delta_M}/\sqrt{M}\end{aligned}$$

and subsequently:

$$\begin{aligned}\|\mathbf{h}_{\mathcal{T}_0}\|_2 &\leq \frac{\|A\mathbf{h}\|_2 + \gamma\, U|\mathcal{T}_0|\sqrt{1+\delta_M}/\sqrt{M}}{\sqrt{1-\delta_{M+|\mathcal{T}_0|}} - \sqrt{1+\delta_M}\sqrt{|\mathcal{T}_0|/M}} \\ &\leq \frac{4\eta f_0\sqrt{M} + \gamma\, f_0|\mathcal{T}_0|\sqrt{1+\delta_M}}{\sqrt{M(1-\delta_{M+|\mathcal{T}_0|})} - \sqrt{|\mathcal{T}_0|(1+\delta_M)}} = Cf_0,\end{aligned}$$

if the denominator is greater than zero. Replacing this result in Eq. (39) and taking $U = f_0$ we get:

$$\|\mathbf{h}\|_2^2 \leq (1 + \frac{|\mathcal{T}_0|}{M})C^2 f_0^2 + 2\gamma\,\frac{|\mathcal{T}_0|^{3/2}}{M}Cf_0^2 + \gamma^2\frac{|\mathcal{T}_0|^2}{M}f_0^2,$$

which is equivalent to (21) and thus completes the proof. ■

ACKNOWLEDGMENT

I. Tošić would like to thank Prof. Bruno Olshausen from the Redwood Center for Theoretical Neuroscience at UC Berkeley for his insightful comments on the intensity-depth model proposed in this work. The authors would also like to thank the associated editor and the anonymous reviewers for their valuable suggestions that have lead to the improved quality of this manuscript.

REFERENCES

[1] T. Ringbeck and B. Hagebeuker, "A 3D time of flight camera for object detection," in *Proc. Opt. 3-D Meas. Tech.*, 2007, pp. 1–10.

[2] T. Oggier *et al.*, "An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger)," *Proc. SPIE*, vol. 5200, pp. 534–545, Feb. 2004.

[3] (2014, Mar.). Microsoft Kinect, Microsoft Corp., Redmond, WA, USA [Online]. Available: http://www.xbox.com/en-US/kinect

[4] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 10–21, Nov. 2007.

[5] B. A. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[6] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–65, 2000.

[7] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[8] I. Tošić, B. A. Olshausen, and B. J. Culpepper, "Learning sparse representations of depth," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 941–952, Sep. 2011.

[9] 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, Nov. 2001.

[10] S. Bakin, *Adaptive Regression and Model Selection in Data Mining Problems*. ACT, Australia: Australian Nat. Univ., 1999.

[11] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. R. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 70, no. 1, pp. 53–71, 2008.

[12] E. J. Candés and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[13] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[14] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, 2006.

[15] S. Jokar and M. E. Pfetsch, "Exact and approximate sparse solutions of underdetermined linear equations," *SIAM J. Sci. Comput.*, vol. 31, no. 1, pp. 23–44, 2008.

[16] T. Tsuchiya, "A convergence analysis of the scaling-invariant primal-dual path-following algorithms for second-order cone programming," *Optim. Methods Softw.*, vol. 11, nos. 1–4, pp. 141–182, 1998.

[17] E. D. Andersen, C. Roos, and T. Terlaky, "On implementing a primal-dual interior-point method for conic quadratic optimization," *Math. Program.*, vol. 95, no. 2, pp. 249–277, 2003.

[18] (2014, Mar.). *IBM ILOG CPLEX Optimizer*. IBM, Armonk, NY, USA [Online]. Available: http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

[19] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[20] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[21] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[22] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar, "Dictionary optimization for block-sparse representations," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2386–2395, May 2012.

[23] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1–2, pp. 89–97, 2004.

[24] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, 2002.

[25] B. Potetz and T. S. Lee, "Scene statistics and 3D surface perception," in *Computational Vision: From Surfaces to Objects*, C. W. Tyler, Ed. London, U.K.: Chapman & Hall, 2010, ch. 1, pp. 1–25.

[26] (2014, Mar.). pmdtechnologies, Siegen, Germany [Online]. Available: http://www.pmdtec.com/

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th ECCV*, 2012, pp. 746–760.

Ivana Tošić (S'03–M'09) received the Ph.D. degree in computer and communication sciences from the Swiss Federal Institute of Technology, Switzerland, and the Dipl.Ing. degree in telecommunications from the University of Niš, Serbia, in 2009 and 2003, respectively.

From 2009 to 2011, she was a Post-Doctoral Researcher with the Redwood Center for Theoretical Neuroscience, University of California at Berkeley, where she studied computational mechanisms of depth perception from binocular vision. Since 2011, she has been a member of research staff at Ricoh Innovations Corporation, Menlo Park, CA, USA, involved in the computational optics and visual processing group. Her research interests lie in the intersection of image processing and computational neuroscience domains, binocular vision, image and 3-D scene representation, depth perception, representation and coding of the plenoptic function, and computational photography.

Dr. Tošić was a Finance Chair for the 2007 Packet Video Workshop and a Publicity Chair for the 2013 Picture Coding Symposium. She has been an Elected Member of the IEEE Image and Multidimensional Signal Processing Technical Committee since 2014. She received a prestigious fellowship from the Serbian Royal Family in 2002, and two post-doctoral fellowships from the Swiss National Science Foundation in 2009 and 2010.

Sarah Drewes received the Ph.D. (Dr. rer. net) degree in mathematics from the Technische Universität Darmstadt, Germany, in 2009, and the Diploma (Dipl.Math., Univer.) degree in mathematics from the Technische Universität München, Germany, in 2005.

From 2010 to 2011, she was a Lecturer and Post-Doctoral Associate with the Computational Optimization Laboratory, Department of Industrial Engineering and Operations Research, University of California, Berkeley. From 2011 to 2013, she was with T-Systems International, Germany, where she conducted operations research projects for network analysis, planning, and optimization at Deutsche Telekom.

Since 2013, she has been a Senior Consultant with The MathWorks GmbH, helping costumers to solve complex optimization problems in a variety of areas, such as communications, finance, or automotive. Her research interests lie in algorithms for complex optimization problems, focusing on mixed integer nonlinear optimization, where both discrete variables as well as model nonlinearities occur.

In 2010, she received the Ruth-Moufang Award for the Outstanding Female Post-Doctoral Researchers from the Department of Mathematics, Technische Universität Darmstadt, Germany.