Dictionary Learning for Stereo Image Representation

Ivana Tošić, *Member, IEEE*, and Pascal Frossard, *Senior Member, IEEE*

***Abstract*—One of the major challenges in multi-view imaging is the definition of a representation that reveals the intrinsic geometry of the visual information. Sparse image representations with overcomplete geometric dictionaries offer a way to efficiently approximate these images, such that the multi-view geometric structure becomes explicit in the representation. However, the choice of a good dictionary in this case is far from obvious. We propose a new method for learning overcomplete dictionaries that are adapted to the joint representation of stereo images. We first formulate a sparse stereo image model where the multi-view correlation is described by local geometric transforms of dictionary elements (atoms) in two stereo views. A maximum-likelihood (ML) method for learning stereo dictionaries is then proposed, where a multi-view geometry constraint is included in the probabilistic model. The ML objective function is optimized using the expectation-maximization algorithm. We apply the learning algorithm to the case of omnidirectional images, where we learn scales of atoms in a parametric dictionary. The resulting dictionaries provide better performance in the joint representation of stereo omnidirectional images as well as improved multi-view feature matching. We finally discuss and demonstrate the benefits of dictionary learning for distributed scene representation and camera pose estimation.**

***Index Terms*—Dictionary learning, multi-view imaging, omnidirectional cameras, sparse approximations.**

I. Introduction

RECENT development of camera network applications has fostered a large interest in multi-view and stereo image processing research. Visual sensor networks capture multiple images of a 3-D scene from different viewpoints. The resulting multi-view images contain rich information about both the structure and the texture of objects in the 3-D scene. The processing of such high dimensional visual information opens many research challenges, such as multi-view compression, 3-D geometry estimation and scene analysis.

Manuscript received June 10, 2009; revised December 10, 2009; accepted September 06, 2010. Date of publication September 30, 2010; date of current version March 18, 2011. This work has been supported in part by the Swiss National Science Foundation under grants 200020–120063 and PBELP2–127847, and by the EU under the FP7 project APIDIS (ICT-216023). The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Arun Ross.

I. Tošić was with the Signal Processing Laboratory LTS4, Ecole Polytechnique Fédérale de Lausanne, Switzerland. She is now with the Redwood Center for Theoretical Neuroscience, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: ivana@berkeley.edu).

P. Frossard is with the Signal Processing Laboratory LTS4, Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: 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.2010.2081679

The representation of 3-D information from multiple views relies on the theory of multiple view geometry [1], which relates image features representing the same 3-D objects in different views. Pixel-based image representation is used in most of the image-based 3-D geometry (depth) estimation methods. However, pixel representations are highly inefficient for image compression. On the contrary, image representations with orthogonal bases are efficient for compression but generally fail to efficiently capture the true geometry of objects in a scene. An alternative solution could be to model intra-view correlation by block translations [2], but these approaches usually suffer from blocking artifacts. Therefore, multi-view imaging requires new representation methods that give good performance both in compression and scene geometry estimation. Sparse image representations with overcomplete geometric dictionaries offer a promising way to solve this problem [3].

An *overcomplete dictionary* is a collection of waveforms called *atoms*, whose size is greater than the signal dimension. Since there is not much restriction on the choice of a dictionary except that it has to span the signal space, one can choose any collection of atoms that is capable of capturing the signal structure. An interesting signal representation is the one that is sparse, i.e., the signal is represented as a linear combination of only few atoms. Sparsity plays a crucial role in compression and in a variety of inverse problems. However, sparsity largely depends on the design of the dictionary, which represents a difficult and challenging problem. Dictionary learning for sparse signal representations has become an extremely active area of research in the last few years, for example in image [4]–[6] and video representation [7]–[9]. Still, there has been no work on learning dictionaries for stereo image representation. Learning such dictionaries could bring significant improvements in applications that require multi-view compression or camera pose and depth estimation for example.

This paper presents a novel method for learning dictionaries of geometric atoms that are adapted to the representation of multi-view images. As the correlation between multi-view images arises from the geometric constraints on objects in the scene, it can be simply described by local geometric transforms of atoms [3]. We propose to learn dictionaries that efficiently describe the content of natural images and simultaneously permit to capture the geometric correlation between multi-view images. We concentrate on the problem of two views and develop a maximum likelihood (ML) method for learning dictionaries that lead to improved image approximation under a sparsity prior, and at the same time give better matching of sparse low-level visual features in stereo views. Learning is based on a new probabilistic model that includes the epipolar geometry relations. The ML optimization problem is cast as an energy minimization problem, which is solved with an Expectation-Maximization (EM) algorithm. The experimental results show the signifi-

cant benefits of stereo dictionary learning for applications such as distributed scene representation and camera pose recovery.

We first overview the related work on dictionary learning in Section II. The stereo image model is introduced in Section III. Section IV presents the optimization problem for learning dictionaries adapted to stereo images, while its solution is given in Section V. Experimental results in omnidirectional imaging are presented in Sections VI and VII.

II. Related Work

One of the earliest work addressing the problem of learning overcomplete dictionaries for image representation was the sparse coding method of Olshausen and Field [4], [10], primarily introduced as a model of early visual sensory coding. This method is based on maximizing the likelihood that a natural image $\mathbf{y}$ arises from the overcomplete dictionary $\mathbf{\Phi}$, when the generative image model is considered as sparse image decomposition into dictionary elements. Therefore, the ML method solves the optimization problem $\mathbf{\Phi}^* = \max_{\mathbf{\Phi}} P(\mathbf{y}|\mathbf{\Phi})$, for $\mathbf{y} = \mathbf{\Phi}\mathbf{a} + \mathbf{e}$, where the coefficient vector $\mathbf{a}$ is considered as a hidden variable and $\mathbf{e}$ is a noise vector. The optimization is solved in two iterative steps: the sparse coding step, where the dictionary is kept fixed and the sparse coefficient vector $\mathbf{a}$ that best approximates the image is found; and a dictionary update step, where $\mathbf{a}$ is kept fixed and the dictionary is updated to maximize the objective maximum likelihood function using gradient descent. The proposed learning method yields dictionary components (atoms) that are localized, oriented and bandpass, and resemble the receptive fields of simple neurons in the primary visual area V1 in mammalian brain. This method has also been extended to natural movies [7]–[9].

The probabilistic approach to dictionary learning has been later adopted by other researchers. Engan *et al.* [11], [12] have introduced a method of optimal directions (MOD), which includes the sparse coding and dictionary update steps that iteratively optimize the objective ML function. Their method differs from the work of Olshausen and Field in two aspects. First, while in [4] the sparse coding step involves finding the equilibrium solution of the differential equation over $\mathbf{a}$, MOD uses either the OMP [11] or the FOCUSS [12] algorithm to find sparse vector $\mathbf{a}$. Second, instead of doing gradient descent, the dictionary $\mathbf{\Phi}$ is updated by the closed-form solution of the minimum energy constraint. These two modifications make the MOD approach faster compared to the ML method of Olshausen and Field. Maximum *a posteriori* (MAP) dictionary learning method, proposed by Kreutz-Delgado *et al.* [5] belongs also to the family of two-step iterative algorithms based on probabilistic inference. Instead of maximizing the likelihood $P(\mathbf{y}|\mathbf{\Phi})$, the MAP method maximizes the posterior probability $P(\mathbf{\Phi}, \mathbf{a}|\mathbf{y})$. This essentially reduces to the same two-step algorithm (i.e., sparse coding-dictionary update), where dictionary update includes an additional constraint on the dictionary that can be for example a unit Frobenius norm of $\mathbf{\Phi}$ or a unit l_2 norm of all atoms in the dictionary. The sparse coding step is performed with FOCUSS [13].

A slightly different family of dictionary learning techniques is based on vector quantization (VQ) achieved by K-means clustering. The VQ approach for dictionary learning has been first proposed by Schmid-Saugeon and Zakhor in Matching Pursuit based video coding [14]. Their algorithm optimizes a dictionary given a set of image patches by first grouping patterns such that their distance to a given atom is minimal, and then by updating the atom such that the overall distance in the group of patterns is minimal. The implicit assumption here is that each patch can be represented by a single atom with a coefficient equal to one, which reduces the learning procedure to K-means clustering. Since each patch is represented by only one atom, the sparse coding step is trivial here. A generalization of the K-means for dictionary learning, called the K-SVD algorithm, has been proposed by Aharon *et al.* in [6]. After the sparse coding step (where any pursuit algorithm can be employed), the dictionary update is performed by sequentially updating each column of $\mathbf{\Phi}$ using a singular value decomposition (SVD) to minimize the approximation error. The update step hence follows a generalized K-means algorithm since each patch can be represented by multiple atoms with different weights.

Finally, there exist other approaches for learning special types of dictionaries, like unions of orthonormal basis [15], shift-invariant dictionaries [16], block-based dictionaries and constrained overlapping dictionaries [17]. A comparison of all state-of-the-art dictionary learning methods is made difficult by the fact that the efficiency of the algorithms differs with the dictionary size and the training data. Advantageously, ML and MAP methods are characterized by the possibility of extending the probabilistic modeling to higher-dimensional data, like videos [8], [9] or stereo images. It is also possible to include different correlated modalities such as audio and visual signals in order to learn audio-visual dictionaries [18]. Because of this property, we have chosen the ML approach for learning parametric dictionaries in stereo imaging.

Even though there exists a plethora of dictionary learning methods for monocular images, there has been little work that target the problem of learning overcomplete dictionaries for stereo imaging or for binocular sensory coding. Hoyer and Hyvärinen [19] have applied the independent component analysis (ICA) to learn the orthogonal basis of stereo images. In their model, each stereo pair is a linear combination of stereo basis functions, which are composed of left and right components. Their algorithm results in Gabor-like basis functions that are tuned to different disparities. Okajima has proposed an Infomax learning approach, where the receptive fields of binocular neurons are learned by maximizing the mutual information between the stereo image model and the disparity [20]. They have obtained results similar to Hoyer and Hyvärinen. The stereo dictionary learning method that we propose in this work differs from previous work because it learns stereo atoms from stereo image pairs *under explicit geometric constraints*. The disparity estimation is included in the probabilistic model of stereo images, thus removing the need for disparity compensation as a preprocessing step (e.g., in [19] eye fixation selection is applied before ICA). However, the main target of this work

is not to study the receptive fields of binocular cells, but rather to design stereo dictionaries that have good properties for both image approximation and 3-D structure estimation.

III. Multi-View Imaging

A. Stereo Image Model

Before deriving a maximum likelihood dictionary learning method for stereo images, we first need to define the stereo image model. We consider two images vectorized into column vectors: the left image $\mathbf{y}_L$ and right image $\mathbf{y}_R$. The images $\mathbf{y}_L$ and $\mathbf{y}_R$ have sparse representations in dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, respectively. Both dictionaries are of size M. The images do not have to be exactly sparse, but they can be approximated by sparse decompositions of m atoms up to an approximation error $\mathbf{e}_L$, resp. $\mathbf{e}_R$. Hence, we have:

$$\begin{aligned}\mathbf{y}_L &= \mathbf{\Phi}\mathbf{a} + \mathbf{e}_L = \sum_{k=1}^{m} a_{l_k}\boldsymbol{\phi}_{l_k} + \mathbf{e}_L \\ \mathbf{y}_R &= \mathbf{\Psi}\mathbf{b} + \mathbf{e}_R = \sum_{k=1}^{m} b_{r_k}\boldsymbol{\psi}_{r_k} + \mathbf{e}_R \end{aligned} \tag{1}$$

where the vectors $\mathbf{a}$ and $\mathbf{b}$ represent the coefficients for the left and the right image, respectively. The index sets $\mathcal{L} = \{l_k\}$, $\mathcal{R} = \{r_k\}$, $k = 1, \ldots, m$ label the atoms that participate in the sparse decompositions of $\mathbf{y}_L$ and $\mathbf{y}_R$, respectively. In other words, $\{l_k\}$, $\{r_k\}$, $k = 1, \ldots, m$ denote the atoms with non-zero coefficients, i.e., $a_{l_k} \neq 0$ and $b_{r_k} \neq 0$. This model assumes that both stereo images are m-sparse, i.e., composed of m atoms, while the atoms in the left and the right image do not have to be identical ($\mathcal{L} \neq \mathcal{R}$). The motivation behind this model is that left and right images record the visual information from the same 3-D environment and typically contain the image projections of the same 3-D scene features, thus the number of sparse components will be approximately the same. However, the sparse components might be different since the images are recorded from different viewpoints. If the dictionary consists of localized and oriented atoms that represent well the edges and objects geometry in general, we can say that stereo images contain similar atoms, but locally transformed (shifted, rotated, etc.). Therefore, we further assume that signals $\mathbf{y}_L$ and $\mathbf{y}_R$ are correlated in the following way:

$$\mathbf{y}_R = \sum_{k=1}^{m} b_{r_k}\boldsymbol{\psi}_{r_k} + \mathbf{e}_R = \sum_{k=1}^{m} b_{r_k} F_{l_k r_k}(\boldsymbol{\phi}_{l_k}) + \mathbf{e}_R \tag{2}$$

where $F_{l_k r_k}(\cdot)$ denotes the transform of an atom $\boldsymbol{\phi}_{l_k}$ in $\mathbf{y}_L$ to an atom $\boldsymbol{\psi}_{r_k}$ in $\mathbf{y}_R$. It is different for each $k = 1, \ldots, m$. This correlation model is a special case of the model introduced in [3] when there are no occlusions. Since they do not participate in stereo matching, occlusions should not be considered for the learning of stereo dictionaries. Therefore, we assume in (2) that the occlusions in the scene are not dominant and that they can be included in the approximation errors $\mathbf{e}_R$, $\mathbf{e}_L$.

Object and atom transforms arising from the change of viewpoint can be usually represented by the 2-D similarity group elements (2-D translation, rotation and isotropic scaling) and anisotropic scaling of the image features [3]. Such transforms are efficiently represented with a parametric dictionary whose construction is built on translation, rotation and scaling of a generating function g defined in the Hilbert space $\mathcal{H}$. Formally, the parametric dictionary $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ is constructed by changing the atom index $\gamma \in \Gamma$ that defines the rotation, translation and scaling transforms applied to the generating function g. This is equivalent to applying a unitary operator $U(\gamma)$ to the generating function g, i.e.,: $g_\gamma = U(\gamma)g$. Finally, the function g_γ needs to be normalized and have the l_2 norm equal to one[1]. Therefore, we define atoms in the structured dictionary as: $\boldsymbol{\phi} = g_\gamma / \|g_\gamma\|_2$.

The multi-dimensional space of parameters γ is continuous and infinite, thus building a dictionary with all possible transforms yields an infinite dictionary. However, in practical cases, only a discrete set $\Gamma = \{\gamma\}$ of transform parameters is used. We further define our dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as structured dictionaries that are built on the same generating function g, but with possibly different sets of parameters: Γ_L for $\mathbf{\Phi}$, and Γ_R for $\mathbf{\Psi}$. To simplify the notation, we introduce the following equivalencies:

$$\begin{aligned}\boldsymbol{\phi}_l &\equiv g_{\gamma_l^{(L)}} \quad \gamma_l^{(L)} \in \Gamma_L \quad \text{for } l = 1, \ldots, M \\ \boldsymbol{\psi}_r &\equiv g_{\gamma_r^{(R)}} \quad \gamma_r^{(R)} \in \Gamma_R \quad \text{for } r = 1, \ldots, M \end{aligned} \tag{3}$$

where we assume that $g_{\gamma_l^{(L)}}$ and $g_{\gamma_r^{(R)}}$ are normalized. An important property of the structured dictionary is that a transform of an atom $\boldsymbol{\phi}_l$ in the left image into an atom $\boldsymbol{\psi}_r$ in the right image reduces to a transform of its transform parameters, i.e.,

$$\boldsymbol{\psi}_r = F_{lr}(\boldsymbol{\phi}_l) = U(\gamma')\boldsymbol{\phi}_l = U\left(\gamma' \circ \gamma_l^{(L)}\right) g. \tag{4}$$

In the following, the transform $F(\cdot)$ that changes atom $\boldsymbol{\phi}_l$ into an atom $\boldsymbol{\psi}_r$ is denoted as $F_{lr}(\cdot)$. Since the parametric dictionary is built on rotation, translation and anisotropic scaling of the generating function, these are the types of atom transforms that we consider. Note also that one can design parametric dictionaries built on perspective transforms where each atom would be described by eight parameters. However, this would increase the dictionary size and the complexity.

B. Multi-View Geometry

The transforms $F_{l_k r_k}, k = 1, \ldots, m$ relating the atoms in the left and right view in (2) are not arbitrary. As the corresponding atoms are approximations of the same features in the 3-D space, the atom transforms have to satisfy multi-view epipolar geometry constraints. The epipolar geometry constraint imposes a geometric relation between 3-D points and their image projections. Consider a point on the left image, given by the coordinates $\mathbf{v}$, and a point $\mathbf{u}$ on the right image. Let these two points represent image projections of the same 3-D point p from two camera positions with a relative pose $(\mathbf{R}, \mathbf{T})$. $\mathbf{R} \in SO(3)$ is the relative orientation between cameras. $\mathbf{T} \in \mathbb{R}^3$ is their relative position. Let further $\mathbf{v}$ lie on the spatial support of atom $\boldsymbol{\phi}_l$ and $\mathbf{u}$ lie on the spatial support of atom $\boldsymbol{\psi}_r = F_{lr}(\boldsymbol{\phi}_l)$, as illustrated in Fig. 1. It can be shown that transforming the atom $\boldsymbol{\phi}_l$ with F_{lr} is equivalent to a linear transform of the coordinate system $Q_{lr}(\cdot)$, i.e., $\mathbf{u} = Q_{lr}(\mathbf{v})$ [21]. The epipolar geometry constraint is then

[1]The vector l_p norm is denoted as $\|\cdot\|_p$.

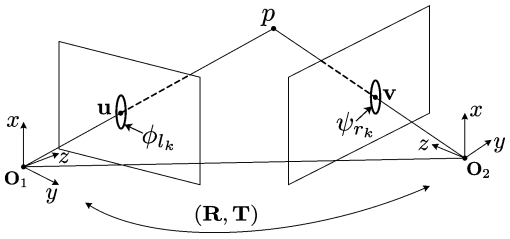


Fig. 1. Epipolar geometry between stereo atoms.

$$[Q_{lr}(\mathbf{v})]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{v} = 0. \quad (5)$$

Matrix $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{v}$ as matrix multiplication. Clearly, Q_{lr} denotes the linear transform of coordinates between atoms $\boldsymbol{\phi}_l = g_{\gamma_l^{(L)}}$ and $\boldsymbol{\psi}_r = g_{\gamma_r^{(R)}}$, thus it depends on the parameters $\gamma_l^{(L)}$ and $\gamma_r^{(R)}$. Knowing these parameters, it is straightforward to derive the analytic form for Q_{lr}. This form is derived in [3], [21] for omnidirectional images.

Note however that the epipolar constraint in (5) is rarely satisfied exactly, due to discrete spatial sampling of images. It can be only evaluated with a certain error (distance) ε_l. The estimated epipolar constraint c_l is thus given as

$$c_l = [Q_{lr}(\mathbf{v})]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{v} + \varepsilon_l = d_l + \varepsilon_l \quad (6)$$

where $d_l = [Q_{lr}(\mathbf{v})]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{v}$. Moreover, since there is uncertainty in epipolar geometry estimation, the epipolar measure is not symmetric. When a point $\mathbf{u}$ in the second image is transformed to a point $\mathbf{v}$ in the first image, we have the epipolar geometry estimate c_r given by

$$c_r = \left[Q_{lr}^{-1}(\mathbf{u})\right]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{u} + \varepsilon_r = d_r + \varepsilon_r. \quad (7)$$

where $d_r = \left[Q_{lr}^{-1}(\mathbf{u})\right]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{u}$. The most likely transforms Q_{lr} (equivalently F_{lr}) in pairs of stereo images are the transforms that give small average epipolar error over all corresponding pairs of points $(\mathbf{u}, \mathbf{v})$. We use this property in the maximum-likelihood learning of stereo dictionaries presented in Section IV.

IV. ML Learning of Dictionaries for Stereo Images

A. Problem Formulation

Following a similar approach as in [4], we formulate the probabilistic framework for the ML learning of overcomplete dictionaries $\bar{\boldsymbol{\Phi}}$, $\bar{\boldsymbol{\Psi}}$ that are used to represent the stereo images $\mathbf{y}_L$ and $\mathbf{y}_R$ respectively. We define the likelihood that stereo images captured by two cameras with a relative pose $\mathbf{R}, \mathbf{T}$ are well represented with a small set of atom pairs related by multi-view geometry constraints. In other words, we want to learn the dictionaries $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ simultaneously. Therefore, we need to maximize the probability that the observed stereo images $\mathbf{y}_L$ and $\mathbf{y}_R$ are well represented by dictionaries $\bar{\boldsymbol{\Phi}}$ and $\bar{\boldsymbol{\Psi}}$ under a sparsity prior, and that the epipolar constraint between all corresponding points on $\mathbf{y}_L$ and $\mathbf{y}_R$ is satisfied, i.e., that the total epipolar distance D between all corresponding points equal to zero, $D = 0$. Therefore, the goal of learning is to find the overcomplete dictionaries $\bar{\boldsymbol{\Phi}}^*$ and $\bar{\boldsymbol{\Psi}}^*$ that are the solutions of the following optimization problem:

$$(\boldsymbol{\Phi}, \boldsymbol{\Psi})^* = \arg\max_{\boldsymbol{\Phi},\boldsymbol{\Psi}} [\log P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \boldsymbol{\Phi}, \boldsymbol{\Psi})] \quad (8)$$

where $\mathbf{y}_L = \boldsymbol{\Phi}\mathbf{a} + \mathbf{e}_L$ and $\mathbf{y}_R = \boldsymbol{\Psi}\mathbf{b} + \mathbf{e}_R$. When images $\mathbf{y}_L$ and $\mathbf{y}_R$ have sparse approximations in $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, respectively, we can approximate the probability in (8) as

$$P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \boldsymbol{\Phi}, \boldsymbol{\Psi}) \approx P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) P(\mathbf{a}, \mathbf{b} | \boldsymbol{\Phi}, \boldsymbol{\Psi}). \quad (9)$$

A similar approximation is proposed in [10] in the learning of dictionaries for natural images. Applying the chain rule and using the fact that $D = 0$ does not bring more information to $\mathbf{y}_L, \mathbf{y}_R$ than $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, we obtain

$$P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \boldsymbol{\Phi}, \boldsymbol{\Psi}) \approx P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) \cdot P(D = 0 | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) P(\mathbf{a}, \mathbf{b} | \boldsymbol{\Phi}, \boldsymbol{\Psi}). \quad (10)$$

We can now rewrite the optimization problem of (8) as

$$(\boldsymbol{\Phi}, \boldsymbol{\Psi})^* = \arg\max_{\boldsymbol{\Phi},\boldsymbol{\Psi}} [\max_{\mathbf{a},\mathbf{b}} (\log P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) \cdot P(D = 0 | \mathbf{a}, \mathbf{b}, \bar{\boldsymbol{\Phi}}, \bar{\boldsymbol{\Psi}}) P(\mathbf{a}, \mathbf{b} | \bar{\boldsymbol{\Phi}}, \bar{\boldsymbol{\Psi}}))] \quad (11)$$

where the objective function consists of three components: 1) the stereo image likelihood $P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$; 2) the epipolar likelihood $P(D = 0 | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$; and 3) the prior on the coefficients $P(\mathbf{a}, \mathbf{b} | \boldsymbol{\Phi}, \boldsymbol{\Psi})$. In the following, we evaluate each one of these three terms of the objective function.

B. The Stereo Image Likelihood

The stereo image likelihood $P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ can be modeled by a Gaussian white noise that describes the approximation error in the image model in (1). We thus have

$$P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = P(\mathbf{e}_L + \mathbf{e}_R) \propto \exp\left(-\frac{1}{2\sigma_I^2}(\|\mathbf{y}_L - \boldsymbol{\Phi}\mathbf{a}\|_2^2 + \|\mathbf{y}_R - \boldsymbol{\Psi}\mathbf{b}\|_2^2)\right) \quad (12)$$

where σ_I^2 is the variance of the Gaussian noise. Note that we have used in (12) the fact that the sum of two zero-mean Gaussian random variables is also a zero-mean Gaussian random variable. In the rest of the paper, we omit the normalization constants of distributions since they do not influence the optimization problem in (8), and we use $\propto$ instead of $=$.

C. The Epipolar Matching Likelihood

We compute now the likelihood that the epipolar constraint D is equal to zero given the stereo image model in (2), i.e., $P(D = 0 | \mathbf{a}, \mathbf{b}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$. This likelihood prioritizes the atoms that correspond to the same 3-D objects in our stereo image model. The epipolar matching of two points on the left and right images, which are the projections of the same point in the 3-D space, has been derived in the Section III-B. The estimation errors ε_l and ε_r of the epipolar constraints c_l and c_r are assumed to be i.i.d. white zero-mean Gaussian noises of variances σ_l^2 and σ_r^2, respectively. We can thus define the conditional probabilities of

the random variables c_l and c_r, given a pair of points $\mathbf{v}$, $\mathbf{u}$ and atoms $\boldsymbol{\phi}_l$, $\boldsymbol{\psi}_r$ as

$$P(c_l|\mathbf{v},\mathbf{u},\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) = P(c_l|d_l) \propto \exp\left(-\frac{(c_l-d_l)^2}{2\sigma_l^2}\right)$$
$$P(c_r|\mathbf{v},\mathbf{u},\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) = P(c_r|d_r) \propto \exp\left(-\frac{(c_r-d_r)^2}{2\sigma_r^2}\right).$$

The atoms $\boldsymbol{\phi}_l$ and $\boldsymbol{\psi}_r$ influence the functions d_l and d_r that are based on the transform Q_{lr}. Since we want to find the probability that two atoms satisfy the epipolar constraint, we are interested in the probability of a particular realization of the random variables c_l and c_r when they are simultaneously equal to zero. Therefore, we define the conditional probability of $c_l = 0$, $c_r = 0$, given $\mathbf{v}$, $\mathbf{u}$, $\boldsymbol{\phi}_l$, $\boldsymbol{\psi}_r$

$$\begin{aligned} &P(c_l=0,c_r=0|\mathbf{v},\mathbf{u},\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) \\ &\quad = P(c_l=0|\mathbf{v},\mathbf{u},\boldsymbol{\phi}_l,\boldsymbol{\psi}_r)P(c_r=0|\mathbf{v},\mathbf{u},\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) \\ &\quad \propto \exp\left(-\frac{d_l^2}{2\sigma_l^2}\right)\exp\left(-\frac{d_r^2}{2\sigma_r^2}\right). \end{aligned} \tag{13}$$

We extend this conditional probability to all pairs of pixels that undergo the transform defined by the atom pair, i.e., for all $\mathbf{v}_i$, $\mathbf{u}_i$, $i = 1,\ldots,q$. Let D_{lr} denote the event that the epipolar constraints are simultaneously respected for all pairs of pixels. In this case, we have $c_l^{[i]} = 0, c_r^{[i]} = 0$ for all $i = 1,\ldots,q$, where $c_l^{[i]} = d_l^{[i]} + \varepsilon_l = [Q_{lr}(\mathbf{v}_i)]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{v}_i + \varepsilon_l$ and $c_r^{[i]} = d_r^{[i]} + \varepsilon_l = \left[Q_{lr}^{-1}(\mathbf{u}_i)\right]^\top \hat{\mathbf{T}}\mathbf{R}\mathbf{u}_i + \varepsilon_r$. Note that small letters within square brackets in the superscript denote the pixel counter parameter. If we assume that the estimation of the epipolar distance for each pixel pair i can be computed independently, we have

$$\begin{aligned} &P(D_{lr}=0|\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) \\ &\quad = \prod_{i=1}^{q} P\left(c_l^{[i]}=0, c_r^{[i]}=0|\mathbf{v}_i,\mathbf{u}_i,\boldsymbol{\phi}_l,\boldsymbol{\psi}_r\right) \\ &\quad \propto \prod_{i=1}^{q} \exp\left(-\frac{\left(d_l^{[i]}\right)^2}{\left(2\sigma_l^{[i]}\right)^2}\right)\exp\left(-\frac{(d_r^{[i]})^2}{2\left(\sigma_r^{[i]}\right)^2}\right) \\ &\quad = \exp\left(-\frac{1}{2\sigma_D^2}\sum_{i=1}^{q}\left(w_l^{[i]}\left(d_l^{[i]}\right)^2 + w_r^{[i]}\left(d_l^{[i]}\right)^2\right)\right) \\ &\quad = \exp\left(-\frac{1}{2\sigma_D^2}W_{lr}\right) \end{aligned} \tag{14}$$

where $w_l^{[i]} = \sigma_D^2/\left(\sigma_l^{[i]}\right)^2$ and $w_r^{[i]} = \sigma_D^2/\left(\sigma_r^{[i]}\right)^2$. The weights $w_l^{[i]}$, $w_r^{[i]}$ permit to control the importance of the epipolar constraints. In particular, they give more importance to the epipolar constraint at points that are closer to the geometric discontinuity represented by the atom, where the estimation of the epipolar constraint is more reliable. This is represented by the function W_{lr} in (14) in order to simplify the notation.

Finally, the probability of the epipolar matching for the stereo image pair is the product of probabilities of epipolar matching for pairs of active atoms. The active atoms participate in the sparse decompositions of the left and right image with their respective coefficients a_l and b_r, which are different from zero. Then, we can model the probability $P(D=0|\mathbf{a},\mathbf{b},\boldsymbol{\Phi},\boldsymbol{\Psi})$ as

$$\begin{aligned} &P(D=0|\mathbf{a},\mathbf{b},\boldsymbol{\Phi},\boldsymbol{\Psi}) \\ &\qquad \propto \exp\left(-\frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)W_{lr}\right) \end{aligned} \tag{15}$$

where $\mathcal{I}$ is the indicator function defined as: $\mathcal{I}(x) = 1$ if $x \neq 0$ and $\mathcal{I}(x) = 0$ if $x = 0$.

D. Prior on the Coefficient Vector Distributions

The last term in our objective function in (11) is the joint distribution $P(\mathbf{a},\mathbf{b}|\boldsymbol{\Phi},\boldsymbol{\Psi})$ of the coefficients $\mathbf{a}$ and $\mathbf{b}$, given the dictionaries $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. We assume that the pairs of coefficients (a_l, b_r) are pairwise independent, which is usually the case when the image approximations are sparse enough. Then, the distribution $P(\mathbf{a},\mathbf{b}|\boldsymbol{\Phi},\boldsymbol{\Psi})$ becomes factorial, and we can write

$$\begin{aligned} P(\mathbf{a},\mathbf{b}|\boldsymbol{\Phi},\boldsymbol{\Psi}) &= \prod_{l=1}^{M}\prod_{r=1}^{M} P(a_l,b_r|\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) \\ &= \prod_{l=1}^{M}\prod_{r=1}^{M} P(b_r|a_l,\boldsymbol{\phi}_l,\boldsymbol{\psi}_r)P(a_l) \\ &= \prod_{l=1}^{M}\prod_{r=1}^{M} P(a_l|b_r,\boldsymbol{\phi}_l,\boldsymbol{\psi}_r)P(b_r). \end{aligned} \tag{16}$$

We compute now the conditional probabilities and the distributions of the coefficients that are used in (16). We assume that pixels keep their intensity values under the local transforms induced by the viewpoint change. This assumption holds in multiview images when the scene is Lambertian, and when the atom transforms correctly represent local transforms. Equivalently, we can write

$$\begin{aligned} \forall\, k \text{ and } \forall\, \mathbf{v} \text{ s.t. } \boldsymbol{\phi}_{l_k}(\mathbf{v}) \neq 0, \Rightarrow \mathbf{y}_L(\mathbf{v}) &= \mathbf{y}_R(Q_{l_k r_k}(\mathbf{v})) \\ &= \mathbf{y}_R(\mathbf{u}) \end{aligned} \tag{17}$$

where $\mathbf{u} = Q_{l_k r_k}(\mathbf{v})$. This means that if the transform $Q_{l_k r_k}$ maps a pixel at position $\mathbf{v}$ on the image $\mathbf{y}_L$ into a pixel at position $\mathbf{u}$ on the image $\mathbf{y}_R$, then those pixels have the same intensity. Under the assumption given in (17), we can use the Lemma 1 in [22], which states the following equality:

$$\langle \mathbf{y}_R, \boldsymbol{\psi}_{r_k}\rangle = \frac{1}{\sqrt{J_{l_k r_k}}}\langle \mathbf{y}_L, \boldsymbol{\phi}_{l_k}\rangle \tag{18}$$

where $J_{l_k r_k} = |\partial\mathbf{u}/\partial\mathbf{v}| = |\partial Q_{l_k r_k}(\mathbf{v})/\partial\mathbf{v}|$ is the Jacobian determinant (or simply the Jacobian) of the linear transform $Q_{l_k r_k}$. Recall that the inner products in (18) correspond to the coefficients b_r and a_l related to the atoms under consideration. Using the sparse image model and the relation in (18) we obtain the following probabilities:

$$\begin{aligned} &P(b_r|a_l,\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) = P(a_l|b_r,\boldsymbol{\phi}_l,\boldsymbol{\psi}_r) \\ &\qquad \propto \exp\left(-\frac{1}{2\sigma_b^2}\left(b_r - \frac{a_l}{\sqrt{J_{lr}}}\right)^2\right) \end{aligned} \tag{19}$$

where σ_b is the standard deviation of the zero-mean Gaussian noise that models the difference between b_r and $a_l/\sqrt{J_{lr}}$. The detailed derivation of (19) is given in Appendix A.

We model now the prior distributions of the coefficients. These distributions depend on an arbitrarily chosen dictionary. However, imposing the independence of the coefficients with respect to the dictionary during learning leads to a dictionary that gives the same prior distribution of coefficients for all types of images. Furthermore, when the prior on coefficients is tightly peaked at zero, the learning leads to a universal dictionary in which all natural images have sparse decompositions. We choose here a different approach than the one proposed in [4], where the coefficients prior is a heavy-tailed Laplace distribution, peaked at zero. Instead, we assume that the coefficients a_l and b_r are drawn from a Bernoulli distribution over the activity of coefficients $\mathcal{I}(a_l)$ and $\mathcal{I}(b_r)$, where $\mathcal{I}$ denotes the indicator function. For $\mathcal{I}(a_l)$ this distribution is

$$P(a_l) = \begin{cases} p & \text{if } \mathcal{I}(a_l) = 1; \\ q & \text{if } \mathcal{I}(a_l) = 0 \end{cases}$$

and similar holds for $\mathcal{I}(b_r)$. Choosing $p \ll q$ introduces a sparsity assumption on the coefficients. These prior distributions represent the case when a large number of coefficients is exactly zero, which is a better sparsity prior than the Laplace distribution where many coefficients are close to, but not exactly equal to zero. When the images are represented by m atoms from a dictionary of size M, we have

$$P(\mathbf{a}) = \prod_{l=1}^{M} P(a_l) = p^m (1-p)^{(M-m)} \tag{20}$$

and the same holds for $P(\mathbf{b})$. Without loss of generality, we pose $p = 1/(1+e^{1/\lambda})$. Therefore, reducing the value of λ increases the level of "sparseness" of coefficients. As the coefficients a_l and b_r for $l, r = 1, \ldots, M$ are independent and identically distributed, the (20) can be rewritten as

$$P(\mathbf{a}) = \left(\frac{e^{1/\lambda}}{1+e^{1/\lambda}}\right)^M \exp\left(\frac{-m}{\lambda}\right) \propto \exp\left(-\frac{\|\mathbf{a}\|_0}{\lambda}\right). \tag{21}$$

A similar expression holds for $\mathbf{b}$, i.e., $P(\mathbf{b}) \propto \exp(-\|\mathbf{b}\|_0/\lambda)$. From (16), (19) and (21), we can write

$$P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) \propto \exp\left(-\frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right) \exp\left(-\frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0)\right) \tag{22}$$

where the term in the second exponent follows from the assumption that the left and the right image are of the same sparsity, i.e., $\|\mathbf{a}\|_0 = \|\mathbf{b}\|_0$.

We have now defined all three components of our objective function in (11). The Section V proposes an approach for numerical optimization of the objective function using the Expectation-Maximization algorithm.

V. EM-Based Energy Minimization Algorithm

A. Energy Minimization Problem

The ML optimization problem of (11) is equivalent to solving the following energy minimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}} \left[\min_{\mathbf{a},\mathbf{b}} E(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})\right] \tag{23}$$

where E denotes the energy function given as

$$\begin{aligned} E \equiv & E(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) \\ = & \frac{1}{2\sigma_I^2}\left(\|\mathbf{y}_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|\mathbf{y}_R - \mathbf{\Psi}\mathbf{b}\|_2^2\right) \\ & + \frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)W_{lr} \\ & + \frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\left(b_r - \frac{a_l}{\sqrt{J_{lr}}}\right)^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \end{aligned} \tag{24}$$

The energy function thus consists in the sum of four main terms that are respectively

1) the data fidelity term, expressed by the energy of the approximation error after sparse approximation of images $\mathbf{y}_L$ and $\mathbf{y}_R$,
2) the epipolar constraint term, measuring the epipolar matching of atoms in sparse decompositions of a stereo image pair,
3) the coefficient similarity term, measuring the correlation of coefficients of stereo atom pairs under a local transform,
4) the sparsity term, expressing the degree of sparsity of the stereo image pair.

We can see that the first three terms depend on the choice of the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, while the last term depends only on the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. Therefore, we group the first three terms into a function $f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$ and express the energy function as

$$E = f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{25}$$

B. EM-Based Algorithm

The energy minimization problem can be solved iteratively by alternating between two steps. In the first step, $(\mathbf{\Phi}, \mathbf{\Psi})$ is kept constant and the energy function is minimized with respect to the coefficients $(\mathbf{a}, \mathbf{b})$. The second step keeps the obtained coefficients $(\mathbf{a}, \mathbf{b})$ constant, while performing the gradient descent on $(\mathbf{\Phi}, \mathbf{\Psi})$ to minimize the energy E, which is equivalent to minimizing $f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$. Therefore, the algorithm iterates between the sparse coding and the dictionary learning steps until convergence. In order to have stable learning, each iteration of the proposed algorithm has to be performed on a set of stereo images $\mathcal{S} = \{(\mathbf{y}_L, \mathbf{y}_R)_p\}_{p=1,\ldots,S_p}$ randomly chosen from a large database of stereo pairs taken from different camera poses. This

algorithm is then equivalent to an Expectation-Maximization algorithm [23], where the images $(\mathbf{y}_L, \mathbf{y}_R)$ are the observed variables, $(\mathbf{a}, \mathbf{b})$ represent hidden variables, and $(\mathbf{\Phi}, \mathbf{\Psi})$ are the parameters [24], [25]. The Expectation (E) step corresponds to the sparse coding (inference) step, while the Maximization (M) step corresponds to the learning step.

1) Minimization With Respect to the Coefficients: In the sparse coding step, for each stereo image pair $(\mathbf{y}_L, \mathbf{y}_R)$ in a randomly chosen set $\mathcal{S}$ we solve for $\mathbf{a}$ and $\mathbf{b}$

$$(\mathbf{a}, \mathbf{b})^* = \arg\min_{\mathbf{a},\mathbf{b}} f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{26}$$

We can see that this problem is similar to the constrained l_0 sparse approximation problem when cast as an unconstrained problem. The multiplier $1/2\lambda$ is the trade-off parameter between minimizing the energy in f and the sparsity of coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. Since finding the global optimum of such a problem is NP-hard, we use the greedy approach to find a locally optimal solution. Although they are not guaranteed to find the sparsest solution for such problems, greedy algorithms have performed quite well with fast convergence in practice [26]–[28]. An advantage of using a greedy approach here is that it leads to a signal approximation using a small set of coefficients different from zero. In contrary, l_1 minimization algorithms leads to many small but non-zero coefficients, which would increase the computation time of the epipolar constraint part of the energy function in (24).

We propose a greedy algorithm that chooses at each iteration k the pair of atoms $\boldsymbol{\phi}_{l_k}, \boldsymbol{\psi}_{r_k}$ that give the minimal value of the function

$$\begin{aligned}(\boldsymbol{\phi}_{l_k}, \boldsymbol{\psi}_{r_k}) = \arg\min_{\boldsymbol{\phi}_l, \boldsymbol{\psi}_r} \Bigg[&\frac{1}{2\sigma_I^2}\left(\left\| \mathbf{h}_l^{[k-1]} - \left\langle \mathbf{h}_l^{[k-1]}, \boldsymbol{\phi}_l \right\rangle \boldsymbol{\phi}_l \right\|_2^2 \right. \\ &\left. + \left\| \mathbf{h}_r^{[k-1]} - \left\langle \mathbf{h}_r^{[k-1]}, \boldsymbol{\psi}_r \right\rangle \boldsymbol{\psi}_r \right\|_2^2 \right) + \frac{1}{2\sigma_D^2} W_{lr} \\ &+ \frac{1}{2\sigma_b^2}\left(\left\langle \mathbf{h}_r^{[k-1]}, \boldsymbol{\psi}_r \right\rangle - \frac{\left\langle \mathbf{h}_l^{[k-1]}, \boldsymbol{\phi}_l \right\rangle}{J_{lr}} \right)^2 \Bigg] \end{aligned} \tag{27}$$

where $\mathbf{h}_l^{[k-1]}$ and $\mathbf{h}_r^{[k-1]}$ are the residues of the left and right images respectively, after $k-1$ iterations[2]. At the beginning, the residues are: $\mathbf{h}_L^{[0]} = \mathbf{y}_L$ and $\mathbf{h}_R^{[0]} = \mathbf{y}_R$ and they are updated at each step k as

$$\begin{aligned} \mathbf{h}_L^{[k]} &= \mathbf{h}_L^{[k-1]} - \left\langle \mathbf{h}_L^{[k-1]}, \boldsymbol{\phi}_{l_k} \right\rangle \boldsymbol{\phi}_{l_k} \\ \mathbf{h}_R^{[k]} &= \mathbf{h}_R^{[k-1]} - \left\langle \mathbf{h}_R^{[k-1]}, \boldsymbol{\psi}_{r_k} \right\rangle \boldsymbol{\psi}_{r_k}. \end{aligned} \tag{28}$$

The coefficients a_{l_k} and b_{r_k} are simply evaluated as

$$\begin{aligned} a_{l_k} &= \left\langle \mathbf{h}_L^{[k-1]}, \boldsymbol{\phi}_{l_k} \right\rangle \\ b_{r_k} &= \left\langle \mathbf{h}_R^{[k-1]}, \boldsymbol{\psi}_{r_k} \right\rangle. \end{aligned} \tag{29}$$

[2]We introduce a slight abuse of notation here, as $[\cdot]$ in the superscript hold an iteration counter, while in Section IV-C they hold a pixel index.

We will refer to this algorithm as Multi-view Matching Pursuit (MVMP)[3], which can be shown to be essentially the Weak Matching Pursuit [29]. The bound of the approximation rate of MVMP can be found in [30].

2) Minimization With Respect to the Atom Scale Parameters: Once the atoms $\boldsymbol{\phi}_{l_k}, \boldsymbol{\psi}_{r_k}, k = 1, \ldots, m$ that participate in the decomposition of images $\mathbf{y}_L$ and $\mathbf{y}_R$ have been found by MVMP, their coefficients are kept fixed while the atoms are updated by minimizing the energy function. Knowing the selected atoms, the energy function at step t becomes

$$\begin{aligned} E^{[t]} = &\frac{1}{2\sigma_I^2}\left(\left\| \mathbf{y}_L - \sum_{k=1}^{m} a_{l_k} \boldsymbol{\phi}_{l_k} \right\|_2^2 + \left\| \mathbf{y}_R - \sum_{k=1}^{m} b_{r_k} \boldsymbol{\psi}_{r_k} \right\|_2^2 \right) \\ &+ \frac{1}{2\sigma_D^2} \sum_{k=1}^{m} W_{lr} + \frac{1}{2\sigma_b^2} \sum_{k=1}^{m} \left(b_{r_k} - \frac{a_{l_k}}{\sqrt{J_k}} \right)^2 \\ &+ \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \end{aligned} \tag{30}$$

Since we have a set $\mathcal{S}$ of stereo images, in the M step we minimize over the average energy $\langle E^{[t]} \rangle_{\mathcal{S}}$. It can be observed that the energy function $E^{[t]}$ is actually an analytic function of the atom parameters $\left\{ \gamma_l^{(L)} \right\}$ and $\left\{ \gamma_r^{(R)} \right\}$ in the case of parametric dictionaries as defined in Section III-A. Hence, we can calculate its derivatives with respect to each parameter. Therefore, one can use the multivariate gradient descent, or the multivariate conjugate gradient method to find the local minimum of $\langle E^{[t]} \rangle_{\mathcal{S}}$ with respect to $\gamma_l^{(L)}$ and $\gamma_r^{(R)}$, given the coefficients $\mathbf{a}$ and $\mathbf{b}$ for each $(\mathbf{y}_L, \mathbf{y}_R) \in \mathcal{S}$.

The sparse coding and the learning steps are iteratively repeated until convergence is achieved. The complexity of the stereo dictionary learning algorithm is twice larger than the complexity of the same algorithm that learns from single images, augmented by the cost of evaluating the epipolar geometry between atoms. However, learning parametric dictionaries has a much smaller cost than learning general type dictionaries [4], [6], as the number of free parameters is drastically reduced from the number of pixels in all atoms to the number of possible atom parameters (different scales in our case). The complexity of learning a sufficient number of atom parameters from stereo images is thus comparable to the complexity of dictionary learning for single images.

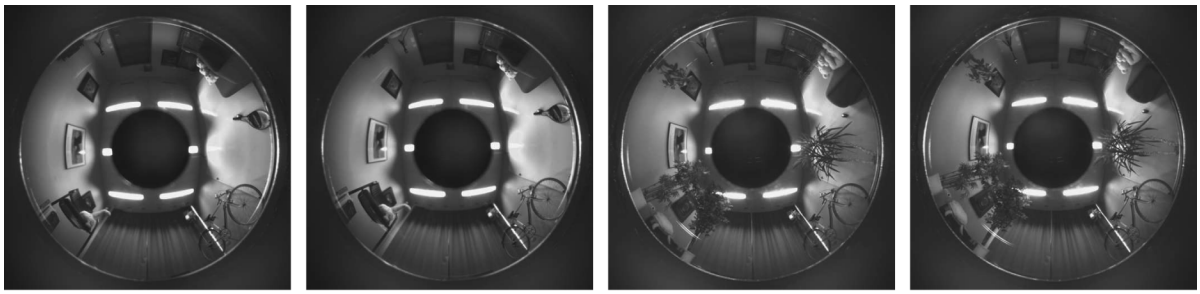
VI. Learning for Stereo Omnidirectional Images

A. Spherical Imaging Framework

The proposed ML stereo dictionary learning method does not put any assumption on the type of cameras used for stereo image acquisition. Since omnidirectional cameras with wide fields of view represent a convenient solution for 3-D scene representation, we perform learning of dictionaries for stereo omnidirectional images.

Omnidirectional images obtained by catadioptric cameras can be appropriately mapped to spherical images [30]. Therefore, we use a dictionary of atoms on the 2-D unit sphere as proposed in [31] in order to represent spherical images. The

[3]Although we take here only two images, we can generalize the algorithm to more than two images by pairwise image correspondence.



(a) (b) (c) (d)

Fig. 2. (a)–(b) Two views from the "Mede" database, first image set. (c)–(d) Two views from the "Mede" database, second image set.

generating function g is defined in the space of square-integrable functions on a unit two-sphere S^2, $g(\theta,\varphi) \in L^2(S^2)$, where θ is the polar angle and φ is the azimuth angle. We use a dictionary of edge-like atoms on the sphere, based on a generating function that is a Gaussian in one direction and its second derivative in the orthogonal direction

$$g(\theta,\varphi) = -\frac{1}{K_A}\left(16\alpha^2\tan^2\frac{\theta}{2}\cos^2\varphi - 2\right) \exp\left(-4\tan^2\frac{\theta}{2}(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi)\right) \quad (31)$$

where K_A is a normalization constant. For the weighting function in the epipolar geometry constraints of (14) we use a Gaussian envelope of the form

$$w(\theta,\varphi) = \frac{1}{K_G}\exp\left(-4\tan^2\frac{\theta}{2}(12\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi)\right). \quad (32)$$

This function gives positive weights to the points on the main (central) lobe of the atom $g(\theta,\varphi)$, while the weights are close to zero outside of the main lobe. The weights are higher towards the axis of the discontinuity represented by the atom. Choosing such a weighting function for the epipolar geometry estimation permits to use only the points that are likely to satisfy the epipolar constraints, and to exclude the points represented by the ripples of the second derivative of the Gaussian. The dictionary is then built by changing the atom parameters $\gamma = (\tau,\nu,\zeta,\alpha,\beta) \in \Gamma$. The triplet (τ,ν,ζ) represents Euler angles that describe the motion of the atom on the sphere by angles τ and ν along θ and φ respectively, and the rotation of the atom around its axis with an angle ζ. The parameters α and β represent anisotropic scaling factors.

The epipolar geometry constraint for the considered stereo image model is given in (5). The transform $\mathbf{u} = Q_{lr}(\mathbf{v})$ that relates a point $\mathbf{v}$ on atom $\boldsymbol{\phi}_l = g_{\gamma_l^{(L)}}$ to its corresponding transformed point $\mathbf{u}$ on the transformed atom $\boldsymbol{\psi}_r = g_{\gamma_r^{(R)}}$ can be defined via the linear transform of the coordinate system. The closed form of this transform can be found in [3], [30].

B. Learning Testbed

We describe now the experimental testbed that we have used for dictionary learning on stereo spherical images. Even if the dictionary is constructed by translation, rotation and scaling of the generating function, we focus on the learning of scaling parameters only. The scaling parameters are the most important parameters since they define the shape of the atoms, which become elongated as the scales become anisotropic. Alternatively, translations and rotations depend highly on the distance and orientations of the cameras, so that learning these parameters is meaningful only when cameras are static (it results in a position-specific dictionary). We want to perform learning of scales $(\alpha^{(L)},\beta^{(L)})$ from the set of atom parameters $\gamma^{(L)}$ and $(\alpha^{(R)},\beta^{(R)})$ from $\gamma^{(R)}$ for atoms that are present in sparse approximations of stereo views and satisfy the epipolar geometry constraint. The energy function of (24) is minimized only with respect to these parameters. We have performed the minimization using the conjugate gradient[4]. The other parameters are kept fixed. The motion parameters (τ,ν) include the positions of all pixels in an image. The rotation parameter ζ has been sampled uniformly between 0 and π with resolution N_r. We have taken the same motion and rotation parameters for the left and right dictionaries.

We have tested the proposed stereo dictionary learning algorithm on our "Mede" omnidirectional multi-view database[5]. The database consists of 54 omnidirectional images of the indoor environment, grouped into two sets: set without plants (27 images), and set with plants (27 images). Different views have been captured by placing cameras on different positions on the floor, without camera rotation. We have formed 216 pairs of images with different distances between the cameras. Since we know the camera positions and the rotation is identity, the relative pose matrices $\mathbf{T}$ and $\mathbf{R}$ are known for each image pair. Sample views are illustrated in Fig. 2.

The first step in the learning algorithm (i.e., the expectation (E) step implemented by MVMP) needs to be performed on a big set of statistically different stereo images for the learning results to be meaningful. In order to limit the complexity of the whole learning process and yet include the image diversity, we select small patches of $N_c \times N_c$ pixels from the spherical images obtained by mapping the omnidirectional images to the unit sphere. As cropping a square image patch from a spherical image is feasible only when its center lies on the equator, we rotate the sphere such that the center of the patch coincides with the equator and then crop it. This rotation is taken into account when we estimate the epipolar geometry. Therefore, in the E

[4]http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/

[5]Database is available upon request.

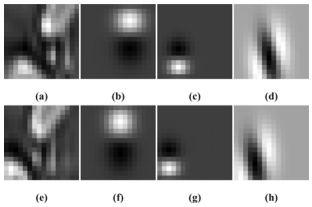


Fig. 3. Example of a pair of stereo patches and their MVMP selected atoms. (a) The left patch, (b)-(d) the first three atoms in the MVMP decomposition of the left patch; (e) right patch, (f)-(h) the first three atoms in the MVMP decomposition of the right patch.

step we form a set $\mathcal{S}$ of S_p pairs of stereo patches. For each $p = 1, \ldots, S_p$ we randomly choose an image pair from the database, then we randomly select a point on the sphere and we extract two patches from two stereo images centered on this point. The MVMP is then performed on each pair of patches independently, and N_{at} atoms are selected. Examples of atoms are shown in Fig. 3(b)–(d) and (f)–(h). The dictionary learning step (M step) is then performed by minimizing the sum of the energy function given by (30) for all patches.

In our experiments, we have taken $S_p = 50$ pairs of patches of size 12×12, that have been obtained by cropping slightly bigger patches of 16×16 in order to avoid border effects. All patches have been normalized to the same variance 0.1 in order to equalize the importance of each patch. Moreover, the patches have been whitened by spherical low-pass filtering in order to flatten the image spectrum and make all frequencies equally important, as proposed in [4]. The number of positions in the dictionary construction is 12×12 and the number of rotations is set to 4. Finally, the pairs of scales have been independently randomly initialized for the left and right dictionary, with 5 pairs of anisotropic scales each in the range $[5, 15]$ (from big to small atoms). The MVMP algorithm selects $N_{at} = 3$ atoms per patch. Since the size of the patches is small, three atoms are usually enough to represent the main geometrical components in the patch. Before starting the learning algorithm, we have performed MVMP on a set of randomly selected patches to estimate the variances $\sigma_D^2 = 2.7 \cdot 10^{-3}$ and $\sigma_b^2 = 0.047$.

C. Learned Dictionaries

We first present the resulting dictionaries obtained by the proposed learning method applied to stereo omnidirectional images. In order to see the influence of the part of objective function that relies on the multi-view constraint, we have introduced a factor ρ in the energy function

$$\begin{aligned}\tilde{E} = &\frac{1}{2\sigma_I^2}\left(\|\mathbf{y}_L - \mathbf{\Phi a}\|_2^2 + \|\mathbf{y}_R - \mathbf{\Psi b}\|_2^2\right) \\ &+ \rho\frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)W_{lr} \\ &+ \rho\frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\left(b_r - \frac{a_l}{\sqrt{J_{lr}}}\right)^2 \\ &+ \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \end{aligned} \tag{33}$$

We can see that for $\rho = 1$, the energy function in (33) is equal to the one in (24). On the other hand, for $\rho = 0$, there are no multi-view constraints in the energy function, and the dictionary learning is based only on minimization of the approximation error and the sparse penalty.

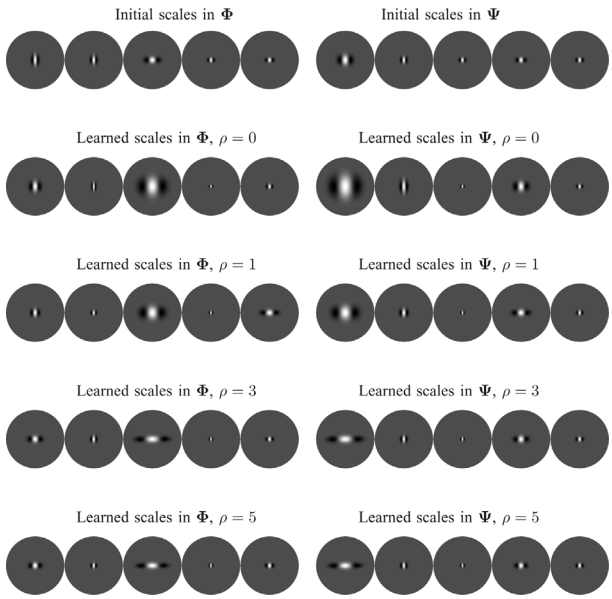


Fig. 4. Initial and learned scales of the atoms for the left and right dictionaries. All atoms are on the North pole.

The initial values of scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$ have been chosen randomly, and they are given in the first two columns of Table I. The atoms built with these initial scales and centered at the North Pole are shown in the first row in Fig. 4. We then learn the scaling parameters with 50 iterations of the EM algorithm. At this point, the change in parameters becomes small and the solution can be considered as stable. The learned scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$ are given in columns 3 to 10 in Table I, for $\rho = 0$, 1, 3, 5. Learning has been initialized with the same initial scales, for all values of ρ. The corresponding atoms are shown in Fig. 4.

We observe first that the learned dictionaries include atoms of different scales; they are therefore able to approximate signals at various scales. When $\rho = 0$, the learned atoms are elongated along the direction of the Gaussian function and narrow in the direction of the second derivative of the Gaussian function. These results are in consistency with the previous work on dictionary learning for image representation in the single view case [4]. However, when we increase ρ and hence include the geometry constraints in the stereo learning, we obtain different results for atoms scales. The atoms become elongated in the direction of the second derivative of the Gaussian function and narrow in the direction of the Gaussian function, which is an opposite effect than in the case where $\rho = 0$. This can also be seen in Fig. 5, which gives a graphical representation of the anisotropy scale ratio β/α for all learned atoms, sorted in the descending order, for $\rho = 0$, 1, 5. Most of the values of β/α (especially the peak values) increase with the increase of ρ, which confirms the anisotropy effect that is visible in Fig. 4. Another effect of the multiview constraint is that for $\rho > 0$ the learned scales generally tend to give smaller atoms than for $\rho = 0$. They are most probably due to the local nature of the epipolar constraint. Namely, the depth of the scene rapidly changes around object

TABLE I
INITIAL AND LEARNED SCALE PARAMETERS FOR THE LEFT AND RIGHT DICTIONARY, FOR DIFFERENT VALUES OF THE PARAMETER ρ

Initial dictionary		Learned dictionary							
		$\rho = 0$		$\rho = 1$		$\rho = 3$		$\rho = 5$	
$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$
13.15	5.98	8.61	6.34	10.82	8.68	6.86	10.17	7.84	10.92
14.06	7.78	22.19	7.30	16.92	13.72	14.84	9.95	16.53	12.36
6.27	10.47	3.40	3.56	3.81	5.05	2.82	10.26	3.17	11.39
14.13	14.58	25.88	22.95	26.00	19.73	25.19	18.21	24.36	17.04
11.32	14.65	14.52	14.78	5.57	11.25	12.73	15.63	11.61	13.92
$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$
6.58	6.42	2.94	2.69	3.58	4.73	2.72	9.36	3.01	10.77
14.71	9.22	12.18	5.04	11.72	8.43	14.79	9.66	15.19	11.00
14.57	14.16	25.93	20.30	25.57	18.94	24.75	17.86	23.94	16.55
9.85	12.92	6.60	6.80	5.70	10.56	6.72	10.13	8.22	11.72
13.00	14.59	15.87	16.05	15.08	14.52	13.08	14.97	13.25	14.85

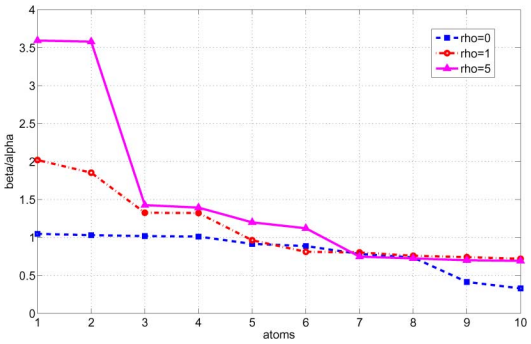


Fig. 5. Anisotropy scale ratio β/α in descending order over all learned atoms, for $\rho = 0, 1, 5$.

boundaries leading to different disparity and epipolar matching. Since the object boundaries are represented by 2-D discontinuities on the image of a 3-D scene, the epipolar geometry is mostly satisfied along the discontinuity and in a limited area. This renders the learned atoms anisotropic and small. It highlights the great importance of the geometric constraints in the learning of dictionaries for stereo images.

VII. APPLICATIONS

A. Distributed Scene Representation

Distributed scene representation with stereo or multi-view cameras corresponds to the problem where each image of the scene is approximated independently from the others. Namely, we do not search for corresponding atoms during sparse approximation using MVMP, but we rather apply MP independently on each image and then match the corresponding atoms for joint reconstruction. The number of correspondences between atoms from different views is very important for scene representation, because these pairs carry the geometric correlation between two images. For example, in distributed multi-view coding, a bigger number of atom stereo pairs directly reduces the required transmission rate and improves the coding performance [3]. If the atoms that form the learned dictionaries $\bar{\mathbf{\Phi}}$ and $\bar{\mathbf{\Psi}}$ represent the statistically optimal atoms for both image approximation and epipolar geometry, one should expect that the learned dictionary results in more corresponding atom pairs than a randomly initialized dictionary, even in distributed settings. In order to verify it, we select randomly two image patches, $\mathbf{y}_L$ and $\mathbf{y}_R$ with the same center from a randomly chosen pair of omnidirectional images in the "Mede" database. The size of the patches is set to 40×40 pixels, which is slightly larger than the size used for learning. After independent MP decompositions of the left and right patch using respectively $\mathbf{\Phi}$ and $\mathbf{\Psi}$ with 10 atoms per patch, the epipolar constraints are evaluated for all possible pairs of left and right atoms $d_A(\boldsymbol{\phi}_l, \boldsymbol{\psi}_r)$, $l = 1, \ldots, 10$, $r = 1, \ldots, 10$. The epipolar measure d_A is equal to half of the value W_{lr} in (14), and represents the epipolar atom distance per view.

Fig. 6(a) plots on the y axis the number of atom pairs $(\boldsymbol{\phi}_l, \boldsymbol{\psi}_r)$ that have the epipolar distance $d_A(\boldsymbol{\phi}_l, \boldsymbol{\psi}_r)$ smaller or equal than the threshold value given on the x-axis. We call this curve the cumulative correspondence number (CCN) curve. The left part of CCN curves with small d_A is more important than the right part, because the correspondences are more reliable when their epipolar distance is smaller. All CCN curves have been averaged over 100 randomly chosen image pairs. We can see that CCN curves for $\rho = 1$, $\rho = 3$ and $\rho = 5$ are all above the CCN curve of the randomly initialized dictionary. This confirms that our learning algorithm produces dictionaries that result in a larger number of correspondences between atoms of different views. On the other hand, for $\rho = 0$, the CCN curve is either close to the CCN curve of the random dictionary, or below it. This shows that designing dictionaries for image approximation without considering the multi-view geometry can lead to suboptimal dictionaries for stereo images. Fig. 6(b) shows the average approximation rate (i.e., energy decay) during the iterations of the MP for images $\mathbf{y}_L$ and $\mathbf{y}_R$. We plot the ratio between the sum of the residues of the left and right images after k iterations and the sum of their initial energies. We can see that for all values of ρ the approximation rate using learned dictionaries is better than using random dictionaries. Moreover, increasing

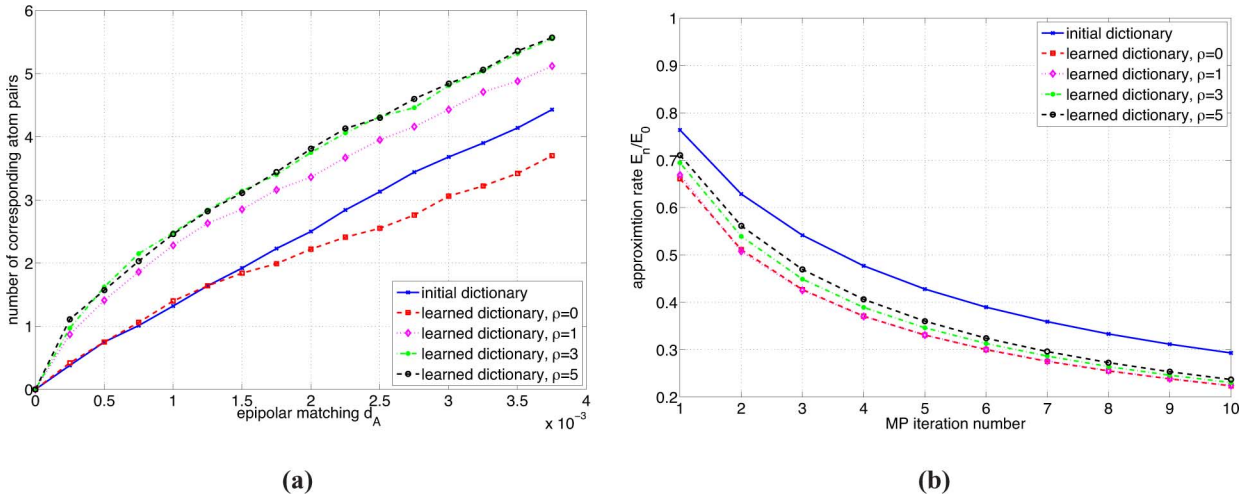


Fig. 6. Performance of the learned dictionaries in distributed settings: (a) Cumulative correspondence number (CCN) curves for initial dictionary and learned dictionaries, for $\rho = 0, 1, 3, 5$; (b) MP energy decay for $\mathbf{y}_L$ and $\mathbf{y}_R$.

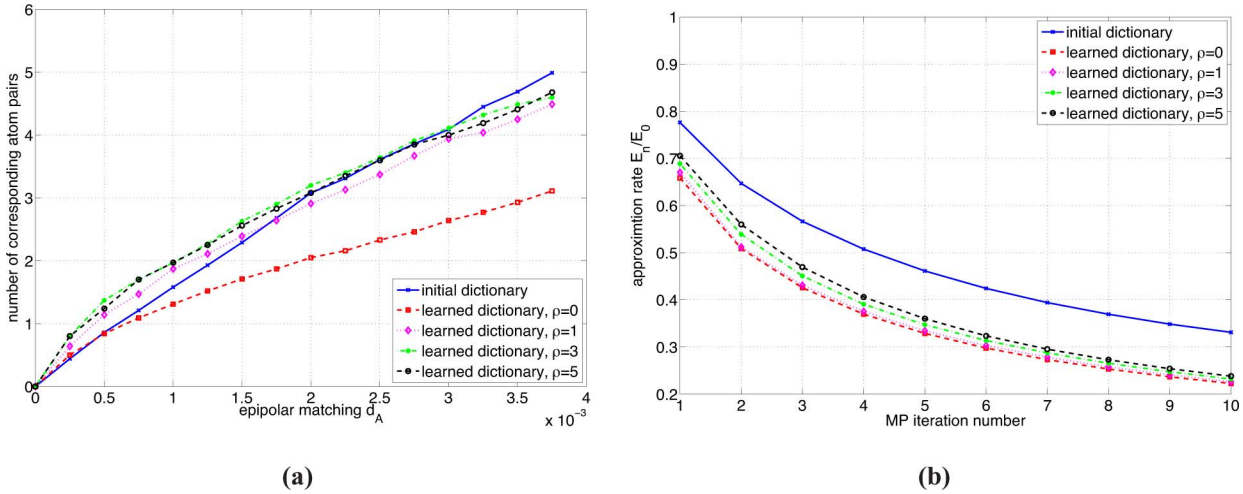


Fig. 7. Performance of the learned dictionaries in distributed settings, with averaging over random initial dictionaries. (a) Average Cumulative correspondence number (CCN) curve for 100 random initial dictionaries and CCN curves for learned dictionaries, for $\rho = 0, 1, 3, 5$; b) MP energy decay for $\mathbf{y}_L$ and $\mathbf{y}_R$.

the value of ρ induces only a minor penalty in the approximation rate; optimizing the dictionaries for stereo matching does not significantly impact the approximation rate.

In order to verify that the superior performance of the learned dictionary over the initial one is not due to the unlucky selection of the initial dictionary, we compare the performance of the learned dictionaries to an average performance of different randomly selected initial dictionaries. We select randomly 100 initial dictionaries and 100 stereo image pairs, and plot the average CCN curve for the initial dictionary. This curve is shown with the blue solid line in Fig. 7(a). We see that the learned dictionaries still give more correspondences than the random ones for small values of d_A. Therefore, they lead to more atom pairs with a better epipolar matching. The comparison of the energy decay in this case is shown in Fig. 7(b), and it is similar to the results in Fig. 6(b).

The learned dictionaries can be beneficial for distributed representation and coding of multi-view images. We evaluate the performance of the distributed coder presented in [3] with new, learned dictionaries. As in [3], we use Lab images for evaluation, which do not belong to the training set "Mede". Since stereo learning results in two dictionaries that are very similar, we use the parameters only of the dictionary $\mathbf{\Phi}$. The distributed coder is based on coding with side information, where image $\mathbf{y}_L$ is independently encoded, while the image $\mathbf{y}_R$ is encoded by coset coding of atom indexes and quantization of their respective coefficients [3]. We use the learned dictionary in Section VI-C for $\rho = 1$ and 16 orientations. The image $\mathbf{y}_L$ is encoded independently at 0.21 bpp with a PSNR of 30.61 dB, while the image $\mathbf{y}_R$ is encoded by coset coding. Fig. 8 shows the rate-distortion (RD) curves for the image $\mathbf{y}_R$, where the dash-dotted line corresponds to the uniformly sampled dictionary [3], and the solid line corresponds to the learned dictionary. The dashed line presents the RD performance of independent coding with MP and the uniform dictionary. We can see that the learned dictionary improves the DSC performance for almost 1 dB at low rate. This confirms our previous observation that the learned dictionary increases the probability of finding correlated atoms between different views. When the number of these correlated atoms decreases, the coder performance satu-

TABLE II
CAMERA POSE ESTIMATION WITH RANDOM INITIAL AND LEARNED DICTIONARIES, FOR TARGET TRANSLATION $\mathbf{T}^\top = [100]$

Target matrices	$\mathbf{T}^\mathsf{T} = [1\ \ 0\ \ 0]$ $\mathbf{R} = \mathbf{I}$								
	learned dictionary			initial dictionary			SIFT		
estimated $\mathbf{T}^\mathsf{T}$, w/o Ransac	[0.9778	0.1519	0.1441]	[0.1910	0.7284	0.6580]	[0.3943	0.7668	0.5065]
error angle [rad] w/o Ransac	**0.2111**			1.3786			1.1655		
estimated $\mathbf{T}^\mathsf{T}$, w/ Ransac	[0.8144	0.5436	0.2032]	[0.7540	0.6067	0.2518]	[0.9708	0.2393	0.0139]
error angle [rad] w/ Ransac	0.6191			0.7167			0.2421		

TABLE III
CAMERA POSE ESTIMATION WITH RANDOM INITIAL AND LEARNED DICTIONARIES, FOR TARGET TRANSLATION $\mathbf{T}^\top = [010]$

Target matrices	$\mathbf{T}^\mathsf{T} = [0\ \ 1\ \ 0]$ $\mathbf{R} = \mathbf{I}$								
	learned dictionary			initial dictionary			SIFT		
estimated $\mathbf{T}^\mathsf{T}$, w/o Ransac	[0.0867	0.9958	0.0310]	[0.8302	0.0484	0.5554]	[0.1048	0.9910	0.0829]
error angle [rad] w/o Ransac	**0.0917**			1.5224			0.1340		
estimated $\mathbf{T}^\mathsf{T}$, w/ Ransac	[0.1143	0.9799	0.1636]	[0.7941	0.4853	0.3661]	[0.4059	0.2053	0.8906]
error angle [rad] w/ Ransac	0.2009			1.0641			1.3640		

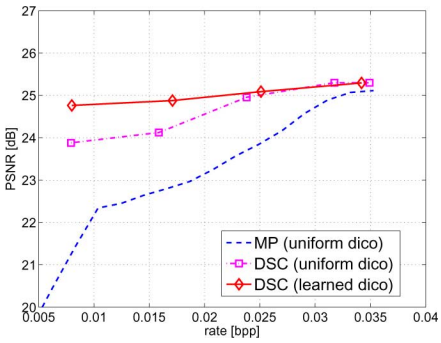


Fig. 8. Rate-distortion performance of the DSC coder for image $\mathbf{y}_R$ from the Lab image dataset [3].

rates irrespectively of the dictionary. In this case, one has to employ a solution that is robust to occlusions [32].

The idea of using disparity learning for efficient distributed coding of stereo images has been previously suggested by Varodayan *et al.* [33]. Their algorithm performs online disparity learning during decoding (it is thus image-dependent) and requires a feedback channel. Our approach is quite different however, because we perform offline learning of general dictionaries that are not image-dependent and we do not require a feedback channel.

B. Camera Pose Estimation

Finally, we discuss the benefits of dictionary learning for the camera pose estimation algorithm proposed in [21]. This algorithm extracts atoms from two views in a distributed fashion, and then finds the atom pairs that are related by local geometric transforms. The selected atom pairs and their transforms are then used to find point correspondences in two views, which are used to estimate the camera pose using the eight-point algorithm [34]. Additionally, we apply Ransac [35] for camera pose estimation in order to be more robust against outliers. We have estimated the camera pose using the initial and learned dictionaries for $\rho = 3$. Two pairs of images from the "Mede" database have been selected, with translation $\mathbf{T} = [100]^\top$ and $\mathbf{T} = [010]^\top$ respectively. For each image in a stereo pair, we have randomly chosen 20 patches of 40×40 pixels, located at the same position in two stereo images. The same image patches are decomposed by MP using initial and learned dictionaries, giving 3 atoms per patch and thus 60 atoms per image. The estimated translation vectors and the angular errors in radians are shown in Tables II and III for both target vectors $\mathbf{T}$. We can see that estimation using the learned dictionary gives significantly better performance than using a randomly initialized dictionary, in terms of the angular error. The learned dictionary leads to a precise estimation of the translation, while the initial dictionary cannot even determine the direction of translation. For completeness, we have also compared our pose estimation method to an approach that uses matching of SIFT features [36] followed by Ransac, where each feature is described by a vector of 128 components and the corresponding spatial position [36]. This is a standard approach to pose estimation [37]. The results of pose estimation using 60 SIFT features per image are reported in the last columns of Tables II and III, with and without Ransac. In both cases, the minimal error with SIFT features is always higher than the error of the learned dictionary without Ransac. Therefore, using the learned dictionary for pose estimation without Ransac gives the smallest error in both examples. Note finally, that Ransac does not necessarily improve the performance of all algorithms, due to the small number of features and thus insufficient statistics. The atom-based approach with a learned dictionary appears to be the best solution when the relative camera pose has to be estimated from a small number of features.

VIII. Conclusions

We have proposed a new method for learning overcomplete dictionaries for representing stereo images. A stereo image model where sparse components are related with local transforms is used as a base for developing a maximum likelihood (ML) method for learning stereo dictionaries. The epipolar geometry constraint has been included in the probabilistic model in order to force the learning algorithm to select atoms that offer good approximation performance and simultaneously satisfy multi-view geometry constraints. The experimental results on omnidirectional images have shown that our method results in dictionaries that give both better stereo matching and approximation properties than randomly selected dictionaries. It leads to important benefits in applications such as distributed scene representation and camera pose estimation.

Appendix

Conditional Coefficient Probabilities: We compute here conditional probabilities $P(b_r|a_l,\boldsymbol{\phi_l},\boldsymbol{\psi}_r)$ and $P(a_l|b_r,\boldsymbol{\phi_l},\boldsymbol{\psi}_r)$.We first replace the expansions for $\mathbf{y}_L$ and $\mathbf{y}_R$ from (1) in (18), and get for all $k=1,\ldots,m$

$$\left\langle \sum_{i=1}^{m} b_{r_i}\psi_{r_i},\boldsymbol{\psi}_{r_k}\right\rangle + \langle \mathbf{e}_R,\boldsymbol{\psi}_{r_k}\rangle = \frac{1}{\sqrt{J_{l_k r_k}}}\left\langle \sum_{i=1}^{m} a_{l_i}\phi_{l_i},\boldsymbol{\phi}_{l_k}\right\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle \mathbf{e}_L,\boldsymbol{\phi}_{l_k}\rangle \quad (34)$$

which can be rewritten as

$$b_{r_k} + \sum_{\substack{i=1\\ i\neq k}}^{m}\langle b_{r_i}\psi_{r_i},\boldsymbol{\psi}_{r_k}\rangle + \langle \mathbf{e}_R,\boldsymbol{\psi}_{r_k}\rangle = \frac{1}{\sqrt{J_{l_k r_k}}}a_{l_k} + \frac{1}{\sqrt{J_{l_k r_k}}}\sum_{\substack{i=1\\ i\neq k}}^{m}\langle a_{l_i}\phi_{l_i},\boldsymbol{\phi}_{l_k}\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle \mathbf{e}_L,\boldsymbol{\phi}_{l_k}\rangle \quad (35)$$

or simply

$$b_{r_k} = \frac{a_{l_k}}{\sqrt{J_{l_k r_k}}} + \eta' \quad (36)$$

where

$$\eta' = \frac{1}{\sqrt{J_{l_k r_k}}}\sum_{\substack{i=1\\ i\neq k}}^{m}\langle a_{l_i}\phi_{l_i},\boldsymbol{\phi}_{l_k}\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle \mathbf{e}_L,\boldsymbol{\phi}_{l_k}\rangle - \sum_{\substack{i=1\\ i\neq k}}^{m}\langle b_{r_i}\psi_{r_i},\boldsymbol{\psi}_{r_k}\rangle - \langle \mathbf{e}_R,\boldsymbol{\psi}_{r_k}\rangle. \quad (37)$$

We will further assume that η' is a small value, since it is a sum of the projection of some noise to a chosen atom, and a linear combination of inner products of a chosen atom with other atoms in the image decomposition. When the image decomposition is sparse and the dictionary is overcomplete, the assumption is usually verified. However, we cannot use directly the expression in (36) to derive the distribution $P(\mathbf{a},\mathbf{b}|\boldsymbol{\Phi},\boldsymbol{\Psi})$ because the sparse support of the stereo images is not known, and hence also the indexes l_k, r_k and $k=1,\ldots,m$. Therefore, we say that an arbitrary stereo atom pair $\boldsymbol{\phi}_l,\boldsymbol{\psi}_r$ and their coefficients a_l, b_r satisfy (36) up to a certain error η_1, which includes also η'. Therefore, we have

$$b_r = \frac{1}{\sqrt{J_{lr}}}a_l + \eta_1 \quad (38)$$

where J_{lr} is the Jacobian of the linear transform of the coordinate system induced by the transform between atoms $\boldsymbol{\phi}_l$ and $\boldsymbol{\psi}_r$. When a_l and b_r are the coefficients of a stereo pair, then they satisfy (38) with a small value of the noise η_1. Otherwise, a_l and b_r are not significant in sparse decompositions of stereo images (according to the model in (1)) and hence the noise η_1 is also small. Therefore, we model η_1 as white Gaussian noise of variance σ_b^2 and get

$$P(\eta_1) = P(b_r|a_l,\boldsymbol{\phi_l},\boldsymbol{\psi}_r) \propto \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right).$$

Although $\boldsymbol{\phi}_l,\boldsymbol{\psi}_r$ are not explicitly contained in the probability expression, they are implicitly there since J_{lr} is evaluated as a Jacobian of a transform between $\boldsymbol{\phi}_l$ and $\boldsymbol{\psi}_r$. Multiplying (36) with $\sqrt{J_{lr}}$, we can get a symmetric relation

$$\begin{aligned} P(\eta_2) &= P(a_l|b_r,\boldsymbol{\phi_l},\boldsymbol{\psi}_r) \\ &\propto \exp\left(-\frac{1}{2\sigma_a^2}(a_l - \sqrt{J_{lr}}b_r)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma_b^2 J_{lr}}(a_l - \sqrt{J_{lr}}b_r)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma_b^2}\left(b_r - \frac{a_l}{\sqrt{J_{lr}}}\right)^2\right) \end{aligned} \quad (39)$$

where we used the fact that variance of the noise $\eta_2 = \sqrt{J_{lr}}\eta_1$ can be evaluated as $\sigma_a = \sigma_b\sqrt{J_{lr}}$. Note that the same expression for $P(a_l|b_r,\boldsymbol{\phi_l},\boldsymbol{\psi}_r)$ can be obtained by considering the inverse transform F_{rl} from atom $\boldsymbol{\psi}_r$ to atom $\boldsymbol{\phi_l}$, because the Jacobian of a linear transform satisfies: $J\left(Q_{lr}^{-1}\right) = 1/J_{lr}$.

Acknowledgment

The authors would like to thank the members of the Redwood Center for Theoretical Neuroscience at UC Berkeley, for the fruitful discussions on ML dictionary learning.

References

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[2] H. Wang, Y. Wexler, E. Ofek, and H. Hoppe, "Factoring repeated content within and among images," *Proc. ACM SIGGRAPH*, pp. 1–10, 2008.
[3] I. Tošić 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.
[4] B. A. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 23, no. 37, pp. 3311–25, 1997.
[5] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.
[6] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[7] B. A. Olshausen, "Learning sparse, overcomplete representations of time-varying natural images," in *Proc. IEEE Int. Conf. Image Process.*, 2003.

[8] B. A. Olshausen, C. Cadieu, B. J. Culpepper, and D. K. Warland, "Bilinear models of natural images," in *Proc. SPIE Conf. Human Vis. Electron. Imag.*, 2007, pp. 1287–1294.
[9] C. Cadieu and B. A. Olshausen, "Learning transformational invariants from time-varying natural images," in *Proc. Conf. Neural Inf. Process. Syst.*, 2008, pp. 209–216.
[10] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
[11] K. Engan, S. Aase, and H. J. Hakon, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 1999, pp. 2443–2446.
[12] K. Engan, B. D. Rao, and K. Kreutz-Delgado, "Frame design using FOCUSS with method of optimal directions (MOD)," in *Proc. Norwegian Signal Process. Symp.*, 1999, pp. 1098–1112.
[13] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A Re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[14] P. Schmid-Saugeon and A. Zakhor, "Dictionary design for matching pursuit and application to motion-compensated video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 880–886, Jun. 2004.
[15] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, pp. 3320–3325, 2003.
[16] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "MoTIF: An efficient algorithm for learning translation invariant dictionaries," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2006.
[17] K. Engan, K. Skretting, and J. Husřy, "Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation," *Digital Signal Process.*, vol. 17, no. 1, pp. 32–49, 2007.
[18] G. Monaci, F. Sommer, and P. Vandergheynst, "Learning bimodal structure in audio-visual data," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, Dec. 2009.
[19] P. Hoyer and A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Netw.: Comput. Neural Syst.*, vol. 11, no. 3, pp. 191–210, 2000.
[20] K. Okajima, "Binocular disparity encoding cells generated through an Infomax based learning algorithm," *Neural Netw.*, vol. 17, no. 7, pp. 953–962, 2004.
[21] I. Tošić and P. Frossard, "Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images," in *Proc. Eur. Signal Process. Conf.*, 2007.
[22] I. Tošić and P. Frossard, "Conditions for recovery of sparse signals correlated by local transforms," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 684–688.
[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[24] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. New York: Kluwer, 1998, pp. 355–368.
[25] B. J. Culpepper, "Learning 'What' and 'Where' From Movies," Master's, UC, Berkeley, 2007.
[26] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[27] I. V. 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.
[28] P. Frossard, "Robust and multiresolution video delivery: From H.26x to matching pursuit based technologies," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2000.
[29] V. N. Temlyakov, "Weak greedy algorithms," *Adv. Computat. Math.*, vol. 12, no. 2–3, pp. 213–227, 2000.
[30] I. Tošić, "On unifying sparsity and geometry for image-based scene represenation," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2009.
[31] I. Tošić, P. Frossard, and P. Vandergheynst, "Progressive coding of 3-D objects based on overcomplete decompositions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1338–1349, Nov. 2006.
[32] I. Tošić and P. Frossard, "Geometry-based distributed coding of multi-view omnidirectional images," in *Proc. IEEE ICIP*, 2008, pp. 2220–2223.
[33] D. Varodayan, Y.-C. Liu, M. Flierl, and B. Girod, "Wyner-Ziv coding of multiview images with unsupervised learning of disparity and gray code," in *Proc. Picture Coding Symp.*, 2007, pp. 1112–1115.
[34] Y. Ma, S. Soatto, J. Košecká, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. New York: Springer, 2004.
[35] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[36] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[37] J. Kosecká and X. Yang, "Global localization and relative pose estimation based on scale-invariant features," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 319–322.

Ivana Tošić (S'04–M'09) received the Dipl.Ing. degree in telecommunications from the University of Niš, Serbia, and the Ph.D. degree in computer and communication sciences from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2003 and 2009, respectively.

She is currently a Postdoctoral Researcher at the Redwood Center for Theoretical Neuroscience, University of California at Berkeley, where she works on the intersection of image processing and computational neuroscience domains. In 2009, she was awarded the Swiss National Science Foundation fellowship for prospective researchers. Her research interests include representation and coding of the plenoptic function, distributed source coding, binocular vision and 3-D object representation and compression.

Pascal Frossard (S'96–M'01–SM'04) received the M.S. and Ph.D. degrees, both 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 at 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 professor 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 has been the General Chair of IEEE ICME 2002 and Packet Video 2007. He has been 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 (2004–2010), the IEEE TRANSACTIONS ON IMAGE PROCESSING (2010-) and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2006-). He is an elected member of the IEEE Image and Multidimensional Signal Processing Technical Committee (2007-), the IEEE Visual Signal Processing and Communications Technical Committee (2006-), and the IEEE Multimedia Systems and Applications Technical Committee (2005-). He has served as 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 received the Swiss NSF Professorship Award in 2003, the IBM Faculty Award in 2005 and the IBM Exploratory Stream Analytics Innovation Award in 2008.