View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Infoscience - École polytechnique fédérale de Lausanne

SCHOOL OF ENGINEERING - STI
SIGNAL PROCESSING LABORATORY LTS4
Ivana Tosic and Pascal Frossard

CH-1015 LAUSANNE

Telephone: +4121 6934712
Telefax: +4121 6937600
e-mail: ivana.tosic@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

DICTIONARY LEARNING IN STEREO IMAGING

Ivana Tosic and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)

Signal Processing Laboratory LTS4 Technical Report

TR-LTS-2009-005

June 11th, 2009

Part of this work has been submitted to the IEEE Transactions on Image Processing.

This work has been supported by the Swiss National Science Foundation under grant 200020-120063, and by the EU under the FP7 project APIDIS (ICT-216023).

Dictionary learning in stereo imaging

Ivana Tošić and Pascal Frossard
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4), Lausanne, 1015 - Switzerland.
{ivana.tosic, pascal.frossard}@epfl.ch
Fax: +41 21 693 7600, Phone: +41 21 693 4712

Abstract

This paper presents a new method for learning overcomplete dictionaries adapted to efficient 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 atoms in two stereo views. A maximum-likelihood method for learning stereo dictionaries is then proposed, which includes a multi-view geometry constraint in the probabilistic modeling in order to obtain dictionaries optimized for the joint representation of stereo images. The dictionaries are learned by optimizing the maximum-likelihood objective function using the expectation-maximization algorithm. We illustrate the learning algorithm in the case of omnidirectional images, where we learn scales of atoms in a parametric dictionary. The resulting dictionaries provide both better performance in the joint representation of stereo omnidirectional images and 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

Sparse approximations, dictionary learning, multi-view imaging, omnidirectional cameras.

I. Introduction

Multiple images of a 3D scene taken from different viewpoints contain information about both 3D structure and texture of the objects in the scene. Therefore, these images give a richer description of the environment compared to a single view. Multi-view images are usually captured by a network of cameras distributed in a 3D scene. Such visual sensor networks can find usage in applications like 3D television, surveillance, robotics or exploration. However, dealing with the high dimensional visual information still poses many challenges, such as multi-view compression, 3D geometry estimation and scene analysis.

Extraction of 3D information from multiple views relies on the theory of the multiple view geometry [1], which relates image features that represent the same 3D objects in different views. Pixel-based image representation is used in most of the image-based 3D geometry estimation methods that build dense depth maps by computing pixel correspondences. However, pixel-based representations are highly inefficient for image coding and compression. On the other hand, image representations with orthogonal bases are efficient for compression, but generally fail to efficiently capture the geometry of objects in a scene and the correlation between views. Therefore, multi-view imaging requires new image representation methods that give good performance in both compression and scene geometry estimation.

This paper addresses the problem of learning dictionaries adapted to the representation of multi-view images. We consider sparse image approximations with overcomplete dictionaries of geometrical atoms. As the correlation between multi-view images arises from the geometric constraints on the objects in the scene, it can be simply described by local transforms of geometric atoms [2]. 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. Dictionary learning for sparse signal representations has become an extremely active area of research in the last few years, when it was realized that adapting the dictionary to a specific task or imposing a certain structure to the dictionary can yield significant improvements of performance in target applications. Researchers have addressed the problem of learning dictionaries for image [3]–[5] and video representation [6]–[8]. To the best of our knowledge there has been however no work on learning dictionaries for multi-view representation. 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 the sparsity prior, and at the same time give better multi-view geometry estimation from sparse low-level visual features. Our method builds upon the ML method for learning overcomplete dictionaries from natural monocular images, introduced by Olshausen and Field [3]. Additionally, the proposed probabilistic approach to learning includes the epipolar geometry in the modeling, and hence matches corresponding atoms within the learning process itself. The optimization problem is cast as an energy minimization problem, that we finally solve with an Expectation-Maximization (EM) algorithm. The experimental results show the significant benefits of stereo dictionary learning for applications such as distributed scene representation and camera pose recovery.

The organization of this paper is as follows. 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 energy minimization solution in given in Section V-B. Experimental results in omnidirectional imaging are presented in Section VI.

We use the following notation convention throughout the paper. Small bold face letters denote vectors, while capital bold face letters denote matrices. Capital L, R letters in the subscript and (L), (R) in the superscript denote the parameters that refer to the left, respectively right, image in a stereo image pair. Small letters within square brackets in the superscript (e.g., $\mathbf{h}^{[k]}$) denote the counter parameter, for example the counter of iterations or the counter of pixels. We denote the vector l_p norm as $\|\cdot\|_p$.

II. Related work

The earliest work addressing the problem of learning overcomplete dictionaries for image representation has appeared in 1997, in the visual neuroscience research domain. It was the work of Olshausen and Field [3], [9], who developed a maximum likelihood (ML) dictionary learning method from natural images under the sparse coding assumption. The goal of the work was to give evidence that the coding in the primary visual area V1 in the human cortex probably follows the sparse image model. Their learning method yielded 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 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 $\mathbf{y} = \mathbf{\Phi}\mathbf{a}$, where $\mathbf{a}$ is considered as a hidden variable. 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. This method has also been extended to time-varying visual stimuli [6]–[8].

The probabilistic inference approach to overcomplete dictionary learning has been later adopted by other researchers. Engan et al. [10], [11] 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 [3] the sparse coding step involves finding the equilibrium solution of the differential equation over $\mathbf{a}$, MOD uses either the OMP [10] or the FOCUSS [11] algorithm to find sparse vector $\mathbf{a}$. Second, the dictionary $\mathbf{\Phi}$ is updated as the solution of the differential equation $\partial E/\partial \mathbf{\Phi} = 0$, where E is the energy function that is in this case equal to the residue $\|\mathbf{y} - \mathbf{\Phi}\mathbf{a}\|_F^2$ and $\|\cdot\|_F$ denotes the Frobenius norm. These two modifications make the MOD approach faster compared to ML method of Olshausen and Field. Maximum a posteriori (MAP) dictionary learning method, proposed by Kreutz-Delgado et al. [4] 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 (sparse coding-dictionary update) algorithm, where dictionary update includes an additional constraint on the dictionary that can be for example unit Frobenius norm of $\mathbf{\Phi}$ or unit l_2 norm of all atoms in the dictionary. The sparse coding step is performed with FOCUSS [12].

A slightly different family of dictionary learning techniques is based on vector quantization 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 [13], [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 [5]. 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 is hence generalized K-means since each patch can be represented by multiple atoms and 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. However, ML and MAP methods are characterized by a flexibility in extending the probabilistic modeling to higher-dimensional data, like videos [7], [8] 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], [19]. Because of this property, we have chosen the ML approach for learning parametric dictionaries in stereo imaging.

Even though there has been recently a great amount of research done in the domain of dictionary learning for single images, there has been no work targeting the problem of learning overcomplete dictionaries for stereo imaging. Learning the binocular cells receptive fields and disparity tuning curves has been, however, widely investigated without the assumption of the sparse coding in overcomplete dictionaries. Hoyer and Hyvärinen [20] 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 resulted in Gabor-like basis functions that are tuned to

different disparities. Okajima has proposed an Infomax learning approach, where the binocular receptive fields are learned by maximizing the mutual information between the stereo image model and the disparity [21]. They have obtained results similar to Hoyer and Hyvärinen. The stereo dictionary learning method that we propose in this work learns stereo atoms from stereo image pairs while simultaneously performing the disparity estimation of the learned image features. The disparity estimation is included in the probabilistic model of stereo images, thus removing the need for disparity estimation as preprocessing step. However, the main target of this work is not to study the receptive fields of binocular cells or their tuning characteristics. The learning strategy that we propose here aims at designing stereo dictionaries that have the optimal properties for both image approximation and disparity or 3D scene structure estimation. To the best of our knowledge, such problem has never been studied in the past.

III. Multi-view imaging

A. Stereo image model

Developing the maximum likelihood dictionary learning method for stereo images requires first a definition of the stereo image model. We consider two images vectorized into column vectors: left image $\mathbf{y}_L$ and right image $\mathbf{y}_R$. 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. $\mathbf{e}_R$. We have:

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

where $\mathcal{L} = \{l_k\}, \mathcal{R} = \{r_k\}, k = 1, ..., m$ label the sets of atoms that participate in the sparse decompositions of $\mathbf{y}_L$ and $\mathbf{y}_R$, respectively. In other words, $\{l_k\}, \{r_k\}, k = 1, ..., m$ denote the atoms for which $a_{l_k} \neq 0$ and $b_{r_k} \neq 0$. This model assumes that both stereo images are m-sparse, i.e., composed of m atoms, but the atoms in the left and the right image do not have to be the same ($\mathcal{L} \neq \mathcal{R}$). Besides the different sparse supports $\mathcal{L}$ and $\mathcal{R}$, sparse image decompositions in Eq. (1) have different vectors of coefficients: $\mathbf{a}$ for the left, and $\mathbf{b}$ for the right image. The motivation behind this model is that left and right images record the visual information from the same 3D environment and typically contain the image projections of the same 3D scene features, thus the number of sparse components will be approximately the same. Moreover, 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:

$$\mathbf{y}_R = \sum_{k=1}^{m} b_{r_k}\psi_{r_k} + \mathbf{e}_R = \sum_{k=1}^{m} b_{r_k} F_{l_k r_k}(\phi_{l_k}) + \mathbf{e}_R, \tag{2}$$

where $F_{l_k r_k}(\cdot)$ denotes a transform of an atom ϕ_{l_k} in $\mathbf{y}_L$ to an atom ψ_{r_k} in $\mathbf{y}_R$, and it differs for each $k = 1, ..., m$. This correlation model is a special case of the model introduced in [2] 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 Eq. (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 additionally anisotropic scaling of the image features [2]. Such transforms are efficiently represented with a parametric dictionary whose construction is built on these transformations. Given a generating function g defined in the Hilbert space, 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 transformations applied to the generating function g. This is equivalent to applying a unitary operator $U(\gamma)$ to the generating function g, i.e.: $g_\gamma = U(\gamma)g$. The unitary operator transforms the generating function by applying a linear transform of the coordinate system in the space $\mathcal{H}$ where the images and the dictionaries are defined. Let $\mathbf{v}$ denote the coordinates in $\mathcal{H}$ and $\mathbf{u}$ denote the coordinates obtained by transforming $\mathbf{v}$ with an arbitrary linear transform Q, i.e., $\mathbf{u} = Q(\mathbf{v})$. The function g is hence transformed into a function g_γ as:

$$g(\mathbf{u}) = g(Q(\mathbf{v})) = F(g(\mathbf{v})) = g_\gamma(\mathbf{v}). \tag{3}$$

Further on, the function g_γ needs to be normalized and have the l_2 norm equal to one. Therefore, we define atoms in the structured dictionary as: $\phi = g_\gamma / \|g_\gamma\|_2$. The multi-dimensional space of parameters γ is continuous and infinite, thus building a dictionary with all possible transforms will yield 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 built on the same

generating function g, but using 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:

$$\begin{aligned} \phi_l &\equiv g_{\gamma_l^{(L)}}, \qquad \gamma_l^{(L)} \in \Gamma_L, \quad \text{for } l = 1, ..., M \\ \psi_r &\equiv g_{\gamma_r^{(R)}}, \qquad \gamma_r^{(R)} \in \Gamma_R, \quad \text{for } r = 1, ..., M, \end{aligned} \tag{4}$$

where we assume that $g_{\gamma_l^{(L)}}$ and $g_{\gamma_r^{(R)}}$ are already normalized, so we drop the norm in the denominator. An important property of the structured dictionary is that a transformation of an atom ϕ_l in the left image into an atom ϕ_r in the right image reduces to a transform of its transform parameters, i.e.,

$$\psi_r = F_{lr}(\phi_l) = U(\gamma')\phi_l = U(\gamma' \circ \gamma_l^{(L)})g. \tag{5}$$

In the following, the transform $F(\cdot)$ that changes atom ϕ_l into an atom ψ_r is denoted as $F_{lr}(\cdot)$. The corresponding linear transform of the coordinate system as $Q_{lr}(\cdot)$.

The transforms $F_{l_k r_k}$, $k = 1, ..., m$ relating the atoms in the left and right view in Eq. (2) are not arbitrary. As the corresponding atoms are images of the same feature in the 3D space, the atom transformations have to satisfy multi-view epipolar geometry constraints.

B. Multi-view geometry

The epipolar geometry constraint imposes a geometric relation between 3D 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 3D point p from two camera positions with relative pose $(\mathbf{R}, \mathbf{T})$. $\mathbf{R} \in SO(3)$ is the relative orientation between cameras and $\mathbf{T} \in \mathbb{R}^3$ is their relative position. The epipolar geometry constraint is then:

$$\mathbf{u}^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{v} = 0. \tag{6}$$

The matrix $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{v}$ as matrix multiplication. In the case where the point $\mathbf{v}$ lies on the atom ϕ_l, as shown on the Fig. 1, our goal is to seek for a transform F_{lr} (and equivalently for Q_{lr}) such that $\mathbf{u} = Q_{lr}(\mathbf{v})$. In other words, the point $\mathbf{u}$ should lie on the atom $\psi_r = F_{lr}(\phi_l)$. The epipolar geometry constraint is then:

$$[Q_{lr}(\mathbf{v})]^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{v} = 0. \tag{7}$$

As formalized in Eq. (3), the transform $Q(\cdot)$ depends on the index parameters γ. In the case above, Q_{lr} denotes the linear transform of coordinates between atoms $\phi_l = g_{\gamma_l^{(L)}}$ and $\psi_r = g_{\gamma_r^{(R)}}$, thus it depends on parameters $\gamma_l^{(L)}$ and $\gamma_r^{(R)}$. Knowing these parameters, it is very simple to derive the analytic form for Q_{lr}. For omnidirectional images Q_{lr} is derived in [2] and also given for completeness in Section VI-A.

The epipolar constraint in Eq. (7) is rarely satisfied exactly, due to discrete spatial sampling of images, and can be only evaluated with a certain error ε_l. The estimated epipolar constraint d_{el} is thus given as:

$$d_{el} = [Q_{lr}(\mathbf{v})]^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{v} + \varepsilon_L = d_L + \varepsilon_L, \tag{8}$$

where $d_L = [Q_{lr}(\mathbf{v})]^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{v}$. 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 in a point $\mathbf{v}$ in the first image, we have the epipolar geometry estimate d_{er} given by

$$d_{er} = [Q_{lr}^{-1}(\mathbf{u})]^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{u} + \varepsilon_R = d_R + \varepsilon_R. \tag{9}$$

where $d_R = [Q_{lr}^{-1}(\mathbf{u})]^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{u}$. The most likely transforms Q_{lr} (and equivalently F_{lr}) in pairs of stereo images are the transforms that give small epipolar errors. We will thus use this fact in the probabilistic framework for the maximum-likelihood learning of stereo dictionaries, presented in the following section.

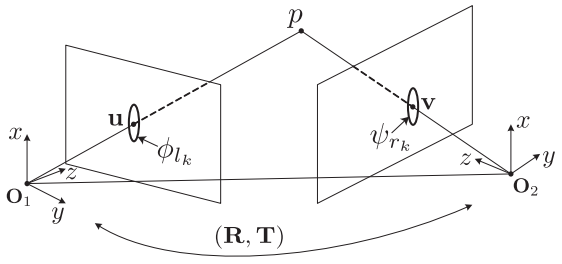


Fig. 1. Epipolar geometry between stereo atoms.

IV. Maximum-likelihood learning of dictionaries for stereo images

A. Problem formulation

Following a similar approach as in [3], we formulate the probabilistic framework for the maximum likelihood learning of overcomplete dictionaries $\mathbf{\Phi}, \mathbf{\Psi}$ that are used to represent stereo images $\mathbf{y}_L$ and $\mathbf{y}_R$ respectively. We want to 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 geometric transforms. In other words, we want to learn the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ simultaneously. Therefore, we need to maximize the probability that the observed stereo images $\mathbf{y}_L$ and $\mathbf{y}_R$ arise from dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ under a sparsity prior, and that the epipolar constraint between all corresponding points on $\mathbf{y}_L$ and $\mathbf{y}_R$ is equal to zero, i.e., $D = 0$. The epipolar geometry constraint is introduced in the probabilistic model in order to maximize the probability that the selected stereo pairs of atoms is conformant with the multi-view geometry. Similarly to the problem proposed in [3], the dictionary learning is performed by minimizing the Kullback-Leibler (KL) divergence between the probability distribution of natural images arising from the image model and the actual distribution of natural images $P^*(\mathbf{y}_L, \mathbf{y}_R)$. This KL divergence is given as:

$$KL = \int\int P^*(\mathbf{y}_L, \mathbf{y}_R) log \frac{P^*(\mathbf{y}_L, \mathbf{y}_R)}{P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi})} \mathrm{d}\mathbf{y}_L \mathrm{d}\mathbf{y}_R. \tag{10}$$

Since $P^*(\mathbf{y}_L, \mathbf{y}_R)$ is constant, minimizing the KL divergence is equivalent to maximizing the log-likelihood

$$\log P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi})$$

that a set of stereo natural images $(\mathbf{y}_L, \mathbf{y}_R)$ arises from the overcomplete sets of functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$. Therefore, the goal of learning is to find the overcomplete dictionaries $\mathbf{\Phi}^*$ and $\mathbf{\Psi}^*$ that are the solutions of the following optimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg \max_{\mathbf{\Phi}, \mathbf{\Psi}} \langle \max_{\mathbf{a}, \mathbf{b}} \log P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi}) \rangle, \tag{11}$$

where $\mathbf{y}_L = \mathbf{\Phi}\mathbf{a}$ and $\mathbf{y}_R = \mathbf{\Psi}\mathbf{b}$. If we marginalize the cost function over $\mathbf{a}$ and $\mathbf{b}$ we obtain

$$P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi}) = \int_{\mathbf{a}, \mathbf{b}} P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) \mathrm{d}\mathbf{a} \mathrm{d}\mathbf{b}. \tag{12}$$

In the rest of this section, we compute the objective function.

B. Coefficient vector distributions

In order to compute the cost function of Eq. (12), we first need to define the joint distribution $P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$ of the coefficients $\mathbf{a}$ and $\mathbf{b}$, given the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$. First, we note that we can decompose the joint distribution of coefficients in two ways:

$$P(a_l, b_r | \phi_l, \psi_r) = P(b_r | a_l, \phi_l, \psi_r) P(a_l) \tag{13}$$

$$P(a_l, b_r | \phi_l, \psi_r) = P(a_l | b_r, \phi_l, \psi_r) P(b_r), \tag{14}$$

where we assume that priors on coefficients in each image $P(a_l)$ and $P(b_r)$ are independent of the atoms. We can therefore write

$$P(a_l, b_r | \phi_l, \psi_r) = \sqrt{P(b_r | a_l, \phi_l, \psi_r) P(a_l | b_r, \phi_l, \psi_r) P(a_l) P(b_r)}. \tag{15}$$

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 $P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$ becomes factorial, and we can write

$$P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) = \prod_{l=1}^{M} \prod_{r=1}^{M} P(a_l, b_r | \phi_l, \psi_r) = \prod_{l=1}^{M} \prod_{r=1}^{M} \sqrt{P(b_r | a_l, \phi_l, \psi_r) P(a_l | b_r, \phi_l, \psi_r) P(a_l) P(b_r)}. \tag{16}$$

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

$$\forall k \text{ and } \forall \mathbf{v} \text{ s.t. } \phi_{l_k}(\mathbf{v}) \neq 0, \ \Rightarrow \ \mathbf{y}_L(\mathbf{v}) = \mathbf{y}_R(Q_{l_k r_k}(\mathbf{v})) = \mathbf{y}_R(\mathbf{u}), \tag{17}$$

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 Eq. (17), we can use the Lemma 1 in [22], which states the following equality:

$$\langle \mathbf{y}_R, \psi_{r_k} \rangle = \frac{1}{\sqrt{J_{l_k r_k}}} \langle \mathbf{y}_L, \phi_{l_k} \rangle, \tag{18}$$

where $J_{l_k r_k} = |\frac{\partial \mathbf{u}}{\partial \mathbf{v}}| = |\frac{\partial Q_{l_k r_k}(\mathbf{v})}{\partial \mathbf{v}}|$ is the Jacobian determinant (or simply the Jacobian) of the linear transform $Q_{l_k r_k}$. Recall that the inner products in Eq. (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 Eq. (18) we obtain the following probabilities:

$$P(b_r|a_l, \phi_l, \psi_r) = P(a_l|b_r, \phi_l, \psi_r) = \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right), \tag{19}$$

where z_b is the normalization factor and 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 Eq. (19) is given in Appendix A.

We model now the distribution of the coefficients. The distribution of the coefficients actually depends on an arbitrarily chosen dictionary. However, imposing the independence of the coefficients with respect to the dictionary during learning actually leads to inferring a dictionary that gives the same prior distribution of coefficients for all types of images. Furthermore, when the prior of coefficients is tightly picked at zero, the learning leads to a universal dictionary in which all natural images have sparse decompositions. We chose here a different approach than the one proposed in [3], where the coefficient distribution is taken to be continuous and 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. These distributions are:

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

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

Choosing $p \ll q$ introduces a sparsity assumption on the coefficients, i.e., it is much more probable that the coefficient takes the value zero than a value greater than zero. If the images can be represented by m components from a dictionary of size M, we get:

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

$$P(\mathbf{b}) = \prod_{r=1}^{M} P(b_r) = p^m (1-p)^{(M-m)}. \tag{21}$$

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 $l, r = 1, .., M$ are independent and identically distributed, the Eqs (20) and (21) can be rewritten as:

$$P(\mathbf{a}) = (\frac{e^{1/\lambda}}{1+e^{1/\lambda}})^M \exp(-m/\lambda) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{a}\|_0}{\lambda}\right), \tag{22}$$

$$P(\mathbf{b}) = (\frac{e^{1/\lambda}}{1+e^{1/\lambda}})^M \exp(-m/\lambda) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{b}\|_0}{\lambda}\right). \tag{23}$$

From Eqs.(19), (20) and (21), we can write:

$$P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{z_b z_\lambda} \exp\left(-\frac{1}{2\sigma_b^2} \sum_{l=1}^{M} \sum_{r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right) \exp\left(-\frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0)\right). \tag{24}$$

We can now go back to the likelihood function in Eq. (12). Since $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ is the product of a zero-mean Gaussian distribution and a discrete distribution tightly peaked at zero, we can approximate the integral in the right hand side of Eq. (12) by its value at the maximum of its argument. A similar approximation is proposed in [9] in the learning of dictionaries for natural images. Thus, Eq. (12) becomes:

$$P(\mathbf{y}_L, \mathbf{y}_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) \approx P(\mathbf{y}_L, \mathbf{y}_R, D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}). \tag{25}$$

Finally, we have

$$\begin{aligned} P(\mathbf{y}_L, \mathbf{y}_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) &\approx P(\mathbf{y}_L, \mathbf{y}_R|D = 0, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) & (26) \\ &= P(\mathbf{y}_L, \mathbf{y}_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}), & (27) \end{aligned}$$

where Eq. (26) follows from the chain rule, and Eq. (27) holds since $D = 0$ does not bring more information to $\mathbf{y}_L, \mathbf{y}_R$ than $\mathbf{\Phi}, \mathbf{\Psi}$. In order to evaluate the likelihood function, we now have to compute $P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$, which is the probability that the epipolar constraint D is equal to zero given the stereo image model in Eq. (2).

C. Epipolar matching probability

We compute now the probability that the atoms during learning correspond to the same 3D objects under 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 3D space, has been derived in the Section III-B. The ε_L and ε_R on the estimation of the epipolar constraints d_{el} and d_{er} are assumed to be i.i.d. white zero-mean Gaussian noises of variances σ_{dl}^2 and σ_{dr}^2, respectively. We have therefore:

$$P(\varepsilon_L) = \frac{1}{z_{dl}} \exp\left(-\frac{\varepsilon_L^2}{2\sigma_{dl}^2}\right), \tag{28}$$

$$P(\varepsilon_R) = \frac{1}{z_{dr}} \exp\left(-\frac{\varepsilon_R^2}{2\sigma_{dr}^2}\right), \tag{29}$$

where z_{dl} and z_{dr} are the normalization factors. Equivalently, we can define the conditional probability of the random variables d_{el} and d_{er}, given a pair of points $\mathbf{v}$, $\mathbf{u}$ and atoms ϕ_l, ψ_r as:

$$P(d_{el}|\mathbf{v},\mathbf{u},\phi_l,\psi_r) = P(d_{el}|d_L) = \frac{1}{z_{dl}} \exp\left(-\frac{(d_{el}-d_L)^2}{2\sigma_{dl}^2}\right), \tag{30}$$

$$P(d_{er}|\mathbf{v},\mathbf{u},\phi_l,\psi_r) = P(d_{er}|d_R) = \frac{1}{z_{dr}} \exp\left(-\frac{(d_{er}-d_R)^2}{2\sigma_{dr}^2}\right). \tag{31}$$

The atoms ϕ_l and ψ_r influence the functions d_L and d_R that are based on the transform Q_{lr}, which depends on the atoms. Since we want to find the probability that two atoms satisfy the epipolar constraint, we are interested in the probability of the particular realization of the random variables d_{el} and d_{er} when they are simultaneously equal to zero. Therefore, we define the conditional probability of $d_{el}=0, d_{er}=0$, given $\mathbf{v},\mathbf{u},\phi_l,\psi_r$ as:

$$\begin{aligned} P(d_{el}=0,d_{er}=0|\mathbf{v},\mathbf{u},\phi_l,\psi_r) &= P(d_{el}=0|\mathbf{v},\mathbf{u},\phi_l,\psi_r)P(d_{er}=0|\mathbf{v},\mathbf{u},\phi_l,\psi_r) \\ &= \frac{1}{z_{dl}z_{dr}} \exp\left(-\frac{d_L^2}{2\sigma_{dl}^2}\right) \exp\left(-\frac{d_R^2}{2\sigma_{dr}^2}\right). \end{aligned} \tag{32}$$

We extend this conditional probability to all the pairs of pixels that undergo the transformation defined by the atom pair, i.e., for all $\mathbf{v}_i, \mathbf{u}_i$, $i=1,...,q$. Let D_{lr} denote the event that the epipolar constraints are simultaneously respected for all pairs of pixels. In this case, $d_{el}^{[i]}=0, d_{er}^{[i]}=0$ for all $i=1,...,q$, where $d_{el}^{[i]} = d_L^{[i]} + \varepsilon_l = [Q_{lr}(\mathbf{v}_i)]^{\mathsf{T}}\hat{\mathbf{T}}\mathbf{R}\mathbf{v}_i + \varepsilon_l$ and $d_{er}^{[i]} = d_R^{[i]} + \varepsilon_l = [Q_{lr}^{-1}(\mathbf{u}_i)]^{\mathsf{T}}\hat{\mathbf{T}}\mathbf{R}\mathbf{u}_i + \varepsilon_r$. If we assume that the estimation of the epipolar distance for each pixel pair i can be computed independently, we have

$$\begin{aligned} P(D_{lr}=0|\phi_l,\psi_r) &= \prod_{i=1}^{q} P(d_{el}^{[i]}=0, d_{er}^{[i]}=0|\mathbf{v}_i,\mathbf{u}_i,\phi_l,\psi_r) \\ &= \prod_{i=1}^{q} \frac{1}{z_{dl}^{[i]}z_{dr}^{[i]}} \exp\left(-\frac{(d_L^{[i]})^2}{(2\sigma_{dl}^{[i]})^2}\right) \exp\left(-\frac{(d_R^{[i]})^2}{2(\sigma_{dr}^{[i]})^2}\right) \\ &= \frac{1}{z_D} \exp\left(-\frac{1}{2\sigma_D^2}\sum_{i=1}^{q}\left(w_l^{[i]}(d_L^{[i]})^2 + w_r^{[i]}(d_R^{[i]})^2\right)\right) = \frac{1}{z_D}\exp\left(-\frac{1}{2\sigma_D^2}W_{lr}\right), \end{aligned} \tag{33}$$

where $z_D = \prod_{i=1}^{q} z_{dl}^{[i]} z_{dr}^{[i]}$, $w_l^{[i]} = \sigma_D^2/(\sigma_{dl}^{[i]})^2$ and $w_r^{[i]} = \sigma_D^2/(\sigma_{dr}^{[i]})^2$. 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 for points that are closer to the geometric discontinuity represented by the atom, where the estimation of the epipolar constraint is more reliable. The function W_{lr} has been introduced in Eq. (33) 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 $P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})$ as:

$$P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) = \frac{1}{z_D}\exp\left(-\frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)W_{lr}\right), \tag{34}$$

where $\mathcal{I}$ is the indicator function that models the distribution of the coeffficients.

We finally compute the last component of the objective function in Eq. (26), which is the probability $P(\mathbf{y}_L,\mathbf{y}_R|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})$ that refers to the approximation error. We use a similar assumption as [3] where this probability is modeled by a Gaussian white noise. We therefore have:

$$P(\mathbf{y}_L,\mathbf{y}_R|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) = P(\mathbf{e}_L+\mathbf{e}_R) = \frac{1}{z_I}\exp\left(-\frac{1}{2\sigma_I^2}(\|\mathbf{y}_L-\mathbf{\Phi}\mathbf{a}\|_2^2 + \|\mathbf{y}_R-\mathbf{\Psi}\mathbf{b}\|_2^2)\right), \tag{35}$$

where z_I is a normalization factor, and σ_I is the variance of the Gaussian noise. Note that we have in Eq. (35) the fact that the sum of two zero-mean Gaussian random variables is also a zero-mean Gaussian random variable.

We can now rewrite the optimization problem of Eq. (11) as:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\max_{\mathbf{\Phi},\mathbf{\Psi}} \left[\max_{\mathbf{a},\mathbf{b}} \left(\log P(\mathbf{y}_L, \mathbf{y}_R | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(D=0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) \right) \right] \quad (36)$$

whose components are given respectively by Eqs. (35), (34) and (24).

V. EM-BASED ENERGY MINIMIZATION ALGORITHM

A. Energy minimization problem

The optimization problem of Eq. (36) can be cast as an energy minimization problem. The ML optimization problem is equivalent to solving the following energy minimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}} \langle \min_{\mathbf{a},\mathbf{b}} E(\mathbf{y}_L, \mathbf{y}_R, D=0, \mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) \rangle, \quad (37)$$

where E denotes the energy function given as:

$$\begin{aligned} E(\mathbf{y}_L, \mathbf{y}_R, D=0, \mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) &= \frac{1}{2\sigma_I^2} (\|\mathbf{y}_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|\mathbf{y}_R - \mathbf{\Psi}\mathbf{b}\|_2^2) \\ &+ \frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l) \mathcal{I}(b_r) W_{lr} \\ &+ \frac{1}{2\sigma_b^2} \sum_{l=1}^{M} \sum_{r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda} (\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \end{aligned} \quad (38)$$

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:

$$E(\mathbf{y}_L, \mathbf{y}_R, D=0, \mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) = f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda} (\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \quad (39)$$

Our target optimization problem in Eq. (37) thus becomes the following energy minimization:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}} \left[\min_{\mathbf{a},\mathbf{b}} \left(f(\mathbf{y}_L, \mathbf{y}_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda} (\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0) \right) \right]. \quad (40)$$

B. EM-based algorithm

We have seen above that the ML learning problem can be written as an energy minimization problem. Inspired by the method proposed in [3], we can solve the energy minimization problem 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 coefficient vector $(\mathbf{a}, \mathbf{b})$. 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 $E(\mathbf{y}_L, \mathbf{y}_R, D=0, \mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$. Therefore, the algorithm iterates between the sparse coding and the dictionary learning steps until convergence. Such an iterative solution for the ML learning problem is actually 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 or latent variables, and $(\mathbf{\Phi}, \mathbf{\Psi})$ are parameters [24], [25]. The EM algorithm iterates between two steps: Expectation (E) and Maximization (M). In the E step the energy is minimized with respect to $(\mathbf{a}, \mathbf{b})$, and it is essentially the sparse coding step. The M step performs minimization with respect to the dictionaries, and it corresponds to the learning step.

1) Minimization with respect to the coefficients: In the sparse coding step, we need to find the coefficients $\mathbf{a}$ and $\mathbf{b}$ such that

$$(\mathbf{a},\mathbf{b})^* = \arg\min_{\mathbf{a},\mathbf{b}} f(\mathbf{y}_L,\mathbf{y}_R,\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{41}$$

We can see that this problem is similar to the constrained l_0 sparse approximation problem when casted as an unconstrained problem. The multiplier $\frac{1}{2\lambda}$ 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 will 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 could lead to many small but non-zero coefficients, which would increase the complexity of computing the epipolar constraint part of the energy function in Eq. (38).

We propose a greedy algorithm that chooses at each iteration k the pair of atoms ϕ_{l_k},ψ_{r_k} that give the minimal value of the function:

$$\begin{aligned}(\phi_{l_k},\psi_{r_k}) &= \arg\min_{\phi_l,\psi_r}[\frac{1}{2\sigma_I^2}(\|\mathbf{h}_l^{[k-1]} - \langle\mathbf{h}_l^{[k-1]},\phi_l\rangle\phi_l\|_2^2 + \|\mathbf{h}_r^{[k-1]} - \langle\mathbf{h}_r^{[k-1]},\psi_r\rangle\psi_r\|_2^2)\\ &+ \frac{1}{2\sigma_D^2}W_{lr} + \frac{1}{2\sigma_b^2}(\langle\mathbf{h}_r^{[k-1]},\psi_r\rangle - \frac{\langle\mathbf{h}_l^{[k-1]},\phi_l\rangle}{J_{lr}})^2],\end{aligned} \tag{42}$$

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. 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:

$$\begin{aligned}\mathbf{h}_L^{[k]} &= \mathbf{h}_L^{[k-1]} - \langle\mathbf{h}_L^{[k-1]},\phi_{l_k}\rangle\phi_{l_k},\\ \mathbf{h}_R^{[k]} &= \mathbf{h}_R^{[k-1]} - \langle\mathbf{h}_R^{[k-1]},\psi_{r_k}\rangle\psi_{r_k}.\end{aligned} \tag{43}$$

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

$$\begin{aligned}a_{l_k} &= \langle\mathbf{h}_L^{[k-1]},\phi_{l_k}\rangle,\\ b_{r_k} &= \langle\mathbf{h}_R^{[k-1]},\psi_{r_k}\rangle.\end{aligned} \tag{44}$$

We will refer to this algorithm as Multi-view Matching Pursuit (MVMP)[1], which can be shown to be essentially the Weak Matching Pursuit [29]. The bound of the approximation rate of MVMP in [30].

2) Minimization with respect to the atom scale parameters: Once the atoms $\phi_{l_k},\psi_{r_k},k=1,...,m$ 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 of EM becomes:

$$\begin{aligned}E^{[t]} &= \frac{1}{2\sigma_I^2}(\|\mathbf{y}_L - \sum_{k=1}^m a_{l_k}\phi_{l_k}\|_2^2 + \|\mathbf{y}_R - \sum_{k=1}^m b_{r_k}\psi_{r_k}\|_2^2) + \frac{1}{2\sigma_D^2}\sum_{k=1}^m W_{lr}\\ &+ \frac{1}{2\sigma_b^2}\sum_{k=1}^m(b_{r_k} - \frac{a_{l_k}}{\sqrt{J_k}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0).\end{aligned} \tag{45}$$

It can be observed that the energy function $E^{[t]}$ is actually an analytic function of the atom parameters $\{\gamma_l^{(L)}\}$ and $\{\gamma_r^{(R)}\}$ in the case of parametric dictionaries as defined in Sec. 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 $E^{[t]}$ with respect to $\gamma_l^{(L)}$ and $\gamma_r^{(R)}$, given the coefficients $\mathbf{a}$ and $\mathbf{b}$.

The sparse coding and the learning steps are iteratively repeated until the convergence is achieved. The inference should be, however, performed from a large set of data, i.e., from different multi-view pairs and with different camera poses. This is achieved by sparse coding of a randomly selected set of stereo pairs yielding sparse coefficients for each pair. Then the energy function is averaged over all pairs to perform the learning step.

VI. Learning for stereo omnidirectional images

A. Implementation for spherical images

The stereo image model given in Eq. (2) does not put any assumption on the type of cameras used for stereo image acquisition. It can be applied to planar or omnidirectional multi-view images by defining the dictionary for the considered type of images, and by introducing the epipolar geometry constraints that are defined for that particular image projection geometry.

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

Since omnidirectional cameras represent a convenient solution for 3D scene representation due to their wide field of view, we perform learning of dictionaries for stereo omnidirectional images.

Omnidirectional images obtained by catadioptric cameras can be appropriately mapped to spherical images [31]. Therefore, we proposed to use a dictionary of atoms on the 2-D unit sphere as proposed in [32] in order to represent spherical images. The 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:

$$g(\theta,\varphi) = -\frac{1}{K_A}\left(16\alpha^2\tan^2\frac{\theta}{2}\cos^2\varphi - 2\right)\exp\left(-4\tan^2\frac{\theta}{2}\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right). \tag{46}$$

For the weighting function in the epipolar geometry constraints of Eq. (33) we use a Gaussian envelope of the form:

$$w(\theta,\varphi) = \frac{1}{K_G}\exp\left(-4\tan^2\frac{\theta}{2}\left(12\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right). \tag{47}$$

It 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 $\gamma = (\tau,\nu,\psi,\alpha,\beta) \in \Gamma$. The triplet (τ,ν,ψ) 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 ψ. Parameters α and β represent anisotropic scaling factors.

Epipolar geometry for the spherical camera model is formulated in the same manner as for the perspective camera model, except that the pixel (point) positions $\mathbf{v}$ and $\mathbf{u}$ in Eq. (6) are expressed in spherical coordinates. The transform $\mathbf{u} = Q_{lr}(\mathbf{v})$ that relates a point $\mathbf{v}$ on atom $\phi_l = g_{\gamma_l^{(L)}}$ to its corresponding transformed point $\mathbf{u}$ on the transformed atom $\psi_r = g_{\gamma_r^{(R)}}$ can be defined via the linear transform of the coordinate system. When the atoms ϕ_l and ψ_r are derived using the parameters $\gamma_l^{(L)} = (\tau_l^{(L)},\nu_l^{(L)},\psi_l^{(L)},\alpha_l^{(L)},\beta_l^{(L)})$ and $\gamma_r^{(R)} = (\tau_r^{(R)},\nu_r^{(R)},\psi_r^{(R)},\alpha_r^{(R)},\beta_r^{(R)})$, respectively, the point $\mathbf{u}$ can be written as:

$$\mathbf{u} = \mathbf{R}^{-1}_{\gamma_r^{(R)}} \cdot \zeta(\mathbf{R}_{\gamma_l^{(L)}} \cdot \mathbf{v}), \tag{48}$$

where $\mathbf{R}_{\gamma_l^{(L)}}$ and $\mathbf{R}_{\gamma_r^{(R)}}$ are rotation matrices given by Euler angles $(\tau_l^{(L)},\nu_l^{(L)},\psi_l^{(L)})$ and $(\tau_r^{(R)},\nu_r^{(R)},\psi_r^{(R)})$, respectively, and $\zeta(\cdot)$ defines the grid transform due to anisotropic scaling. If the spherical coordinates of $\mathbf{v}$ are denoted as (θ_2,φ_2) (unit sphere is assumed), and the spherical coordinates of $\mathbf{u}$ are (θ_1,φ_1), the function $\zeta(\cdot)$ is a pair of transforms $\zeta_p(\varphi_1)$ and $\zeta_t(\theta_1,\varphi_1,\zeta_p(\varphi_1))$ given by:

$$\begin{aligned}
\varphi_2 &= \zeta_p(\varphi_1) = \arctan\left(\frac{\alpha_r^{(R)}\beta_l^{(L)}\sin\varphi_1}{\alpha_l^{(L)}\beta_r^{(R)}\cos\varphi_1}\right) \\
\theta_2 &= \zeta_t(\theta_1,\varphi_1,\varphi_2) = \arctan\left[\tan\frac{\theta_1}{2}\sqrt{\frac{(\alpha_l^{(L)})^2\cos^2\varphi_1 + (\beta_l^{(L)})^2\sin^2\varphi_1}{(\alpha_r^{(R)})^2\cos^2\varphi_2 + (\beta_r^{(R)})^2\sin^2\varphi_2}}\right].
\end{aligned} \tag{49}$$

This is followed by the transform of spherical coordinates (θ_2,φ_2) to Euclidean coordinates $\mathbf{u}_t$. Finally we obtain $\mathbf{u}$ as $\mathbf{R}^{-1}_{\gamma_r^{(R)}}\mathbf{u}_t$.

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 a generating function, we focus on the learning of scaling parameters only. The scaling parameters are the most important parameters since they directly define the shape of the atoms, which become elongated as the scales become anisotropic. On the other hand, translations and rotations depend highly on the distance and orientations of cameras, so learning these parameters is meaningful only when cameras are static and 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 Eq. (38) is minimized only with respect to these four parameters. We have performed the minimization using the conjugate gradient[2]. The other parameters are kept fixed. The motion parameters (τ,ν) include the positions of all pixels in an image. The rotation parameter ψ sampled uniformly between 0 to π with resolution a 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[3]. The database consists of 54 omnidirectional images of the indoor environment, grouped into two sets: one set without plants (27

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

[3]Database is available upon request.

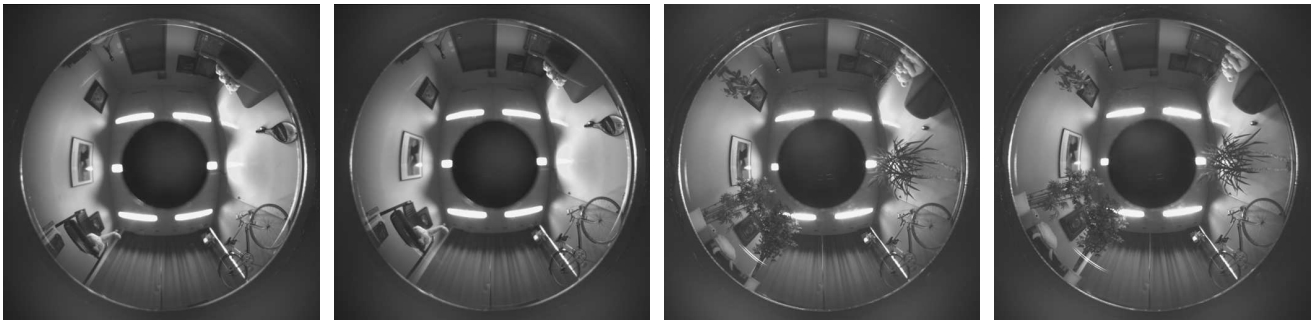


Fig. 2. (a),(b): Two views from the "Mede" database, first image set. (c),(d): Two views from the "Mede" database, second image set.

images), and one 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 taken in the same set, 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 on Figure 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 step we form a set of S_p pairs of stereo patches. For each $p = 1, ..., S_p$ 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 on the 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 Eq. (45) for all patches.

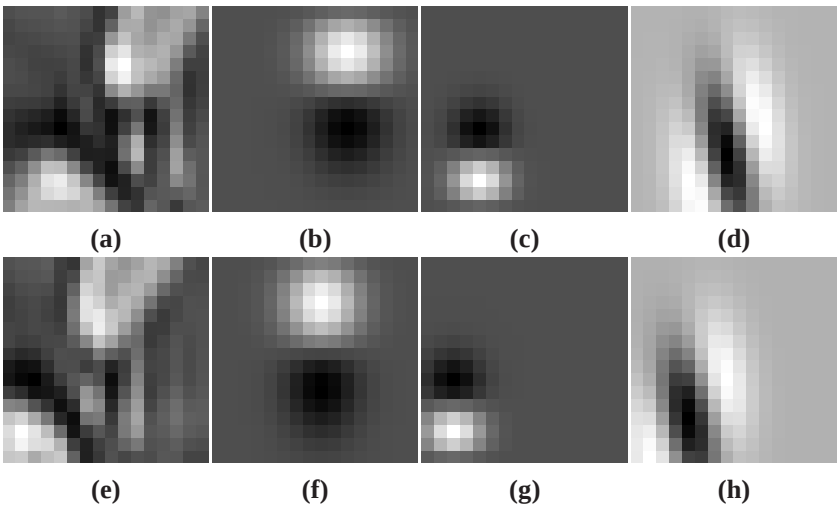


Fig. 3. Example of a pair of stereo patches and their MVMP selected atoms: a) left patch, b)-d) the first three atoms in the MVMP decomposition of the left patch; e) right patch, f)-h) the first three atoms in the MVMP decomposition of the right patch.

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 filtering [33] in order to flatten the image spectrum and make all frequencies equally important, as proposed in [3]. The number of positions in the dictionary construction is 12×12, 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 illustrate in this section the dictionary that has been learned on 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:

$$\begin{aligned}\tilde{E}(\mathbf{y}_L, \mathbf{y}_R, D=0, \mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) &= \frac{1}{2\sigma_I^2}(\|\mathbf{y}_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|\mathbf{y}_R - \mathbf{\Psi}\mathbf{b}\|_2^2) \\ &+ \rho\frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)D_E(\gamma_l^{(L)}, \gamma_r^{(R)}) \\ &+ \rho\frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \end{aligned} \quad (50)$$

We can see that for $\rho = 1$, the energy function in Eq. (50) is equal to the one in Eq. (38). On the other hand, for $\rho = 0$, there are no multi-view constraints in the energy function, and dictionary learning is based only on minimization of the residual energy of both stereo images under sparse representations.

The initial values of scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$ for the learning algorithm 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 on Fig. 4. We then learn the scaling parameters with 50 iterations of the EM algorithm. At this point, the change in parameters became small and the solution can be considered to be stable. The columns 3 to 10 in Table I give the values of the learned scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$. The results are given for $\rho = 0, 1, 3, 5$, where the same initial scales have been used for all values of ρ. The corresponding atoms are shown in Fig. 4.

TABLE I
INITIAL AND LEARNED SCALE PARAMETERS FOR THE LEFT AND RIGHT IMAGE, 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

We observe first that the learned dictionaries include atoms of different scales and 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 [3]. 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 $\rho = 0$. In addition, for $\rho > 0$ the learned scales generally tend to give smaller atoms than for $\rho = 0$. These two effects result from the multi-view geometry constraints in the dictionary learning process. They are most probably due to the local nature of the epipolar constraint. Namely, the depth of the scene rapidly changes around object boundaries leading to different disparity and epipolar matching. Since the object boundaries are represented by 2D discontinuities on the image of a 3D scene, the epipolar geometry is mostly satisfied along the discontinuity and in a limited area. This makes the learned atoms become anisotropic and small. This highlights the great importance of the geometric constraints in the learning of dictionaries for stereo images.

D. Application to 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

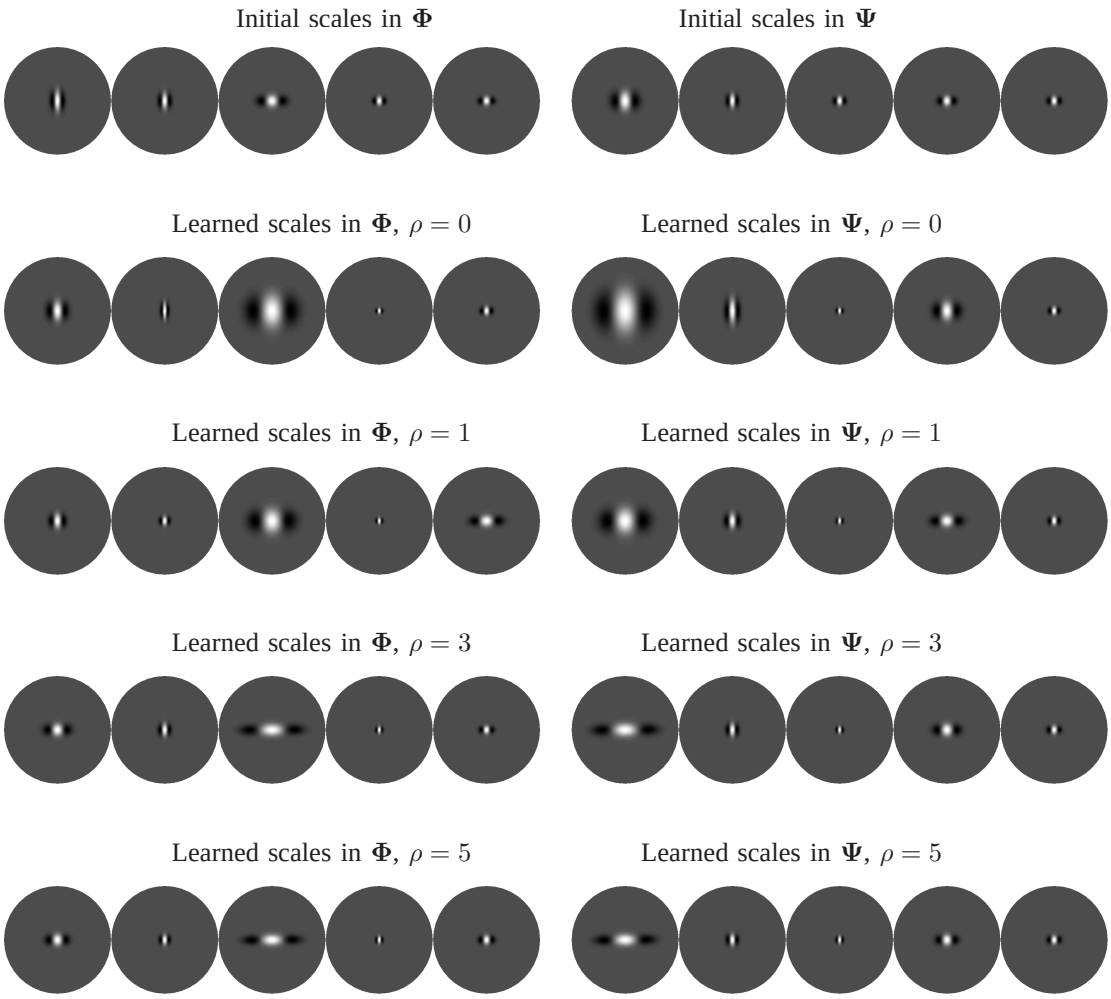
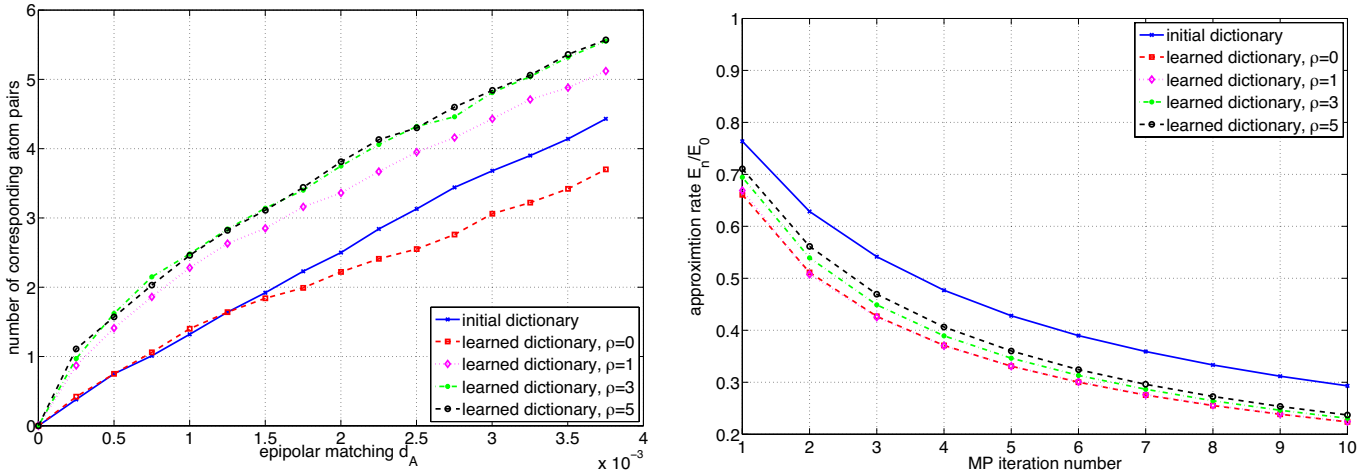


Fig. 4. Initial and learned scales of the atoms for the left and right dictionaries. All atoms are on the North pole.

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 the 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 [2]. If the atoms that form the learned dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ 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 decomposition of the left and right patch using 10 atoms per patch, the epipolar constraints are evaluated for all possible pairs of left and right atoms $d_A(\phi_l, \psi_r)$, $l = 1, ..., 10$, $r = 1, ..., 10$. The epipolar measure d_A is equal to half of the value W_{lr} in Eq. (33), and represents the epipolar atom distance per view.

Fig. 5(a) plots on the y-axis the number of atom pairs (ϕ_l, ψ_r) that have the epipolar distance $d_A(\phi_l, \psi_r)$ 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 smaller 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 algorithms really produces dictionaries that result into 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 multiview geometry can lead to suboptimal dictionaries for stereo images. Fig. 5 (b) shows the average approximation rate (i.e., energy decay) during

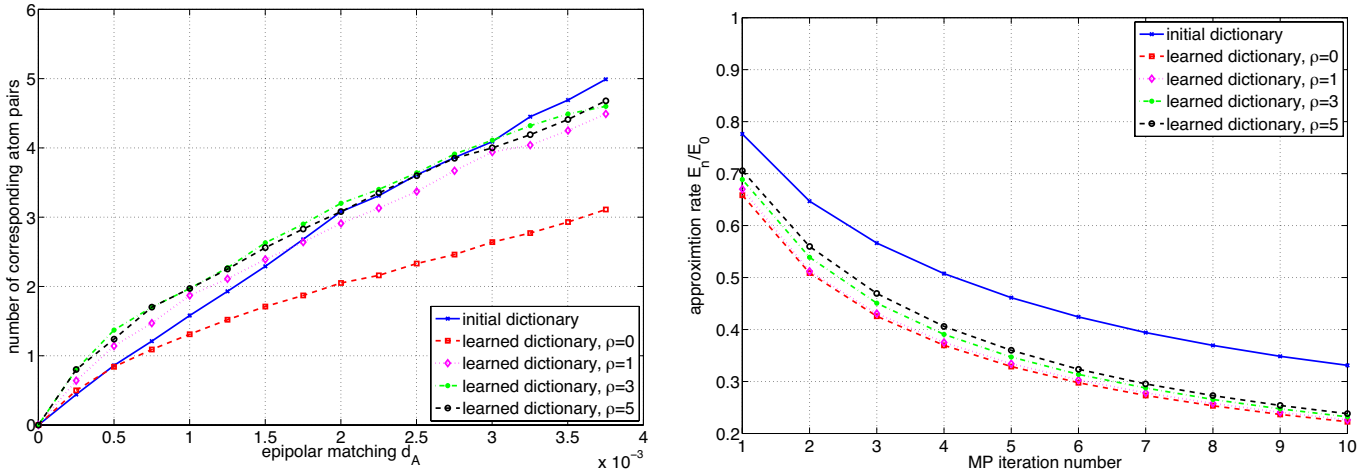
the iterations of the MP for images $\mathbf{y}_L$ and $\mathbf{y}_R$. We plot the ratio between the sum of the residues of the left and right images after k iterations and the sum of their initial energies, versus the iteration number k. We can see that for all values of ρ the approximation rate using learned dictionaries is better than using random dictionaries. Moreover, increasing ρ slows down the approximation rate for a very small amount, hence optimizing the dictionaries for stereo matching does not induce a big penalty on the approximation rate.



(a) **(b)**

Fig. 5. 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 $\mathbf{y}_R$.

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 on Fig. 6 (a). We see that the learned dictionaries still give more correspondences than the random ones for the 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 on Fig. 6 (b), and it is similar to the results of Fig. 5 (b). Since learned dictionaries result in a higher number of correspondences with no penalty in the approximation rate, these dictionaries can be beneficial for distributed scene representation and distributed coding.



(a) **(b)**

Fig. 6. 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 $\mathbf{y}_R$.

E. Application to camera pose estimation

Finally, we discuss how the proposed learning algorithm can be useful for camera pose estimation [34]. We have performed an experiment where the camera pose has been estimated using the initial and learned dictionaries for $\rho = 3$. Two pairs of

images from "Mede" database have been selected, with translation $\mathbf{T} = [1 \ 0 \ 0]^T$ and $\mathbf{T} = [0 \ 1 \ 0]^T$ respectively. For each image pair, we have randomly chosen 20 patches of 40×40 pixels. The same image patches are decomposed by MP using initial and learned dictionaries, giving 3 atoms for each patch and thus 60 atoms per image. Matching atoms from left and right images and estimation of the camera pose has been performed using the algorithm proposed in [34]. 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, from which we estimate the camera pose using the eight-point algorithm [35]. Additionally, we apply Ransac [36] for camera pose estimation in order to be more robust to outliers. The estimated translation matrices are shown in Table II and Table III for both target matrices $\mathbf{T}$. We can see that using the learned dictionary for pose estimation gives significantly better performance than using a randomly initialized dictionary. The learned dictionary leads to a precise estimation of the translation, while initial dictionary cannot even determine the direction of the camera motion. Moreover, applying Ransac does not improve the performance in the case of the learned dictionary. This leads to the conclusion that all points on the learned atoms give reliable matches and no gross outliers.

TABLE II
CAMERA POSE ESTIMATION WITH RANDOM INITIAL AND LEARNED DICTIONARIES, FOR TARGET TRANSLATION $\mathbf{T}^\mathsf{T} = [1 \ 0 \ 0]$.

Target matrices	$\mathbf{T}^\mathsf{T} = [1 \ 0 \ 0]$ $\mathbf{R} = \mathbf{I}$					
	learned dictionary			initial dictionary		
estimated $\mathbf{T}^\mathsf{T}$, without Ransac	[0.9778	-0.1519	0.1441]	[0.1910	-0.7284	0.6580]
estimated $\mathbf{T}^\mathsf{T}$, with Ransac	[0.8144	-0.5436	0.2032]	[0.7540	0.6067	0.2518]

TABLE III
CAMERA POSE ESTIMATION WITH RANDOM INITIAL AND LEARNED DICTIONARIES, FOR TARGET TRANSLATION $\mathbf{T}^\mathsf{T} = [0 \ 1 \ 0]$.

Target matrices	$\mathbf{T}^\mathsf{T} = [0 \ 1 \ 0]$ $\mathbf{R} = \mathbf{I}$					
	learned dictionary			initial dictionary		
estimated $\mathbf{T}^\mathsf{T}$, without Ransac	[-0.0385	0.9951	0.0907]	[0.9067	0.3484	0.2376]
estimated $\mathbf{T}^\mathsf{T}$, with Ransac	[-0.1317	0.9424	0.3075]	[0.9003	0.3934	0.1866]

VII. CONCLUSIONS

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

VIII. ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation under grant 200020-120063, and by the EU under the FP7 project APIDIS (ICT-216023). The authors would like to thank the members of the Redwood Center for Theoretical Neuroscience at UC Berkeley, for the fruitful discussions on the ML dictionary learning methods.

APPENDIX A
CONDITIONAL PROBABILITIES $P(b_r|a_l, \phi_l, \psi_r)$ AND $P(a_l|b_r, \phi_l, \psi_r)$

We first replace the expansions for $\mathbf{y}_L$ and $\mathbf{y}_R$ from Eq. (1) in Eq. (18), and get for all $k = 1, ..., m$:

$$\langle \sum_{i=1}^{m} b_{r_i} \psi_{r_i}, \psi_{r_k} \rangle + \langle \mathbf{e}_R, \psi_{r_k} \rangle = \frac{1}{\sqrt{J_{l_k r_k}}} \langle \sum_{i=1}^{m} a_{l_i} \phi_{l_i}, \phi_{l_k} \rangle + \frac{1}{\sqrt{J_{l_k r_k}}} \langle \mathbf{e}_L, \phi_{l_k} \rangle, \tag{51}$$

which can be rewritten as:

$$b_{r_k} + \sum_{\substack{i=1 \\ i \neq k}}^{m} \langle b_{r_i} \psi_{r_i}, \psi_{r_k} \rangle + \langle \mathbf{e}_R, \psi_{r_k} \rangle = \frac{1}{\sqrt{J_{l_k r_k}}} a_{l_k} + \frac{1}{\sqrt{J_{l_k r_k}}} \sum_{\substack{i=1 \\ i \neq k}}^{m} \langle a_{l_i} \phi_{l_i}, \phi_{l_k} \rangle + \frac{1}{\sqrt{J_{l_k r_k}}} \langle \mathbf{e}_L, \phi_{l_k} \rangle, \tag{52}$$

or simply:

$$b_{r_k} = \frac{a_{l_k}}{\sqrt{J_{l_k r_k}}} + \eta', \tag{53}$$

where:

$$\eta' = \frac{1}{\sqrt{J_{l_k r_k}}} \sum_{\substack{i=1 \\ i \neq k}}^{m} \langle a_{l_i} \phi_{l_i}, \phi_{l_k} \rangle + \frac{1}{\sqrt{J_{l_k r_k}}} \langle \mathbf{e}_L, \phi_{l_k} \rangle - \sum_{\substack{i=1 \\ i \neq k}}^{m} \langle b_{r_i} \psi_{r_i}, \psi_{r_k} \rangle - \langle \mathbf{e}_R, \psi_{r_k} \rangle. \tag{54}$$

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 Eq. (53) to derive the distribution $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ because the sparse support of the stereo images is not known, and hence also the indexes l_k, r_k and $k = 1, ..., m$. Therefore, we say that an arbitrary stereo atom pair ϕ_l, ψ_r and their coefficients a_l, a_r satisfy Eq. (53) up to a certain error η_1, which includes also η'. Therefore, we have:

$$b_r = \frac{1}{\sqrt{J_{lr}}} a_l + \eta_1, \tag{55}$$

where J_{lr} is the Jacobian of the linear transform of the coordinate system induced by the transform between atoms ϕ_l and ψ_r. When a_l and b_r are the coefficients of a stereo pair, then they satisfy Eq. (55) 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 Eq. (1)) and hence the noise η_1 is also small. Therefore, we can model the noise η_1 with a white Gaussian noise of variance σ_b^2 and get:

$$P(\eta_1) = P(b_r|a_l, \phi_l, \psi_r) = \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 \right). \tag{56}$$

Although ϕ_l, ψ_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 ϕ_l and ψ_r. Multiplying Eq. (53) with $\sqrt{J_{lr}}$, we can get a symmetric relation:

$$\begin{aligned} P(\eta_2) = P(a_l|b_r, \phi_l, \psi_r) &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_a^2} (a_l - \sqrt{J_{lr}} b_r)^2 \right) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2 J_{lr}} (a_l - \sqrt{J_{lr}} b_r)^2 \right) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 \right), \end{aligned} \tag{57}$$

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 the conditional probability $P(a_l|b_r, \phi_l, \psi_r)$ would be obtained if we consider the inverse transform F_{rl} from atom ψ_r to atom ϕ_l because the Jacobian of the linear transform satisfies: $J(Q_{lr}^{-1}) = 1/J_{lr}$.

REFERENCES

[1] Hartley R.I. and Zisserman A., *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[2] Tošić I. and Frossard P., "Geometry-Based Distributed Scene Representation With Omnidirectional Vision Sensors," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1033–1046, 2008.

[3] Olshausen B. and Field D., "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–25, 1997.

[4] Kreutz-Delgado K., Murray J., Rao B., Engan K., Lee T.-W. and Sejnowski T. J., "Dictionary Learning Algorithms for Sparse Representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[5] Aharon M., Elad M. and Bruckstein A., "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311– 4322, 2006.

[6] Olshausen B. A., "Learning sparse, overcomplete representations of time-varying natural images," in *Proceedings of the IEEE International Conference on Image Processing*, 2003.

[7] Olshausen B. A., Cadieu C., Culpepper B. J., and Warland D. K., "Bilinear Models of Natural Images," in *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging*, 2007.

[8] Cadieu C. and Olshausen B. A., "Learning Transformational Invariants from Time-Varying Natural Images," in *Proceedings of the Conference on Neural Information Processing Systems*, 2008.

[9] Olshausen B. A. and Field D. J., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996, springer-Verlag New YORK INC.

[10] Engan K., Aase S. and Hakon Husoy J., "Method of optimal directions for frame design," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[11] Engan K., Rao B. D. and Kreutz-Delgado K., "Frame design using FOCUSS with method of optimal directions (MOD)," in *Proceedings of the Norwegian Signal Processing Symposium*, 1999.

[12] Gorodnitsky I. and Rao B., "Sparse Signal Reconstruction from Limited Data Using FOCUSS: a Re-weighted Minimum Norm Algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[13] Schmid-Saugeon P. and Zakhor A., "Dictionary design for matching pursuit and application to motion-compensated video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 6, pp. 880– 886, 2004.

[14] ——, “Learning dictionaries for matching pursuits based video coders,” in *Proceedings of the IEEE International Conference on Image Processing*, 2001.
[15] Gribonval R. and Nielsen M., “Sparse representations in unions of bases,” *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.
[16] Jost P., Vandergheynst P., Lesage S. and Gribonval R., “MoTIF: an efficient algorithm for learning translation invariant dictionaries,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.
[17] Engan K., Skretting K. and Husřy J., “Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation,” *Digital Signal Processing*, vol. 17, no. 1, pp. 32–49, 2007.
[18] Monaci G., Sommer F. and Vandergheynst P., “Learning sparse generative models of audiovisual signals,” in *Proceedings of the European Signal Processing Conference*, 2008.
[19] ——, “Learning sparse generative models of audiovisual signals,” *submitted to IEEE Transactions on Neural Networks*, 2008.
[20] Hoyer P. and Hyvärinen A., “Independent component analysis applied to feature extraction from colour and stereo images,” *Network: Computation in Neural Systems*, vol. 11, no. 3, pp. 191–210, 2000.
[21] Okajima K., “Binocular disparity encoding cells generated through an Infomax based learning algorithm,” *Neural Networks*, vol. 17, no. 7, pp. 953–962, 2004.
[22] Tošić I. and Frossard P., “Conditions for recovery of sparse signals correlated by local transforms,” *Proceedings of the IEEE International Symposium on Information Theory*, 2009.
[23] Dempster A. P., Laird N. M. and Rubin D. B., “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38.
[24] Neal R. M. and Hinton G. E., *A view of the EM algorithm that justifies incremental, sparse, and other variants*. in *Learning in Graphical Models*, edited by M. I. Jordan, Dordrecht: Kluwer Academic Publishers, 1998, pp. 355–368.
[25] Culpepper B. J., “Learning ‘what’ and ‘where’ from movies,” Master’s thesis, UC Berkeley, 2007.
[26] Mallat S. G. and Zhang Z., “Matching Pursuits With Time-Frequency Dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
[27] Figueras i Ventura R. M., Vandergheynst P. and Frossard P., “Low rate and flexible image coding with redundant representations,” *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 726–739, 2006.
[28] Frossard P., “Robust and Multiresolution Video Delivery: From H.26x to Matching Pursuit Based Technologies,” PhD Thesis, EPFL, 2000.
[29] Temlyakov V. N., “Weak greedy algorithms,” *Advances in Computational Mathematics*, vol. 12, no. 2-3, pp. 213–227, 2000.
[30] Tošić I., “On unifying sparsity and geometry for image-based scene represenation,” PhD Thesis, EPFL, 2009.
[31] Tošić I. and Frossard P., *Spherical imaging in omni-directional camera networks*. in *Multi-Camera Networks, Principles and Applications*, edited by Aghajan H. and Cavallaro A., Academic press, 2009.
[32] Tošić I., Frossard P. and Vandergheynst P., “Progressive Coding of 3-D Objects Based on Overcomplete Decompositions,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 11, pp. 1338–1349, 2006.
[33] Tošić I., Bogdanova I., Frossard P. and Vandergheynst P., “Multiresolution Motion Estimation for Omnidirectional Images,” in *Proceedings of the European Signal Processing Conference*, 2005.
[34] Tošić I. and Frossard P., “Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images,” in *Proceedings of the European Signal Processing Conference*, 2007.
[35] Ma Y., Soatto S., Košecká J. and Sastry S. S., *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, 2004.
[36] Fischler M. and Bolles R., “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.