Learning of Stereo Visual Dictionaries

Ivana Tošić and Pascal Frossard
Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
{ivana.tosic, pascal.frossard}@epfl.ch

***Abstract*—We present a novel method for learning overcomplete dictionaries that are optimized for stereo image representation. Our learning algorithm optimizes the construction of dictionaries for both efficient image approximation and epipolar matching between pairs of images. The multi-view geometry constraint is included in the probabilistic modeling that permits to maximize the likelihood that natural stereo images are efficiently represented with the selected dictionaries. Experimental results of dictionary learning for stereo omnidirectional images show that the multi-view constraints significantly influence the construction of the dictionary, which leads to atoms with high anisotropic characteristics.**

I. Introduction

Multiple images of a scene taken from different viewpoints contain information about both 3D structure and texture of the scene, giving a richer perception of the environment compared to a single view. However, dealing with this high dimensional visual information poses many challenges, such as multi-view compression and geometry estimation. The most important requirement in these challenges is to have an appropriate multi-view image model. The multi-view image model based on sparse image approximations with overcomplete dictionaries of geometrical atoms has shown good performance in distributed multi-view coding [1]. In this model, each image is approximated by a linear combination of meaningful features that represent the visual information of the scene. Atoms in different views are related with local geometric transforms that satisfy the multi-view geometry constraints. However, the choice of the dictionary in [1] is empirical and not optimized for multi-view imaging. Certainly, by adapting the dictionary to the case of multi-view imaging, we can expect to achieve better performance in various applications.

This paper targets the problem of learning a dictionary adapted to the multi-view image representation model. Maximum likelihood (ML) dictionary learning for natural monocular images has been introduced by Olshausen and Field in 1997 [2]. However, there has been little work targeting the problem of learning stereo overcomplete dictionaries. Hoyer and Hyvärinen have applied independent component analysis (ICA) to learn the orthogonal basis of stereo images [3]. Their algorithm resulted in Gabor-like basis functions tuned to different disparities. Okajima has proposed a learning approach that maximizes the mutual information between the stereo image model and the disparity [4]. They have obtained results similar to Hoyer and Hyvärinen. The stereo learning methods in [3] and [4] have been primarily designed for the purpose of studying the receptive fields of binocular cells in the primary visual cortex.

We study here the design of stereo dictionaries that have the optimal properties for both image approximation and disparity or 3D scene structure estimation. We assume a sparse stereo image model and we learn overcomplete dictionaries. Motivated by the good performance of the ML methods for monocular images, we develop a novel ML method for learning stereo dictionaries optimized for both image approximation and epipolar matching. We include the epipolar geometry in the probabilistic modeling and hence match pairs of atoms within the learning process itself. The experimental results show that the dictionary atoms learned by our algorithm present high anisotropy characteristics and substantially differ from atoms in single view learning. This illustrates the importance of disparity matching for efficient stereo representations.

II. Stereo image model

Developing the ML dictionary learning method for stereo images requires a definition of the stereo image model. Since stereo images capture the same scene from different viewpoints, they are correlated by local transforms of image components. If we decompose each image into sparse components that capture the objects in the scene, we can assume that the most prominent components are present in both images, possibly under different local transforms [1]. Let us consider two images: left image y_L and right image y_R, which have sparse representations in dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, respectively. The images are approximated by sparse decompositions of m atoms up to an approximation error (e_L, resp. e_R), i.e.:

$$\begin{aligned} y_L &= \mathbf{\Phi a} = \sum_{k=1}^{m} a_{l_k} \phi_{l_k} + e_L, \\ y_R &= \mathbf{\Psi b} = \sum_{k=1}^{m} b_{r_k} \psi_{r_k} + e_R, \end{aligned} \tag{1}$$

where the set of indices $\{l_k\}$ and $\{r_k\}$, $k = 1, ..., m$ label the atoms that participate in the sparse decompositions of y_L and y_R, respectively. The atoms are ordered in both expansions, i.e., pairs of corresponding atoms are indexed with the same counting parameter k. The vectors $\mathbf{a}$ and $\mathbf{b}$ denote the atom coefficients in the sparse representation of the left and right image, respectively. We further assume that stereo images

contain similar atoms that are locally transformed:

$$y_R = \sum_{k=1}^{m} b_{r_k}\psi_{r_k} + e_R = \sum_{k=1}^{m} b_{r_k} F_{l_k r_k}(\phi_{l_k}) + e_R, \quad (2)$$

where $F_{l_k r_k}(\cdot)$ denotes the transform of an atom ϕ_{l_k} in y_L to an atom ψ_{r_k} in y_R, and it differs for each $k = 1, ..., m$. This model assumes that both stereo images are m-sparse, i.e., composed of m atoms. The motivation behind this is that left and right images typically contain image projections of the same 3D scene features, thus the number of sparse components will be approximately the same.

Due to the change of viewpoint on the 3D scene, various types of transforms are introduced in the image projective space. Most of these transforms can be represented by the 2-D translation, rotation and anisotropic scaling of the image features. Such transforms are efficiently represented with a parametric dictionary whose construction is built on these transforms. 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 rotation, translation and scaling parameters applied to the generating function g. This is equivalent to applying a unitary operator $U(\gamma)$ to the generating function g, i.e.: $g_\gamma = U(\gamma)g$. We define the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as structured dictionaries built on the same generating function g, but using different sets of parameters: Γ_L for $\mathbf{\Phi}$, and Γ_R for $\mathbf{\Psi}$. To simplify the notation, we introduce the following equivalencies: $\phi_l \equiv g_{\gamma_l^{(L)}}$, $\gamma_l^{(L)} \in \Gamma_L$, and $\psi_r \equiv g_{\gamma_r^{(R)}}$, $\gamma_r^{(R)} \in \Gamma_R$. When the dictionaries are defined this way, the transform of one atom ϕ_l to another atom ψ_r reduces to a transform of its parameters, i.e.,

$$F_{lr}(\phi_l) = F_{lr}(g_{\gamma_l^{(L)}}) = U(\gamma')g_{\gamma_l^{(L)}} = g_{\gamma_r^{(R)}} = \psi_r. \quad (3)$$

III. Stereo dictionary learning

We now formulate the probabilistic framework for the maximum likelihood learning of overcomplete dictionaries $\mathbf{\Phi}, \mathbf{\Psi}$ that are used to represent stereo images y_L and 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 by a set of atom pairs related by geometric transforms, under the sparsity prior. In other words, we want to simultaneously learn the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ that approximate well the stereo images y_L and y_R, given the sparse stereo image model in Eq. (2). Moreover, we want to maximize the probability that the stereo images given by the model (2) satisfy the epipolar constraint, i.e., that the epipolar distance between all corresponding points on y_L and y_R is equal to zero ($D = 0$). Maximization of this probability is crucial for learning dictionaries that have atoms with good epipolar matching properties, which is important in applications involving scene geometry estimation.

Formally, we want to solve the following optimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\max_{\mathbf{\Phi},\mathbf{\Psi}} \langle \max_{\mathbf{a},\mathbf{b}} \log P(y_L, y_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) \rangle. \quad (4)$$

Marginalizing over $\mathbf{a}$ and $\mathbf{b}$ we have that:

$$P(y_L, y_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) = \\ = \int\int P(y_L, y_R, D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) \mathrm{d}\mathbf{a}\mathrm{d}\mathbf{b}. \quad (5)$$

We first need to define the joint distribution of coefficients $\mathbf{a}$ and $\mathbf{b}$, given dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, denoted as $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$. Let us 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 atom transforms correctly represent the local object transforms. Under this assumption, for a stereo atom pair ϕ_l, ψ_r, corresponding to the same object in the scene and linked with a transform F_{lr}, the following equality holds (see Lemma 1 in [5]):

$$\langle y_R, \psi_r \rangle = \frac{1}{\sqrt{J_{lr}}} \langle y_L, \phi_l \rangle, \quad (6)$$

where J_{lr} is the Jacobian of the transform F_{lr}. Using the sparse image model and Eq. (6) we obtain the following probabilities:

$$\begin{aligned} P(b_r|a_l, \phi_l, \psi_r) &= P(a_l|b_r, \phi_l, \psi_r) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right), \end{aligned} \quad (7)$$

where z_b is the normalization factor and σ_b^2 is the variance of the zero-mean Gaussian noise that models the difference between b_r and $a_l/\sqrt{J_{lr}}$. We further assume that pairs of coefficients (a_l, b_r) are independent, which is usually the case when image decompositions are sparse enough. Then, the distribution $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ is factorial, i.e.:

$$P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = \prod_{l=1}^{M}\prod_{r=1}^{M} P(a_l, b_r|\phi_l, \psi_r) = \\ P(\mathbf{a})P(\mathbf{b}) \prod_{l=1}^{M}\prod_{r=1}^{M} \sqrt{P(b_r|a_l, \phi_l, \psi_r)P(a_l|b_r, \phi_l, \psi_r)}, \quad (8)$$

where we assume that priors on coefficients in each image $P(a_l)$ and $P(b_r)$ are independent of the atoms. Although in reality the distribution of the coefficients would depend on an arbitrarily chosen dictionary, imposing the independence of the coefficients with respect to the dictionary during learning would actually lead to inferring a dictionary that gives the same prior distribution of coefficients for all types of images.

For modeling the priors on coefficients, we assume that the coefficients a_l and b_r are i.i.d. and drawn from a Bernoulli distribution over the activity of coefficients, where a coefficient is different from zero with probability p and equal to zero with probability q. Thus, for $p \ll q$ the Bernoulli distribution can well model the prior on the sparse coefficients $\mathbf{a}$ and $\mathbf{b}$. If we take $p = 1/(1 + e^{1/\lambda})$, we have:

$$P(\mathbf{a}) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{a}\|_0}{\lambda}\right) \text{ and } P(\mathbf{b}) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{b}\|_0}{\lambda}\right),$$

where $\|\cdot\|_0$ denotes the l_0 norm and λ controls the level of "sparseness" of coefficients. For sparse vectors $\mathbf{a}$ and $\mathbf{b}$,

the probabilities $P(\mathbf{a})$ and $P(\mathbf{b})$ are highly peaked at zero. Thus, we can approximate the probability $P(\mathbf{a},\mathbf{b}|\mathbf{\Phi},\mathbf{\Psi})$ by its value at the maximum, since it is a product of a zero-mean Gaussian distribution and discrete distributions tightly peaked at zero [2]. Eq. (5) then becomes:

$$P(y_L, y_R, D=0|\mathbf{\Phi},\mathbf{\Psi}) \approx P(y_L, y_R|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) \cdot \cdot P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) P(\mathbf{a},\mathbf{b}|\mathbf{\Phi},\mathbf{\Psi}), \quad (9)$$

where we have used the fact that $D=0$ does not bring more information to y_L, y_R than $\mathbf{\Phi},\mathbf{\Psi}$. To evaluate our likelihood function, we next need to find the probability that the epipolar distance D is equal to zero given the stereo image model in Eq. (2), i.e., we need to find $P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})$. The probability of epipolar matching for the stereo image pair can be modeled by the product of probabilities of epipolar matching for pairs of atoms that participate in sparse decompositions of the left and the right image, i.e. whose coefficients a_l and b_r are different from zero. If the epipolar estimation error is assumed to be Gaussian of zero mean and variance σ_D^2, we can model the probability $P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})$ as:

$$P(D=0|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) = \frac{1}{z_D}\exp\left(-\frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)D_E(\phi_l,\psi_r)\right). \quad (10)$$

where $\mathcal{I}$ denotes the indicator function, z_D is the normalization factor and $D_E(\phi_l,\psi_r)$ is the epipolar distance between stereo atoms. This distance can be easily evaluated by summing over the epipolar distances between points paired by the local transform between the corresponding atoms.

At this point, we have defined all components of the objective maximum likelihood function in Eq. (9), except $P(y_L, y_R|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})$. This probability can be modeled by a Gaussian white noise of variance σ_I^2:

$$P(y_L, y_R|\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) = P(e_L + e_R) = \frac{1}{z_I}\exp\left(-\frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2)\right), \quad (11)$$

where we have used the fact that the sum of two zero-mean Gaussian random variables is also a zero-mean Gaussian random variable, and z_I is the normalization factor. We can now rewrite the ML learning problem in Eq. (4) as the following energy minimization problem:

$$(\mathbf{\Phi},\mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}}\langle\min_{\mathbf{a},\mathbf{b}} E(\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi})\rangle, \quad (12)$$

where E denotes the energy function given as:

$$E(\mathbf{a},\mathbf{b},\mathbf{\Phi},\mathbf{\Psi}) = \frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2) + \frac{1}{2\sigma_D^2}\sum_{l=1}^{M}\sum_{r=1}^{M}\mathcal{I}(a_l)\mathcal{I}(b_r)D_E(\phi_l,\psi_r) + \frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \quad (13)$$

The energy function thus consists of four summation terms: 1) the approximation error term; 2) the epipolar constraint term; 3) the coefficient similarity term; and 4) the sparsity term. Unfortunately, the obtained energy function is not convex. We use the Expectation-Maximization (EM) algorithm to find a local minimum. EM alternates between two steps:

- **E step**, that minimizes the energy over the coefficients $\mathbf{a}$ and $\mathbf{b}$, while keeping the dictionaries fixed. Coefficients are found using a modified version of the Matching Pursuit (MP) algorithm. It selects the atoms that give the minimal value of the energy function, and then removes the contribution of those atoms from the stereo images. Thus, it selects m atoms for each of the stereo images.
- **M step**, that minimizes the energy over the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, while keeping the coefficients fixed. This can be performed using the conjugate gradient method.

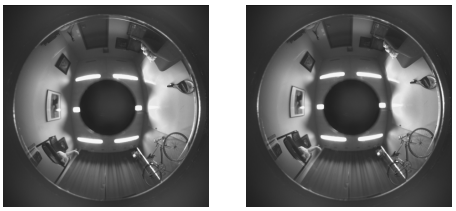
In the first iteration, the dictionaries are initialized randomly. The following iterations take the values for the coefficients and the dictionaries from the previous iteration. The E step and M steps are iteratively repeated until the convergence is achieved. The learning should be performed from a large set of data, i.e., from different multi-view image pairs.

IV. Experimental results

In practical applications, the benefits of the stereo image model in Eq. (2) depend on the discretization of the dictionary parameters: translations, rotations and scaling. Among those, the scaling parameters are the most important since they directly define the shape of atoms. As translations and orientations are highly dependent on the position of the sensors, we choose here to focus on learning only the scaling parameters of the atoms. A parametric dictionary is then constructed by applying to the generating function the learned scales and a discretized set of translations and orientations.

The stereo image model given in Eq. (2) does not put any assumption on the type of cameras used for stereo image acquisition. As omnidirectional cameras are suitable for capturing 3D scenes, we perform the learning for stereo omnidirectional images mapped to spherical images. For representing spherical images, we use the formulation of a dictionary on the 2-D unit sphere [1]. The generating function is a Gaussian in one direction and its second derivative in the orthogonal direction. We have tested the proposed stereo dictionary learning algorithm on our "Mede" database, which consists of 54 multi-view omnidirectional images of an indoor environment. Two views from the database are shown in Fig. 1. From three different scenes, we have formed 216 pairs of images with different translation $\mathbf{T}$ between cameras, while the rotation $\mathbf{R}$ is identity. The database is constructed from a variety of images such that the learned dictionaries can be used afterwards on images outside the training set.

We learn here five pairs of scaling parameters. 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 in Table I. The atoms of the initial scales are shown in the first row in Fig. 2. The whole



(a) **(b)**

Fig. 1. Two views from the "Mede" database.

dictionary is built from these atoms by shifting them at all pixel locations and rotating in four orientations. To see the influence of the part of the objective function that relies on the multi-view constraint, we have introduced a factor ρ that multiplies the second and the third term in the energy function. When $\rho = 0$ the learning takes into account only the image approximation term, while increasing ρ puts more importance on the multi-view correlation terms. From Fig. 2 we can see that for $\rho = 0$, the learned atoms are more elongated along the Gaussian direction, while narrower on the direction of the second derivative of the Gaussian. These results are in consistency with the previous work on dictionary learning for image representation [2]. However, when we increase ρ we obtain different results for atoms scales (see Table I). The atoms become more elongated along the direction of the Gaussian second derivative and narrower in the direction of the Gaussian. In addition, for $\rho > 0$ the learned scales generally tend to give smaller atoms than for $\rho = 0$. These two effects of including multi-view geometry in the dictionary learning process are due to the local nature of the epipolar constraint. Namely, the depth of the scene changes rapidly around object boundaries leading to different disparity and epipolar matching in these areas. Since the object boundaries are represented by 2D discontinuities on the image of a 3D scene, the epipolar geometry is satisfied along the discontinuity and in a limited area. This makes the learned atoms become anisotropic and small and different from the learned atoms in the single-view case. Therefore, the geometric constraints need to be introduced in the probabilistic model for stereo dictionary learning. Finally, we observe that the dictionaries for both images are very similar since the learning strategy is completely symmetric.

V. Conclusion

This paper proposes a novel method for learning the overcomplete dictionaries that have optimal performance in representing stereo images. The stereo image model based on sparse approximations over parametric dictionaries has served as basis for developing a maximum likelihood (ML) method for stereo dictionary learning. The experimental results have shown that that the obtained atoms significantly differ from atoms typically obtained by learning from single-view images. Therefore, one has to consider the geometric constraints in designing efficient representation strategies for stereo images.

TABLE I
Initial and learned scale parameters for the left and the right image, for different values of the parameter ρ.

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

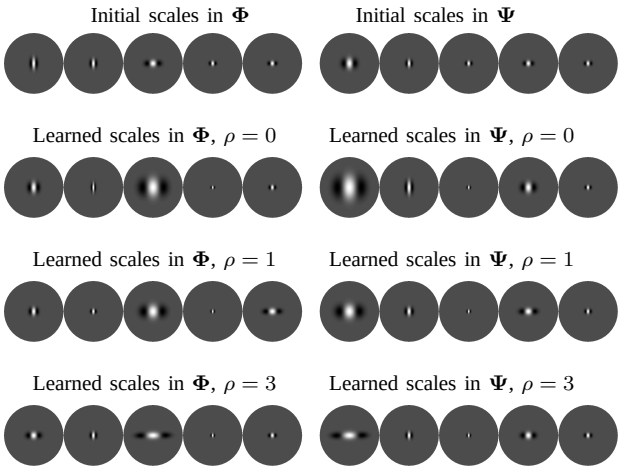


Fig. 2. Subset of atoms in the initial and learned dictionaries for the left and right images. All atoms are on the North pole.

VI. 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 dictionary learning research.

References

[1] 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.

[2] 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.

[3] 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.

[4] Okajima K., "Binocular disparity encoding cells generated through an Infomax based learning algorithm," *Neural Networks*, vol. 17, no. 7, pp. 953–962, 2004.

[5] 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.