Spatial intra-prediction based on mixtures of sparse representations

Angelique Dremeau, Mehmet Turkan, Cedric Herzet, Christine Guillemot, Jean-Jacques Fuchs
INRIA Centre Rennes - Bretagne Atlantique, Campus universitaire de Beaulieu, 35000 Rennes, France

***Abstract*—In this paper, we consider the problem of spatial prediction based on sparse representations. Several algorithms dealing with this problem can be found in the literature. We propose a novel method involving a mixture of sparse representations. We first place this approach into a probabilistic framework and then derive a practical procedure to solve it. Comparisons of the rate-distortion performance show the superiority of the proposed algorithm with regard to other state-of-the-art algorithms.**

***Index Terms*—Sparse representations, prediction, inpainting.**

I. Introduction

Intra-prediction is an important tool in image and video compression to deal with the spatial correlation of natural images. The idea of intra-prediction is to infer the value of some unknown image blocks from the knowledge of those already decoded. It is widely recognized that accurate prediction can significantly decrease the overall coding rate and this type of technique has thus been integrated to the latest video codec H.264 [1].

Recently, the prediction problem (and the closely-related "inpainting" problem) has been placed in the framework of sparse representations. Sparse representations aim at describing a signal (*e.g.*, an image block) as the combination of a small number of atoms chosen from an overcomplete dictionary. For a proper choice of the dictionary, it has been shown that such decompositions can offer very good performance in prediction or inpainting problems, see *e.g.*, [2], [3], [4], [5], [6], [7].

In [2] and [3], Guleryuz considers an overcomplete dictionary made up of orthonormal bases and proposes an iterative implementation of the sparse representation problem applied to inpainting. Another approach is presented by Elad *et al.* in [4]. The proposed implementation involves a different type of dictionary, made up of atoms capturing either "cartoon" or "texture" areas. Elad *et al.* add also a total variation (TV) penalty term to the standard sparse representation problem. Finally, in [5], Fadili *et al.* introduce an implementation based on the expectation-maximization (EM) algorithm.

Several contributions also consider the problem of prediction based on sparse representations in the context of image/video coding, see *e.g.,* [6], [7]. These contributions mainly distinguish by the choice of the dictionary used to "sparsely" represent the signal and the choice of the data used for the prediction. In [6], Martin *et al.* consider an overcomplete dictionary made up of discrete real Fourier and cosine functions, while Türkan *et al.* [7] construct a dictionary from image patches taken in a large causal area and consider seven possible causal neighborhoods.

The common features of the prediction methods mentioned above is the use of one *single* dictionary[1] in the sparse representation problem. In contrast, this paper considers the option of using a *mixture* of dictionaries: the vector is assumed to arise from a "multi-source" process where each source defines sparse signals over a *particular* dictionary. We will emphasize that prediction based on this model leads to estimates which are a weighted mixture of sparse representations in *each* of the considered dictionaries.

II. Sparse representation and standard prediction

Sparse representations aim at describing a signal as the combination of a small number of atoms chosen from an overcomplete dictionary. Formally, this problem can be formulated as follows. Let $\mathbf{D}\in\mathbb{R}^{N\times M}$ be a dictionary with $N\leq M$ and $\mathbf{y}\in\mathbb{R}^{N}$ an observed signal. We want to find the vector $\mathbf{x}\in\mathbb{R}^{M}$ such that:

$$\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 \quad \textit{subject to} \quad \|\mathbf{x}\|_0 \leq L, \tag{1}$$

where $\|\mathbf{x}\|_0$ denotes the l_0-norm, *i.e.,* the number of nonzero coefficients in $\mathbf{x}$ and L is a given constant. Note that problem (1) is also often expressed in its Lagrangian version:

$$\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_0, \tag{2}$$

where μ is a Lagrangian multiplier. Finding the exact solution of (2) is an intractable problem. Therefore, numerous suboptimal (but tractable) algorithms have been devised in the literature to address the SR problem. In particular, the Orthogonal Matching Pursuit (OMP) algorithm (see [8]) solves problem (1) and builds up the sparse vector $\mathbf{x}$ by successively adding a nonzero coefficient. The Global Matched Filter (GMF) algorithm also known as Basis Pursuit (BP) algorithm (see [9] and [10]) approximates the l_0-norm by the l_1-norm in (2) and can thus find an approximation of $\mathbf{x}$ by standard convex optimization procedures.

The idea of prediction based on sparse representations relies on the assumption that the missing data we want to predict and the observed data have a sparse representation in a given dictionary. Sparsity defines thus a prior on the signal made up of the concatenation of the observed and the missing data. This can be formalized as follows. Let $\mathbf{y}=[\mathbf{y}_o^T,\mathbf{y}_m^T]^T$ be the

[1] Although the dictionary can be chosen from a set of dictionaries or made of atoms of different nature (*e.g.* DCT, wavelets, curvelets, etc.).

Given σ_o^2, calculate:

1. $\forall i \in \{1, \dots, P\}$,

$$\mathbf{x}_i^\star = \underset{\mathbf{x}_i}{\arg\min} \left\{ \frac{1}{2\sigma_o^2} \|\mathbf{y}_o - \mathbf{D}_o^i \mathbf{x}_i\|_2^2 + \lambda_i \|\mathbf{x}_i\|_0 \right\}, \qquad (5)$$

2. $$\mathbf{y}_m^\star = \sum_{i=1}^{P} w_i^\star \mathbf{D}_m^i \mathbf{x}_i^\star, \qquad (6)$$

$$\begin{aligned} \textit{with} \quad w_i^\star &\triangleq p(c = i | \mathbf{y}_o, x^\star), \\ &\propto \exp\left(-\frac{1}{2\sigma_o^2} \|\mathbf{y}_o - \mathbf{D}_o^i \mathbf{x}_i^\star\|_2^2 - \lambda_i \|\mathbf{x}_i^\star\|_0\right) p(c = i). \qquad (7) \end{aligned}$$

Table I
MAIN EQUATIONS OF THE PROPOSED ALGORITHM

concatenation of $\mathbf{y}_o \in \mathbb{R}^{N_o}$, observed data, and $\mathbf{y}_m \in \mathbb{R}^{N_m}$, missing data. $\mathbf{y}$ is assumed to have a sparse representation in dictionary $\mathbf{D}$. The missing data $\mathbf{y}_m$ is thus estimated as:

$$\mathbf{y}_m^\star = \mathbf{D}_m \mathbf{x}^\star, \qquad (3)$$

where the sparse representation $\mathbf{x}^\star$ is calculated from the observed data as

$$\mathbf{x}^\star = \underset{\mathbf{x}}{\arg\min} \ \|\mathbf{y}_o - \mathbf{D}_o \mathbf{x}\|_2^2 + \mu \|\mathbf{x}\|_0, \qquad (4)$$

$\mathbf{D}_o \in \mathbb{R}^{N_o \times M}$ (resp. $\mathbf{D}_m \in \mathbb{R}^{N_m \times M}$) is the dictionary whose rows correspond to the elements in $\mathbf{y}_o$ (resp. $\mathbf{y}_m$).

III. INPAINTING USING A SET OF DICTIONARIES

In this section, we derive a prediction method in which the sought vector is assumed to have a sparse representation in *one* out of P dictionaries. We first expose a probabilistic framework suited to the modelization of such situations. We then propose a practical algorithm for prediction which exploits this framework.

A. A probabilistic framework

We consider a set of P dictionaries $\mathcal{D} = \{\mathbf{D}^i\}_{i=1}^P$ with $\mathbf{D}^i \in \mathbb{R}^{N \times M_i}$ $\forall i$. Let moreover x denote the set of representation vectors of $\mathbf{y}$ in each dictionary $\mathbf{D}^i$, *i.e.,*

$$x = \{\mathbf{x}_i\}_{i=1}^P. \qquad (8)$$

Based on these definitions, we consider the following model for $\mathbf{y}$:

$$p(\mathbf{y}) = \sum_{i=1}^{P} \int_{\mathbb{R}^{M_i}} p(\mathbf{y}|x, c = i)\, p(x|c = i)\, p(c = i)\, d\mathbf{x}_i, \qquad (9)$$

with

$$p(\mathbf{y}|x, c = i) = \mathcal{N}(\mathbf{D}^i \mathbf{x}_i, \Sigma), \qquad (10)$$

$$p(x|c = i) \propto \exp\{-\lambda_i \| \mathbf{x}_i \|_0\}, \qquad (11)$$

where $\lambda_i > 0$ and $\propto$ denotes equality up to a normalization factor [2]. $\mathcal{N}(\mu, \Gamma)$ denotes a Gaussian distribution with mean μ and covariance Γ. Hereafter, we consider that Σ is a diagonal matrix[3] with:

$$\Sigma_{jj} = \begin{cases} \sigma_o^2 & \textit{if element } (j) \textit{ is in } \mathbf{y}_o, \\ \sigma_m^2 & \textit{if element } (j) \textit{ is in } \mathbf{y}_m, \end{cases} \qquad (12)$$

The model (9)-(11) can be interpreted as follows: $\mathbf{y}$ is assumed to be a noisy combination of vectors from *one* (among P) dictionary; the choice of the dictionary is indexed by c. Sparsity is encouraged via prior (11) which penalizes $\mathbf{x}_i$'s with many nonzero elements. $p(\mathbf{y})$ can therefore be understood as a mixture of Gaussians $\mathcal{N}(\mathbf{D}^i \mathbf{x}_i, \Sigma)$ where each element is weighted by a factor depending on the sparsity of $\mathbf{x}_i$ and the prior probability $p(c = i)$.

B. MMSE Prediction from a Mixture of Dictionaries

We now propose a practical method to infer the value of $\mathbf{y}_m$ from the observation of $\mathbf{y}_o$. We look for the solution of the following minimum mean-square estimation (MMSE) problem:

$$\mathbf{y}_m^\star = \int_{\mathbf{y}_m} \mathbf{y}_m \, p(\mathbf{y}_m | x = x^\star, \mathbf{y}_o)\, d\mathbf{y}_m. \qquad (13)$$

where

$$x^\star = \underset{x}{\arg\max} \log p(\mathbf{y}_o, x). \qquad (14)$$

Given the knowledge of x, (13) is therefore the optimal estimator in terms of mean-square reconstruction error.

The goal function in (14) can be rewritten as

$$p(\mathbf{y}_o, x) = \sum_{c} \int_{\mathbf{y}_m} p(\mathbf{y}, x, c)\, d\mathbf{y}_m. \qquad (15)$$

Taking model (9)-(11) into account, the ith term of the summation over c only depends on $\mathbf{x}_i$. Therefore, the joint optimization problem (14) over x reduces to P individual optimization problems over $\mathbf{x}_i$. The solution of (14) is expressed as $x^\star = \{\mathbf{x}_i^\star\}_{i=1}^P$ with:

$$\mathbf{x}_i^\star = \underset{\mathbf{x}_i}{\arg\min} \left\{ \frac{1}{2\sigma_o^2} \|\mathbf{y}_o - \mathbf{D}_o^i \mathbf{x}_i\|_2^2 + \lambda_i \|\mathbf{x}_i\|_0 \right\} \ \forall i, \qquad (16)$$

where $\mathbf{D}_o^i \in \mathbb{R}^{N_o \times M_i}$ denotes the restriction of $\mathbf{D}^i$ to rows corresponding to elements in $\mathbf{y}_o$.

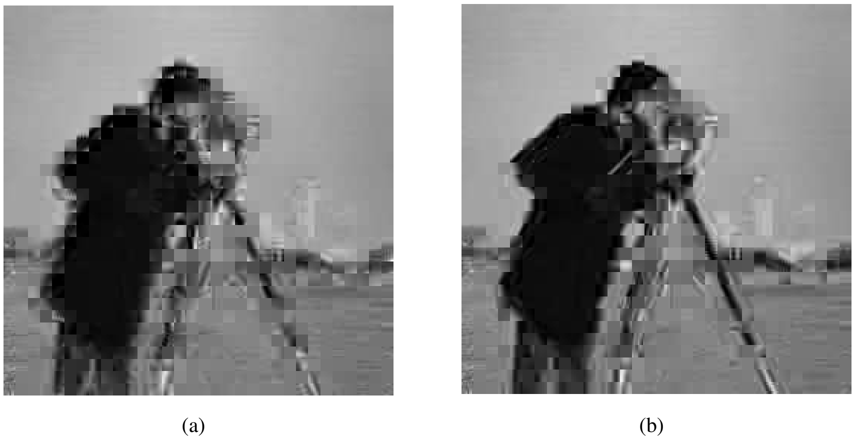
It is interesting to note that (16) has the form of a standard sparse representation problem (2). More particularly, $\mathbf{x}_i^\star$ can be regarded as the sparse representation of $\mathbf{y}_o$ in dictionary $\mathbf{D}_o^i$.

On the other hand, taking model (9)-(11) into account, it is easy to see that $p(\mathbf{y}_m | x = x^\star, \mathbf{y}_o)$ is a mixture of Gaussians:

$$p(\mathbf{y}_m | x = x^\star, \mathbf{y}_o) = \sum_{i=1}^{P} p(c = i | \mathbf{y}_o, x^\star)\, \mathcal{N}(\mathbf{D}_m^i \mathbf{x}_i^\star, \sigma_m^2 \mathbf{I}),$$

[2] Note that (11) is actually improper since the normalization factor is equal to ∞. This technical problem does however not lead to any particular issue in the rest of the paper.

[3] For a sake of simplicity, we consider in this paper a unique Σ for all dictionaries, but the general case where the noise variance depends on the dictionary is straightforward.



(a) (b)

Figure 1. Spatial prediction result for "Cameraman" with the standard method based on sparse representations 1(a) and the proposed method 1(b)

where $\mathbf{D}_m^i \in \mathbb{R}^{N_m \times M_i}$ denotes the restriction of $\mathbf{D}^i$ to rows which correspond to elements in $\mathbf{y}_m$. Therefore, (13) writes

$$\mathbf{y}_m^\star = \sum_{i=1}^{P} p(c = i|\mathbf{y}_o, x^\star)\ \mathbf{D}_m^i \mathbf{x}_i^\star. \tag{17}$$

According to equations (3)-(4), $\mathbf{D}_m^i \mathbf{x}_i^\star$ is the estimate of $\mathbf{y}_m$ if one single dictionary $\mathbf{D}^i$ is considered. Hence, $\mathbf{y}_m^\star$ can be interpreted as a weighted combination of estimates in different dictionaries. The weighting coefficients, $p(c = i|\mathbf{y}_o, x^\star)$, give the probability that the observed vector $\mathbf{y}_o$ has been generated as a sparse combination of atoms from the ith dictionary. These a posteriori probabilities can be computed as:

$$p(c|\mathbf{y}_o, x^\star) \propto \exp(-\frac{1}{2\sigma_o^2}\|\mathbf{y}_o - \mathbf{D}_o^c \mathbf{x}_c^\star\|_2^2 - \lambda_c \|\mathbf{x}_c^\star\|_0)\, p(c). \tag{18}$$

The implementation of (13)-(14) is summarized in Table I.

The complexity of the proposed algorithm is dominated by the P operations (16). This complexity is the same as that of solving the standard sparse representation problem (2) with a dictionary made up of the concatenation of the P considered dictionaries.

Note that if we make the assumption $P = 1$, then the equivalence between (3)-(4) and (13)-(14) is straightforward by taking model (9)-(11) into account. We thus recognize the standard formulation of the inpainting problem based on sparse representations as a particular case of the method proposed in this paper.

IV. Implementation and results

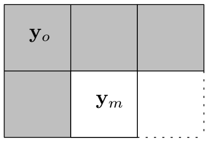


Figure 2. Illustration of an image block prediction: block to predict $\mathbf{y}_m$ and the causal neighborhood considered, $\mathbf{y}_o$.

In this section, we apply the proposed algorithm to the problem of image intra-prediction. We consider the spatial prediction context illustrated in Fig. 2: prediction is performed on each 8×8 pixel block (white block in Fig. 2) from the 4 nearest causal 8×8 pixel blocks (grey blocks in Fig. 2). We compare the performance of the proposed approach with two other prediction algorithms: a H.264-like predictive scheme and the standard prediction based on sparse representations (3)-(4).

In the rest of this section, we first detail the choice of model parameters and the encoding scheme in section IV-A and IV-B, respectively. Performance of the predictive schemes is then illustrated in section IV-C.

A. Model parameters

The parameters characterizing model (9)-(11) are defined as follows. We assume that the distribution of c is uniform:

$$\forall i \in \{1, \ldots, P\},\ \ p(c = i) = \frac{1}{P}. \tag{19}$$

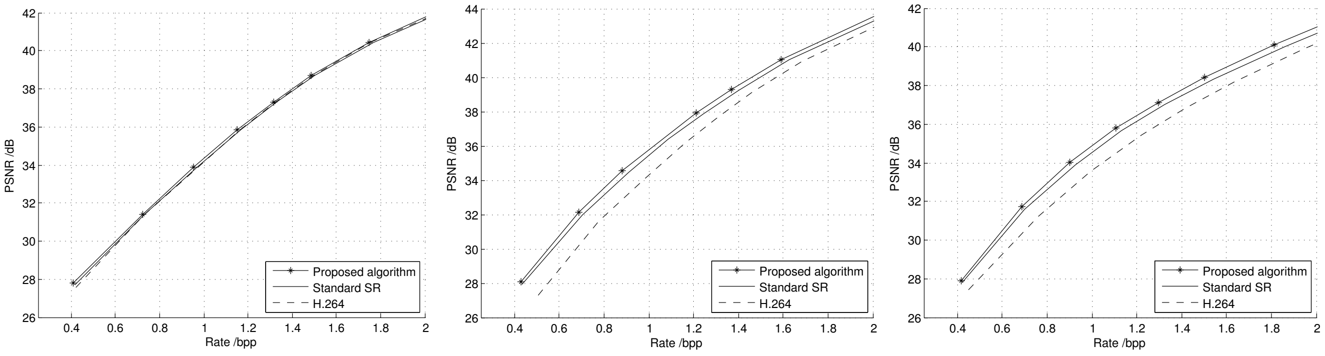
The computation of the a posteriori probabilities (18) reduces thus to the following expression:

$$p(c|\mathbf{y}_o, x^\star) \propto \exp(-\frac{1}{2\sigma_o^2}\|\mathbf{y}_o - \mathbf{D}_o^c \mathbf{x}_c^\star\|_2^2 - \lambda_c \|\mathbf{x}_c^\star\|_0). \tag{20}$$

The dictionaries used to "sparsely" represent the data are the directional DCTs (DDCT) introduced by Zeng and Fu in [11] and later extended in [12] by Dremeau *et al.*. We generate 7 directional DCTs corresponding to the prediction modes in H.264 (DC, vertical and horizontal modes are included in the classical DCT). Note that the directional DCTs are orthonormal bases.

We set $\lambda_i = \log N\ \forall i$, assuming that the result established by Donoho and Johnstone in [13] is still valid when skipping some rows of an orthogonal dictionary.

The choice of the value of σ_o^2 is closely related to the sparsity on x. On the other hand, this latter strongly impacts the reconstruction quality on $\mathbf{y}_m$. Now, the "best" sparsity on x with regard to the reconstruction of $\mathbf{y}_m$ can be extremely



(a)

(b)

(c)

Figure 3. PSNR versus Rate for "Cameraman" 3(a), "Roofs" 3(b) and "Barbara" 3(c)

varying from a predicted block to another. An additional information on the "best" sparsity level has then to be transmitted to the decoder. In the standard methods based on sparse representations, the sent information corresponds to the iteration number used in pursuit algorithms ([6], [7]). In our case, sending the "best" value of σ_o^2 is very costly since this variable is continuously-valued. We thus define the noise variance σ_o^2 as follows: for a given number of nonzero coefficients L,

$$\forall i \in \{1, \ldots, P\}, \tag{21}$$

$$\mathbf{x}_i^\star = \underset{\mathbf{x}_i}{\arg\min} \ \|\mathbf{y}_o - \mathbf{D}_o^i \mathbf{x}_i\|_2^2 \quad \textit{subject to} \quad \|\mathbf{x}_i\|_0 \leq L,$$

$$\sigma_o^2 = \frac{1}{N_o} \sum_{i=1}^{P} \frac{1}{P} \|\mathbf{y}_o - \mathbf{D}_o^i \mathbf{x}_i^\star\|_2^2, \tag{22}$$

where N_o is the number of pixels in $\mathbf{y}_o$.

Hence, the definition of σ_o^2 reduces to the knowledge of L, which fixes the iteration number of pursuit algorithms used to solve (21). In order to maximize the reconstruction quality on $\mathbf{y}_m$, the iteration number is then optimized under a distortion criterion on the block to predict (Mean-Square Error between original and predicted block). We will precise in the next section the coding aspect of this specification.

B. Prediction and encoding scheme

To initialize the prediction, the first top row and first left column of 8×8 pixel blocks are encoded with JPEG algorithm. For the standard prediction based on sparse representations we use the OMP algorithm whose iteration number varies between 1 and 8. The iteration number is optimized under a distortion criterion on the block to predict (Mean-Square Error between original and predicted block) and is then Huffman encoded. A similar process is used for the definition of the noise variance σ_o^2 in the proposed algorithm, as we discussed in previous section.

The residual between the original block and its prediction is encoded with an algorithm similar to JPEG. A uniform quantization matrix is used with step size equal to 16. It is weighted by a quality factor increasing from 10 to 90 with a step size equal to 10.

C. Performance analysis

We evaluate and compare three different prediction algorithms:

- "H.264" implements a spatial prediction similar the one used in H.264 on 4×4 pixel blocks but extended to 8×8 pixel blocks (to be fair with the sparse-representation-based algorithms, the prediction modes are also chosen according to a distortion criterion on the block to predict),
- "Standard SR" uses the standard prediction based on sparse representations (3)-(4) with a dictionary made up of the concatenation of the 7 directional DCTs, *i.e.,* $\mathbf{D} = [\mathbf{D}^1, \ldots, \mathbf{D}^i, \ldots, \mathbf{D}^P]$,
- "Proposed algorithm" implements the prediction algorithm defined in Table I.

Fig. 1 illustrates the prediction improvement brought by the proposed algorithm compared with the standard prediction method based on sparse representations. The gain is thus visually perceptible on the example of "Cameraman": we clearly notice that the geometric structures (see *e.g.,* the arm or the camera stand) are particularly well recovered by the proposed algorithm.

Fig. 3 represents the Rate-PSNR performance achieved by the three algorithms "H.264", "Standard SR" and "Proposed algorithm" for images "Barbara", "Roofs" and "Cameraman". As far as these three images are concerned, we can observe that the proposed prediction algorithm outperforms the standard approach based on sparse representations and the H.264-like prediction. Thus the proposed algorithm leads to a gain up to 2 dB (for "Roofs") with regard to the H.264-like prediction and a less important but still significative gain of 0.5 dB (for "Roofs" and "Barbara") with regard to the standard method based on sparse representations.

Note that all techniques aiming at adapting the support of the prediction (causal area, in grey in Fig. 2) developed within the framework of prediction methods based on sparse representations (see *e.g.,* [7]) can also be applied to the proposed algorithm. This can possibly lead to an additional performance improvement.

Moreover, the chosen directional DCTs allow here a H.264-like apprehension of the prediction problem, but other dictionaries can also be considered. A set of dictionaries well-adapted to texture content like Gabor transforms or wavelet packets ([14]) on the one hand and to cartoon content like curvelets ([15]) or bandelets ([16]) on the other hand could lead to a better representation of the local image characteristics and thus possibly to an improvement of the prediction performance.

V. Conclusion

In this paper, we address the prediction problem based on sparse representations using a set of dictionaries. This problem is placed in a probabilistic framework by considering the data as realizations of a mixture of Gaussians. The prediction task is then reformulated as a MMSE estimation problem and a procedure is derived to solve it. The proposed algorithm is shown to give enhanced performance with regard to previously-proposed algorithms.

References

[1] ITU-T Rec. H.264 ISO/IEC 14496-10 (AVC), *Advanced Video Coding for Generic Audiovisual Services*, March 2005.

[2] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part 1 - theory," *IEEE Trans. On Image Processing*, vol. 15, pp. 555–571, 2004.

[3] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part 2 - adaptive algorithms," *IEEE Trans. On Image Processing*, vol. 15, pp. 555–571, 2004.

[4] M. Elad, J-L Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *Journal on Applied and Computational Harmonic Analysis (ACHA)*, vol. 19, pp. 340–358, November 2005.

[5] M.J. Fadili and J-L. Starck, "An em algorithm for sparse representation-based image inpainting," in *Proc. IEEE Int'l Conference on Image Processing (ICIP).*, September 2005, vol. 2, pp. II – 61–4.

[6] A. Martin, J-J. Fuchs, C. Guillemot, and D. Thoreau, "Sparse representations for image prediction," in *Proc. European Signal Processing Conference (EUSIPCO), Poznan, Poland.*, September 2007.

[7] M. Turkan and C. Guillemot, "Sparse approximation with adaptive dictionary for image prediction," in *Proc. IEEE Int'l Conference on Image Processing (ICIP)*, November 2009, pp. 25–28.

[8] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 1993, pp. 40–44.

[9] J.J. Fuchs, "On the application of the global matched filter to doa estimation with uniform circular arrays," *IEEE Trans. On Signal Processing*, vol. 49, no. 4, pp. 702–709, 2001.

[10] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp*, vol. 20, no. 1, pp. 3361, 1999.

[11] B. Zeng and J. Fu, "Directional discrete cosine transforms - a new framework for image coding," *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 18, no. 3, pp. 305–313, March 2008.

[12] A. Dremeau, C. Herzet, C. Guillemot, and J.J Fuchs, "Sparse optimization with directional dct bases for image compression," in *Proc. IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 1290–1293.

[13] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1993.

[14] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. On Image Processing*, vol. 2, no. 2, pp. 160–175, April 1993.

[15] E. J. Candes and D. L. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," Tech. Rep., Stanford University CA - Dept of Statistics, 2000.

[16] E. LePennec and S. Mallat, "Sparse geometric image representations with bandelets," *IEEE Trans. On Image Processing*, vol. 14, no. 4, pp. 423–438, April 2005.