DEPARTMENT OF ELECTRICAL ENGINEERING
SIGNAL PROCESSING LABORATORY
Professor M. Kunt - Director

EPFL–DE–LTS CH–1015 LAUSANNE SWITZERLAND

Telephone: +41-21-693-2626 Secretariat: +41-21-693-2601
Telefax: +41-21-693-7600
Email: kunt@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Redundancy-Driven A Posteriori Matching Pursuit Quantization

Pascal Frossard, Pierre Vandergheynst and Murat Kunt

Internal Report LTS 00.01

Lausanne, October, 2000

Redundancy-Driven A Posteriori Matching Pursuit Quantization [1]

Pascal Frossard, Pierre Vandergheynst and Murat Kunt

Signal Processing Laboratory
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

pascal.frossard@epfl.ch

CONTENTS

[1] This work has been partly supported by Hewlett-Packard.

Abstract

This paper studies quantization error in the context of Matching Pursuit coded streams. The quantization noise is shown to depend on error on both coefficients and indexes. It is moreover influenced by the redundancy of the Matching Pursuit dictionary. A novel general formulation of the structural redundancy in overcomplete decompositions is shown to enhance the accuracy of classical redundancy factors. The redundancy factor drives the decay of Matching Pursuit coefficients energy and is therefore used to design an optimal *a posteriori* quantization scheme for multi-resolution Matching Pursuit coding. This exponentially upper-bounded quantization of Matching Pursuit coefficients clearly outperforms uniform quantization schemes in the practical case of image coding. The atom selection error is then studied for structured dictionaries and more particularly dictionaries built upon common Gabor functions. This analysis is used to quantify the error in the case of a simple scalar quantization of structured atom indexes. Finally, it is shown that *a posteriori* scalar indexes quantization should be preferably avoided, thus emphasizing the need for a careful dictionary design as a trade-off between compression efficiency and code size.

Keywords

Matching Pursuit, Structured Dictionaries, Quantization, Atom Selection Error, Redundancy, Gabor Decomposition.

I. INTRODUCTION

Non-orthogonal transforms presents several interesting properties which position them as an interesting alternative to orthogonal transforms like DCT or wavelet based schemes. Decomposing a signal over a redundant dictionary improves the compression efficiency, especially at low bit rate where most of the signal energy is captured by only few elements. Moreover, signals resulting from decomposition over redundant dictionary are more robust to noise [1]. The main limitation of non-orthogonal transforms is however the encoding complexity, since the number of possible decompositions becomes infinite. Matching Pursuit algorithms [2] provide an interesting way to iteratively decompose the signal in its most important features with a limited complexity. It outputs a stream composed of both atoms or basis functions and their respective coefficients.

The aim of this paper is to study the effects of quantization onto reconstruction of Matching Pursuit streams. Since the Matching Pursuit coefficients generally take on real values, quantization is necessary to reduce the bandwidth needed to transmit them. Quantization errors have been studied in [3–6] in the context of overcomplete frame expansions and consistent Matching Pursuit. Similarly, in the case of real or even very large dictionaries, atoms indexes should also be quantized to limit the bit rate. This paper studies the quantization error of both the coefficients and atoms for a general Matching Pursuit decomposition. This study focuses particularly on the transmission of Matching Pursuit streams over heterogeneous networks. In this case, Matching Pursuit streams is computed once and then quantized several times to satisfy different rate constraints.

The set of functions that forms the dictionary plays a crucial role in Matching Pursuit coding. A very redundant dictionary generally allow to capture the main features of the signal with only a few dictionary functions. However, the coding rate of the function parameters obviously increases with the size of the dictionary. Meanwhile, the relative importance of each component of the decomposition also directly depends on the dictionary. The chance to find a function that closely fits the input signal, and thus capture most of its energy, grows with the dictionary size. A new formulation of the structural

redundancy of dictionaries is therefore proposed to quantify the compression properties of a dictionary. The redundancy factor leads the energy decay rate of the residual signal coded by Matching Pursuit, which has been proven to be upper-bounded by an exponential curve [7–9]. In [10, 11] a redundancy formulation has been proposed in the context of frame expansion, but the same computation can not be applied to Matching Pursuit decomposition. The formulation proposed here is however general enough to be applied to any structured dictionary, and even in the context of frame expansions.

The quantization error is then studied for *a posteriori* quantization schemes, in contrast with usual schemes [12–14] where the encoder uses the quantized coefficient to update the residual signal. In *a posteriori* quantization, the quantized version of the elements does not influence the decomposition. Hence the Matching Pursuit stream can be computed only once and then quantized several times to serve different constraints. The signal reconstruction error due to quantization is shown to depend on the error on both the coefficients and the atoms. Moreover, the contribution of each Matching Pursuit elements to the quantization error clearly depends on its position within the encoded stream, and hence on the redundancy of the dictionary. Based on the characterization of the energy decay curve, an exponentially bound quantization scheme is proposed for the Matching Pursuit coefficients. This scheme truly outperforms uniform quantization schemes, especially at low bit rates. The scheme is shown to achieve very good results in the practical case of image compression. The index quantization error in structured dictionaries, and more particularly 2D Gabor dictionaries, will then be studied. It will finally be shown that index quantization can not be performed in an optimal way. This conclusion still emphasizes the importance of a correct design of the dictionary to avoid the need of index quantization.

The paper is organized as follows: Section II first overviews Matching Pursuit algorithm and describes the construction of structured dictionaries. Section III then proposes a general formulation of the redundancy of Matching Pursuit dictionaries. An approximation is provided for uniformly distributed atoms. An upper-bound on the signal reconstruction error is given in Section IV for quantization error. Section V studies the *a posteriori* quantization of Matching Pursuit coefficients and proposes an exponential quantization which takes part of the properties of the encoding. Section VI studies a structured dictionary index quantization scheme driven by the redundancy of the dictionary. It therefore focuses on the error due to wrong atom selection at reconstruction. Finally, concluding remarks are given in Section VII.

II. Matching Pursuit and Structured Dictionaries

A. Matching Pursuit Overview

In contrast to orthogonal transforms, overcomplete expansions of signals are not unique. The number of feasible decompositions is infinite, and finding the best solution under a given criteria is a NP-complete problem. In compression, one is interested in representing the signal to be coded with the smallest number of elements, that is in finding the solution with most of the energy on only a few functions. Matching Pursuit is one of the sub-optimal approaches that greedily approximates the solution to this NP-complete problem.

Matching Pursuit (MP) is an adaptive algorithm that iteratively decomposes any function f in the

Hilbert space $\mathcal{H}$ in a possibly redundant dictionary of functions called *atoms* [2]. Let $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ be such a dictionary with $\|g_\gamma\| = 1$ and Γ represents the set of possible indexes. The function f is first decomposed as follows :

$$f = \langle g_{\gamma_0}|f\rangle\, g_{\gamma_0} + \mathcal{R}f\,, \tag{1}$$

where $\langle g_{\gamma_0}|f\rangle\, g_{\gamma_0}$ represents the projection of f onto g_{γ_0} and $\mathcal{R}f$ is a residual component. Since all elements in $\mathcal{D}$ have by definition a unit norm, it is easy to see from eq. (1) that g_{γ_0} is orthogonal to $\mathcal{R}f$, and this leads to

$$\|f\|^2 = |\langle g_{\gamma_0}|f\rangle|^2 + \|\mathcal{R}f\|^2\ . \tag{2}$$

To minimize $\|\mathcal{R}f\|$, one must choose g_{γ_0} such that the projection coefficient $|\langle g_{\gamma_0}|f\rangle|$ is maximum. The pursuit is carried out by applying iteratively the same strategy to the residual component. After N iterations, one has the following decomposition for f :

$$f = \sum_{n=0}^{N-1} \langle g_{\gamma_n}|\mathcal{R}^n f\rangle g_{\gamma_n} + \mathcal{R}^N f\,, \tag{3}$$

where $\mathcal{R}^N$ is the residual of the N^{th} step with $\mathcal{R}^0 f = f$. Similarly, the energy $\|f\|^2$ is decomposed into :

$$\|f\|^2 = \sum_{n=0}^{N-1} |\langle g_{\gamma_n}|\mathcal{R}^n f\rangle|^2 + \|\mathcal{R}^N f\|^2\ . \tag{4}$$

It can be shown now that the Matching Pursuit (MP) scheme recovers the projection of f onto the space defined by the dictionary atoms $\{g_\gamma\}_{\gamma\in\Gamma}$. Let $\mathbf{V}$ be the closed linear span of $\mathcal{D}$. The following theorem, proved in [2], shows that MP recovers the projection of f onto $\mathbf{V}$:

Theorem 1: The residue $\mathcal{R}f$ defined by eq. (3) satisfies

$$\lim_{n\to+\infty} \|\mathcal{R}^n f - \mathbf{P}_\mathbf{W} f\| = 0\,, \tag{5}$$

where $\mathbf{W}$ is the orthogonal complement of $\mathbf{V}$ and $\mathbf{P}_\mathbf{W}$ is the associated projection operator.
This also means that we can write

$$\mathbf{P}_\mathbf{V} f = \sum_{n=0}^{\infty} \langle g_{\gamma_n}|\mathcal{R}^n f\rangle\, g_{\gamma_n} \tag{6}$$

and

$$\|\mathbf{P}_\mathbf{V} f\|^2 = \sum_{n=0}^{\infty} |\langle g_{\gamma_n}|\mathcal{R}^n f\rangle|^2\ . \tag{7}$$

This is an important issue. Indeed, when the dictionary is complete (i.e., $\mathbf{V} = L^2(\mathbb{R})$), f can be exactly decomposed: we have $\mathbf{P}_\mathbf{V} f = f$, and thus $\mathbf{P}_\mathbf{W} f = 0$. In this case, the original function can be expressed

as :

$$f = \sum_{n=0}^{+\infty} \langle g_{\gamma_n} | \mathcal{R}^n f \rangle \, g_{\gamma_n} \tag{8}$$

Designing complete dictionaries is thus an important issue that we address in the next section using an original geometrical technique. Although Matching Pursuit places very few restrictions on the dictionary, the latter is strongly related to convergence speed and thus to coding efficiency. In this paper, the convergence speed has to be understood as the ability for the MP to capture most of the input signal energy in just a few iterations. Moreover, the decoder needs the dictionary to reconstruct the signal. Dictionaries built on elementary functions are thus preferred, since the encoder transmits only the parameters or indexes of these functions instead of the complete functions. Finally, any collection of arbitrarily sized and shaped functions can be used as dictionary, as long as completeness is respected. The remaining of this section presents the design of structured dictionaries which will be used within this paper.

B. Group theoretical design of dictionaries

Let G be a locally compact group and $\mathcal{U}$ a unitary representation of G in a given Hilbert space $\mathcal{H}$. Definitions and properties of group representations can be found in Appendix A. Let $\mathcal{U}(\gamma)f$ denote the action of the unitary operator $\mathcal{U}(\gamma)$ on the function f. The unitary representation is said to be *i*rreducible if it does not admit any non trivial invariant subspaces. That is if $\mathcal{S} \subset \mathcal{H}$ is such that

$$\mathcal{U}(\gamma)f \in \mathcal{S}\,, \quad \forall f \in \mathcal{S} \text{ and } \forall \gamma \in G$$

then either we have $\mathcal{S} = \mathcal{H}$ or $\mathcal{S} = \{0\}$. In particular, this means that the orbit of any $f \in \mathcal{H}$ under G is a dense subspace of $\mathcal{H}$:

$$\mathcal{O}_f = \{\mathcal{U}(\gamma)f \,|\, \forall \gamma \in G\}\,, \quad \overline{\mathcal{O}_f} = \mathcal{H}\,. \tag{9}$$

This shows that we can use unitary irreducible group representations (UIR) to design complete dictionaries of a Hilbert space. Such constructions are well-known to mathematical physicists as they are the first step in designing *coherent state* systems [15]. Such a construction induces covariance of the orbit, and thus of the dictionary, with respect to group action. Using the notation $f_\gamma = \mathcal{U}(\gamma)f$, this property is summarized by the following relation :

$$\mathcal{U}(\gamma_0)f_\gamma = \mathcal{U}(\gamma_0 \circ \gamma)f \tag{10}$$

where $\circ$ denotes the group law. The projection of a target signal f onto such a dictionary using the unitarity of $\mathcal{U}$ yields to the following equation :

$$\begin{aligned} \langle g | f_\gamma \rangle &= \langle g | \mathcal{U}(\gamma)f \rangle \\ &= \langle \mathcal{U}(\gamma^{-1})g | f \rangle\,. \end{aligned}$$

Hence, the projection on the dictionary is covariant [2] with respect to G. This also means that f_γ can be reconstructed from the coefficients of f by applying the group law to the dictionary elements selected by Matching Pursuit :

$$f_\gamma = \sum_{n=0}^{+\infty} \langle g_{\gamma_n} | \mathcal{R}^n f \rangle g_{\gamma \circ \gamma_n} \, . \tag{11}$$

If we truncate this expansion, the norm of the residual remains unchanged since the representation is unitary.

Example 1 (2-D wavelet dictionary) Let us consider the similitude group of $\mathbb{R}^2$,

$$\mathrm{SIM}(2) = \mathbb{R}^2 \rtimes \left(\mathbb{R}_*^+ \times \mathrm{SO}(2) \right) .$$

This group is composed of translations, dilations and rotations in the plane. It can be checked [15] that the following representation is a UIR of $\mathrm{SIM}(2)$:

$$\mathcal{U}(\vec{b}, a, \theta) g(\vec{x}) = \frac{1}{a} g\left(a^{-1} r_{-\theta} (\vec{x} - \vec{b}) \right) . \tag{12}$$

This representation is also the basic ingredient in the construction of the 2-D Continuous Wavelet Transform [16].

C. Semi-Structured Dictionaries

In some cases the operations we would like to apply to a given atom do not form a group. Nevertheless there are interesting cases where a subset of these operations do actually form a genuine group and some properties of the previous section still hold, though in a weaker form. Suppose our dictionary is spanned by a family of unitary operators U_α :

$$\mathcal{D} = \{ U_\alpha g \, , \ \alpha \in A \} \, , \tag{13}$$

for a given set of indices A. Then suppose we have a unique decomposition :

$$U_\alpha = \mathcal{U}(\gamma_\alpha) \bullet R_\alpha \, , \tag{14}$$

where $\bullet$ denote the composition of operators, $\mathcal{U}$ is a unitary representation of a given group G and R_α is a unitary operator belonging to $U\left(\mathcal{H}\right)$. The dictionary $\mathcal{D}$ is then invariant under G and MP coefficients are covariant under the action of G onto $\mathcal{H}$. We call such a dictionary *semi-structured*.

Example 2 (Anisotropic Refinement) Let the index set A contain translations, rotations and two dilations, one for each principal direction :

$$A = \left\{ \vec{b} \in \mathbb{R}^2 , \theta \in \mathrm{SO}(2), (a_1, a_2) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+ \right\} . \tag{15}$$

The action of U_α on an atom g is defined by :

$$U_\alpha g = \mathcal{U}(\vec{b}, \theta) D(a_1, a_2) g \, , \tag{16}$$

[2] Here covariant means that the projection coefficients are transformed according to the group law.

where $\mathcal{U}$ is a representation of the Euclidean group :

$$\mathcal{U}(\vec{b},\theta)g(\vec{x}) = g\big(r_{-\theta}(\vec{x}-\vec{b})\big)\,, \tag{17}$$

and D acts as a dilation operator :

$$D(a_1,a_2)g(\vec{x}) = \frac{1}{\sqrt{a_1 a_2}} g\big(\frac{x_1}{a_1},\frac{x_2}{a_2}\big)\,, \tag{18}$$

where $\vec{x} = (x_1, x_2)$. When $a_1 = a_2$ one obviously gets back to the similitude group of the plane and, avoiding rotations, we end up with a group studied by Bernier and Taylor [17]. In general the set A can not be endowed with a group structure, but the factorisation property of eq. (16) guarantees covariance with respect to the Euclidean group. This dictionary offers a simple way to implement an anisotropic refinement process on atoms. This allows to efficiently capture the geometry of edges (and other elongated structures) in images.

III. Redundancy and Convergence of Matching Pursuit

A. *Convergence of Matching Pursuit*

The convergence speed of Matching Pursuit corresponds to its ability to extract the maximum signal energy in a few iterations. In other words, it corresponds to the decay rate of the residue and thus the coding efficiency of Matching Pursuit. The convergence speed depends directly on the dictionary set. The decay rate of the residual energy can thus be bounded once the dictionary is known, even without a priori information about the input signal.

The approximation error decay rate in Matching Pursuit have been shown to be bounded by an exponential [2, 9]. In other words, the decay of the residue norm is faster than an exponential decay curve whose rate depends on the dictionary only. From [7], there exists $\lambda > 0$ such that for all $N \geq 0$

$$\|\mathcal{R}^N f\| \leq 2^{-\lambda N}\|f\|\,, \tag{19}$$

or, equivalently,

$$\|\mathcal{R}^{m+1} f\| \leq 2^{-\lambda}\,\|\mathcal{R}^m f\|\,, \forall m\,. \tag{20}$$

The decay rate[3] can be written as [7]

$$2^{-\lambda} = (1-\alpha^2\,\beta^2)^{\frac{1}{2}}, \tag{21}$$

where $\alpha \in (0,1]$ is an optimality factor. This factor depends on the algorithm that, at each iteration, searches for the best atom in the dictionary. The optimality factor α is set to one when MP browses the complete dictionary at each iteration. The parameter β depends on the dictionary construction. It represents the ability of the dictionary functions to capture features of any input function f. The upper-bound on the coefficient norm is reached in the worst case where the input function is the farthest from

[3]The decay parameter λ decreases when the size of the signal space increases. However, at the limit of infinite dimensional spaces, the convergence is not exponential any more [18].

any dictionary vector. Hence, β is defined by :

$$\sup_{\gamma} |\langle f, g_{\gamma_n} \rangle| \geq \beta \|f\|. \tag{22}$$

This relationship is further developed in the next sections.

B. New formulation of the redundancy factor

The parameter β can be interpreted as the cosine of the maximum angle between any direction in $\mathcal{H}$ and its closest direction of any L-dimensional vector of the dictionary [2]. It characterizes the redundancy of the dictionary and tends to one when the size of the complete dictionary increases. In this section, a new formulation of β is provided. It allows to compute the redundancy factor of any over-complete dictionary in $\mathcal{H}$.

For each dictionary vector g_{γ_i}, one can define its projection neighborhood as the subspace of $\mathcal{H}$ which admits g_{γ_i} as closest direction.

Definition 1: The projection neighborhood $\mathcal{V}_{\gamma_i}$ of the vector g_{γ_i} is the subspace of $\mathcal{H}$ defined by

$$\mathcal{V}_{\gamma_i} = \{x \in \mathcal{H} \mid |\langle x|g_{\gamma_i}\rangle| \geq |\langle x|g_{\gamma_j}\rangle|, \forall j \neq i\} \, . \tag{23}$$

The projection neighborhood $\mathcal{V}_{\gamma_i}$, as represented in Fig. 1 and 2, corresponds to the intersection of couples of infinite convex polyhedral cones situated symmetrically with their apexes at the origin[4]. The polyhedral cones are specified by a pair of half-space constraints for each element of the dictionary. The half-space constraints can be expressed as

$$\langle x \mid g_{\gamma_i} + g_{\gamma_j} \rangle \, \langle x \mid g_{\gamma_i} - g_{\gamma_j} \rangle > 0, \; \forall j \neq i \, . \tag{24}$$

The parameter β corresponds thus to the cosine of the maximum possible angle, over all dictionary vectors, between g_{γ_i} and any direction in its projection neighborhood $\mathcal{V}_{\gamma_i}$. It can be written as

$$\beta = \min_{i} \inf_{x \in \mathcal{V}_{\gamma_i}} \langle x \mid g_{\gamma_i} \rangle \, . \tag{25}$$

The direction which has the maximum angle with g_{γ_i} is the intersection of two (L-1)-dimensional hyperplanes specified by the half-space constraints of eq. (24). Instead of browsing the complete $\mathcal{V}_{\gamma_i}$ region, one can look for such an intersection with the maximum angle to g_{γ_i}, but within $\mathcal{V}_{\gamma_i}$.

The procedure can be formulated as follows. First pick a vector g_{γ_j} different from g_{γ_i}. Let $g_{\gamma_{i+j}}$ and $g_{\gamma_{i-j}}$ respectively represent the normalized sum and difference of g_{γ_i} and g_{γ_j}:

$$g_{\gamma_{i+j}} = \frac{g_{\gamma_i} + g_{\gamma_j}}{\|g_{\gamma_i} + g_{\gamma_j}\|} \tag{26}$$

and

$$g_{\gamma_{i-j}} = \frac{g_{\gamma_i} - g_{\gamma_j}}{\|g_{\gamma_i} - g_{\gamma_j}\|} \tag{27}$$

[4] A similar geometrical representation is proposed in [4] for consistent reconstruction.

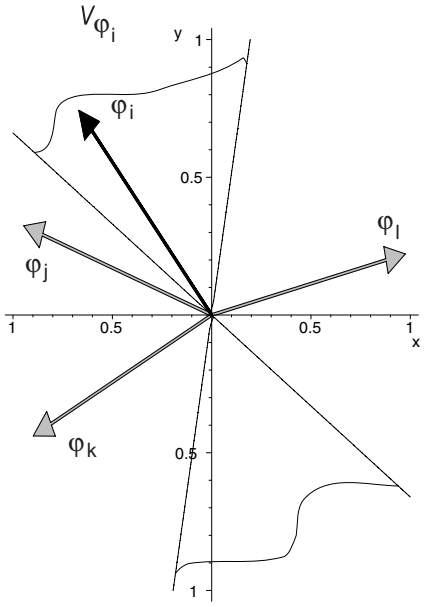

Fig. 1. Representation of the projection neighborhood $\mathcal{V}_{\gamma_i}$ in $\mathbb{R}^2$.

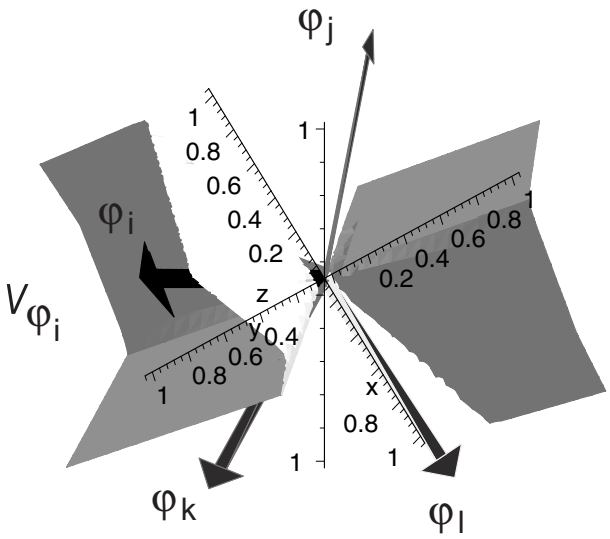

Fig. 2. Representation of the projection neighborhood $\mathcal{V}_{\gamma_i}$ in $\mathbb{R}^3$.

It has to be noted that $g_{\gamma_{i+j}}$ and $g_{\gamma_{i-j}}$ are orthogonal, i.e., $\langle\ g_{\gamma_{i+j}}\ |\ g_{\gamma_{i-j}}\ \rangle = 0$. We call $\mathcal{H}_{i+j}$ the hyperplane of dimension (L-1) defined by $g_{\gamma_{i+j}}$ and orthogonal to $g_{\gamma_{i-j}}$. Similarly let $\mathcal{H}_{i-j}$ be the hyperplane defined by $g_{\gamma_{i-j}}$ and orthogonal to $g_{\gamma_{i+j}}$. Both hyper-planes define frontiers of $\mathcal{V}_{\gamma_i}$ since they satisfy

$$|\langle x|g_{\gamma_i}\rangle| = |\langle x|g_{\gamma_j}\rangle|\,, \tag{28}$$

or equivalently

$$\begin{cases} \langle x|g_{\gamma_i} - g_{\gamma_j}\rangle = 0 & \text{if } \langle x|g_{\gamma_i}\rangle \, \langle x|g_{\gamma_j}\rangle \geq 0 \\ \langle x|g_{\gamma_i} + g_{\gamma_j}\rangle = 0 & \text{otherwise.} \end{cases} \tag{29}$$

Now define $\mathcal{I}_\alpha$ as the direction defined by the intersection $\cap_{l\in\alpha}\mathcal{H}_{i\pm l}$, where α is the (L-1)-tuple $\{k_1 \dots k_{L-1}\}$, with $k_l \neq i$. Let $\mathcal{I}_\alpha^\perp$ be the orthogonal complement to $\mathcal{I}_\alpha$. Hence, the (L-1)-dimensional hyperplane $\mathcal{I}_\alpha^\perp$ can be written as

$$\mathcal{I}_\alpha^\perp = span\,(g_{\gamma_{i+k_1}} \dots g_{\gamma_{i+k_{L-1}}}). \tag{30}$$

The orthogonal projection P_α^i of g_{γ_i} onto $\mathcal{I}_\alpha$ is equivalent to the residue of its orthogonal projection onto $\mathcal{I}_\alpha^\perp$. Therefore,

$$P_\alpha^i = g_{\gamma_i} - \sum_{l=1}^{L-1} \langle g_{\gamma_i}|\widehat{g_{\gamma_{i+k_l}}}\rangle \, \widehat{g_{\gamma_{i+k_l}}}\,, \tag{31}$$

where $g_{\gamma_{i+k_l}}$ have been orthogonalized using the Gram-Schmidt procedure. The cosine of the angle between $\mathcal{I}_\alpha$ and g_{γ_i} is given by

$$\beta_\alpha^i = \frac{|\langle g_{\gamma_i}|P_\alpha^i\rangle|}{\|P_\alpha^i\|}\,. \tag{32}$$

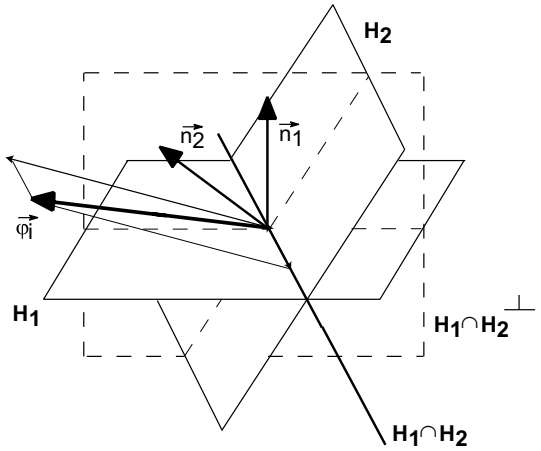

Fig. 3. The orthogonal projection of φ_i onto the intersection $H_1 \cap H_2$ is equivalent to the residue of the orthogonal projection of φ_i onto the orthogonal complement to the intersection $H_1 \cap H_2$. This representation in $\mathbb{R}^3$ generalizes to n-dimensional case.

Finally, the structural redundancy factor β is given by searching, for each vector g_{γ_i} of the dictionary, the intersection $\mathcal{I}_\alpha$ belonging to the projection neighborhood $\mathcal{V}_{\gamma_i}$ and forming the maximum angle with g_{γ_i}. In other words,

$$\beta = \min_i \min_{\mathcal{I}_\alpha \in \mathcal{V}_{\gamma_i}} \beta_\alpha^i\,. \tag{33}$$

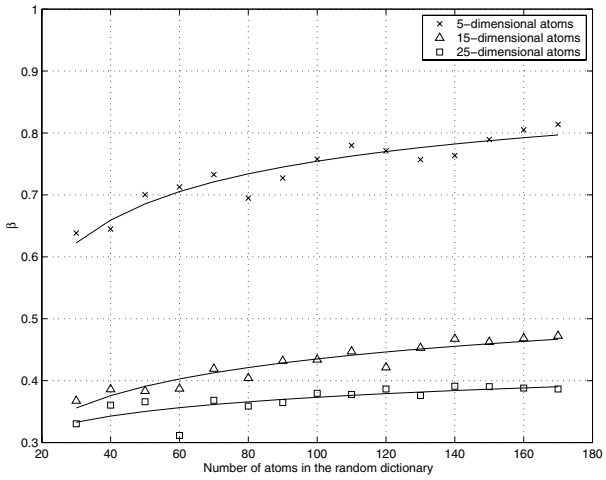

Fig. 4. Redundancy factor β as a function of the number of vectors N_v in a random dictionary. Exponential fitting function of the form: $\beta = 1 - A\, N_v^B$.

Fig. 4 represents the evolution of the redundancy factor β as a function of the size of the dictionary. The dictionary is formed of random atoms of dimension 5, 15 and 25 respectively. We can see that the redundancy factor obviously increases with the size of the dictionary. Finally, we can conjecture that the evolution of β is exponential with the number of vectors N_v. In other words,

$$\beta = 1 - A\, N_v^B \ . \tag{34}$$

Fig. 5 depicts the redundancy factor as a function of the ratio of the number of dictionary vectors N_{vec} and the size N of the vectors. It shows that the slope of the exponential fitting curves are quite similar. Therefore, for large signals, the number of vectors needed to reach a high redundancy factor explodes and becomes much bigger than the simple ratio $\frac{N_{vec}}{N}$. The redundancy definition generally used in the context of frame expansions provides therefore only a coarse qualitative approximation of the redundancy of MP dictionaries.

Finally, Fig. 6 shows that the residual energy is clearly upper-bounded by the exponential curve computed from the redundancy factor. Moreover, this upper-bound does not depend on the input function f, but, as such, may become one of its major weaknesses. It does not permit to estimate the true convergence of Matching Pursuit, since the latter does not only depend on the dictionary, but also on the input signal. In other words, the convergence will indeed be faster for an input signal whose main characteristics are represented by single dictionary vectors. However one can easily include probability-weighting in the computation of β, provided that the statistics about the input signal are known *a priori*.

C. Uniformly distributed atoms

The exact computation of β proposed in the previous section becomes heavy for a very large dictionary or a high dimensional signal, like, for example, an image. Therefore, in the case of uniformly distributed

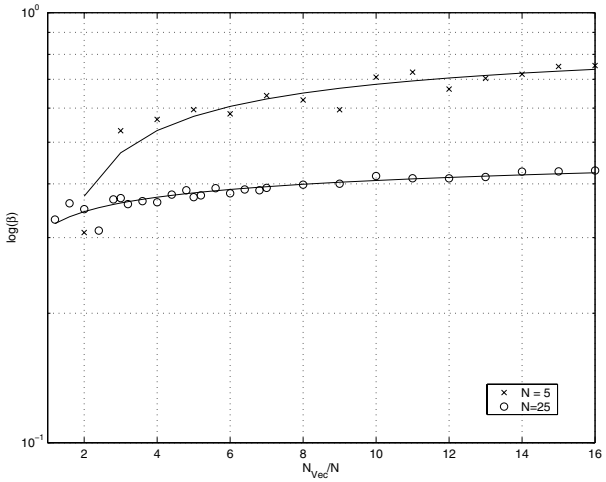

Fig. 5. Redundancy factor β as a function of the ratio between the number of vectors N_v in a random dictionary and the size N of the vectors. Exponential fitting function of the form: $\beta = 1 - A\ N_v^B$.

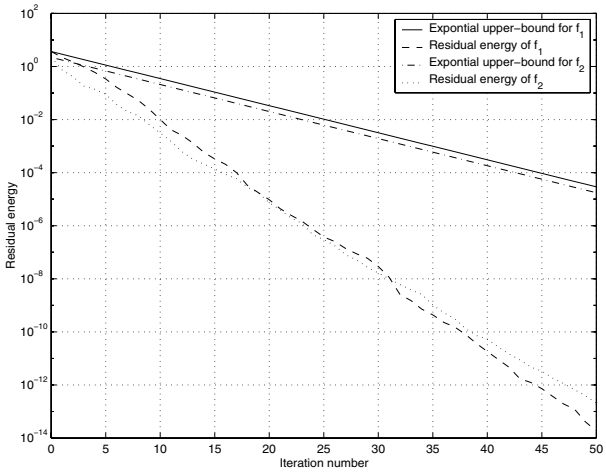

Fig. 6. Residual energy decay for Matching Pursuit of two random signals (f_1, f_2) of length 10. The dictionary is built of 50 random vectors. The exponential upper-bounds are computed with the redundancy factor β. The optimality factor α is equal to 1, since a full search over the complete dictionary has been performed at each iteration.

atoms on the L-dimensional unit sphere [19, 20] one can roughly approximate β as

$$\tilde{\beta} = \min_j \max_{i \neq j} |\langle g_i, g_j \rangle| \ . \tag{35}$$

The assumption of uniformly distributed atoms is also valid for very redundant dictionaries. The so-computed redundancy factor is however larger than the exact bound given by eq. 33. The upper-bound onto the coefficient norm may thus be under-estimated in this case. However the difference between both β values decreases rapidly with the size of the dictionary. For very redundant dictionary, this difference becomes small enough for $\tilde{\beta}$ to be used in practical cases. Moreover the upper-bound onto the coefficient decay is generally very loose since it is computed in the worst case where the input signal is at each step

the farthest possible from any dictionary vector.

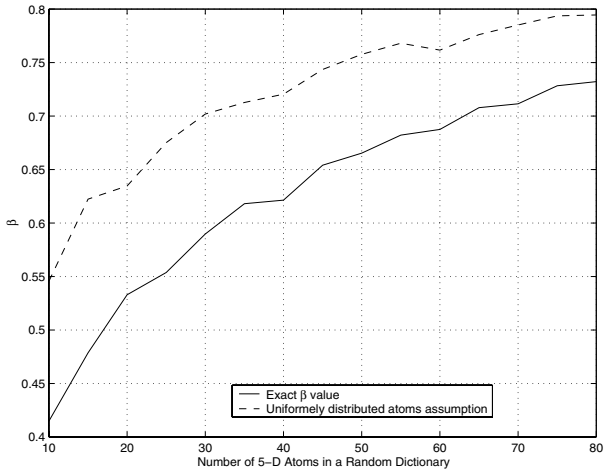

Fig. 7. Redundancy factor β and its approximation through the uniform vectors distribution as a function of the size of a random 5-dimensional vectors dictionary. The values are averaged on 100 different random dictionaries.

Fig. 7 represents the evolution of β and its approximation of a function of the size of the random dictionary. The curves have been averaged on 100 different dictionaries. They clearly show that the approximated β values are always above the exact value, and that the difference between both values decreases with the size of the dictionary, as expected.

Finally, the β parameter characterizes the redundancy of the dictionary for uniformly distributed atoms, independently from the function f to approximate. A value of β close to one implies a highly redundant dictionary, while a zero value indicates that the dictionary is not complete. In the same time, β characterizes the importance of the atom index and thus the coefficient in carrying information. Indeed, the importance of the atom index increases with the redundancy of the dictionary (i.e., the β value), while the relative importance of the coefficient decreases. In contrast, most of the information is carried on by the coefficients in orthogonal transforms.

IV. Signal Reconstruction Error

In practical cases an exact reconstruction of the signal is rarely possible. First, the number of iterations is generally limited due to computational cost and bandwidth constraints. Thus, the signal is approximated by the first N iterations of the Matching Pursuit stream. Note also the exact signal reconstruction may need an infinite number of iterations in MP with redundant dictionary. Second, the Matching Pursuit elements (i.e., coefficients and dictionary indexes) have often to be quantized to reduce the overall size of the signal representation. Indeed, the MP coder generally outputs real coefficients, which clearly need to be quantized to limit the dedicated rate. Moreover, if the dictionary is very large to ensure a good convergence, or even if it contains real parameters, the atoms should also be quantized before transmission. The framework that is considered here is an *a posteriori* quantization, which is very useful in the context where a stream is encoded once and then quantized several times to meet different constraints. The

Matching Pursuit is indeed the heaviest part of the encoding. Therefore, a different encoding for each client would be too demanding. Moreover, Matching Pursuit systems are particularly well suited for asymmetric applications, where the same stream serves several possibly heterogeneous devices. Third and finally, transmission noise can also alter the signal reconstruction. Transmission errors can be studied in the same way as quantization error, under certain conditions. All the results of this section are given for quantization error, though most of them can easily be translated into transmission error.

Let c_{γ_n} denote a MP coefficient, or equivalently the inner product $\langle g_{\gamma_n} | \mathcal{R}^n f \rangle$. The squared reconstruction error Δ between the signal approximation f_N and its reconstructed version $\tilde{f}$ can be written as

$$\Delta = \|f_N - \tilde{f}\|^2 = \| \sum_{n=0}^{N-1} c_{\gamma_n} \, g_{\gamma_n} - \sum_{n=0}^{N-1} \tilde{c}_{\gamma_n} \, g_{\tilde{\gamma_n}} \|^2 , \tag{36}$$

where $\tilde{c}_{\gamma_n}$ and $g_{\tilde{\gamma_n}}$ are respectively the distorted coefficient and atom. Let $\xi_n = c_{\gamma_n} - \tilde{c}_{\gamma_n}$ denote the error on the coefficient. Similarly, let E_n be the error on the n^{th} atom, with $E_n = g_{\gamma_n} - g_{\tilde{\gamma}_n}$. The reconstruction error can thus be decomposed further as

$$\begin{aligned} \Delta &= \| \sum_{n=0}^{N-1} \left(c_{\gamma_n} \, g_{\gamma_n} - (c_{\gamma_n} - \xi_n) \, g_{\tilde{\gamma_n}} \right) \|^2 \\ &= \| \sum_{n=0}^{N-1} \left(c_{\gamma_n} \, E_n + \xi_n \, g_{\tilde{\gamma_n}} \right) \|^2 . \end{aligned} \tag{37}$$

By triangular inequality theorem, Δ can be bounded by

$$\begin{aligned} \Delta &\leq \| \sum_{n=0}^{N-1} c_{\gamma_n} \, E_n \|^2 + \| \sum_{n=0}^{N-1} \xi_n \, g_{\tilde{\gamma_n}} \|^2 \\ &\leq \sum_{n=0}^{N-1} |c_{\gamma_n}|^2 \, \|E_n\|^2 + \sum_{n=0}^{N-1} |\xi_n|^2 \, \|g_{\tilde{\gamma_n}}\|^2 . \end{aligned} \tag{38}$$

Finally, since $\|g_\gamma\| = 1$, the error bound can be written as

$$\Delta \leq \sum_{n=0}^{N-1} |c_{\gamma_n}|^2 \, \|E_n\|^2 + \sum_{n=0}^{N-1} |\xi_n|^2 . \tag{39}$$

Eq. (39) provides an upper-bound to the reconstruction error, due to either quantization or transmission. The errors on the coefficients and indexes can thus be separated in finite dimensions. The reconstruction error is bounded by the sum of errors on both the coefficients and the atoms, weighted by the respective contributions. The first term of eq. (39) corresponds to the atom selection error multiplied by the respective coefficients. For highly redundant dictionaries, only the first few atoms participate to the reconstruction error, since the coefficient norm rapidly decreases. Meanwhile, if the atoms are correctly decoded (i.e., the first term is zero), the reconstruction error can be bounded by the sum of the error on

the coefficients. Due to the exponential convergence of MP, this error depends generally almost only on the first coefficients, i.e., the most important ones.

It has to be noted here that consistent reconstruction algorithms [4] can improve the quality of the reconstructed signal. Such concealment algorithms results in a reconstructed signal which would produce the same MP decomposition sequence as the original. However, such algorithms are not considered in this paper, due to complexity constraints on the decoder. In our case, the receiver simply reconstructs the signal from the available coefficients and atom indexes, without error detection or concealment algorithms.

In the next sections, errors on both coefficients and atoms indexes are studied in the context of quantization. The atom selection error is analyzed in details for the common Gabor atoms.

V. Quantization of Matching Pursuit Coefficients

A. A Posteriori Coefficients Quantization

This section uses Matching Pursuit properties in the quantization of coefficients. Recall that the purpose of the *a posteriori* quantization scheme lies in the context of completely asymmetric applications. Due to the encoding complexity, the Matching Pursuit stream is computed only once and then differently quantized to match various receiver requirements. The coefficient quantization cannot be included in the Matching Pursuit decomposition in contrast to [3–6, 12, 13], since this would imply dedicated encoding for each end-user. While *a posteriori* quantization provides flexibility for the quantization, its distortions are however not compensated by further MP iterations. The quantization error propagation can nevertheless be kept small for sufficiently fine quantization. Fig. 8 shows that distortions resulting

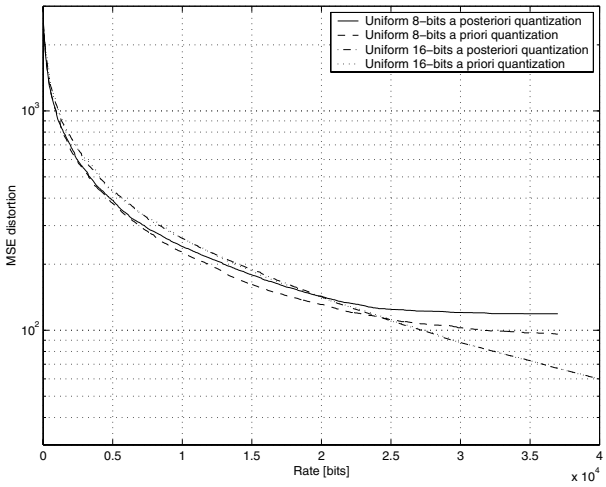

Fig. 8. Comparison of in-out the feedback loop uniform coefficient quantization (8 bits), for *Lena* image. The coding rate is dependent on the number of MP iterations.

from a posteriori quantization are slightly higher than those of the in-loop quantization, depending on the quantization step size however. Indeed, in the latter case, MP adapts to the quantized coefficient, and the residue corresponds to the stream used for reconstruction. However, such a scheme is less flexible

than the *a posteriori* quantization since the stream is targeted for a given rate. Moreover, it can be seen that the distortions increases mainly for the highest iteration orders, where coefficients become smaller. For the first few coefficients, the relative quantization error is small in both cases. Finally, for an overall fine quantization the error is similar in both quantization schemes. Note also that in-loop quantization may slow down the convergence speed of the MP.

The aim of quantization is to offer the best possible reconstruction quality for a given bit budget. The optimal solution minimizes the reconstruction error given by eq. (36), where the atoms $\|E_n\| = 0$ are correctly decoded. We propose a sub-optimal solution using the exponential decay of the residual energy. This interesting property directly drives the quantization of the coefficients in two ways. Intuitively, the quantization error on a Matching Pursuit element depends first on the iteration. Therefore, the highest iteration elements can be more coarsely quantized than the first elements. Second, the number of Matching Pursuit elements can also be adapted to the available bandwidth, since the highest iterations carry very low energy.

B. Coefficient Norm Decay

Since Matching Pursuit elects at each step only the most important atom from an overcomplete dictionary, the coefficients have a decreasing importance along the iteration number. Recall that the coefficient c_{γ_n} corresponds to the scalar product $|\langle g_{\gamma_n} | \mathcal{R}^n f \rangle$. A Matching Pursuit iteration can be thus characterized as

$$\mathcal{R}^n f = c_{\gamma_n} g_{\gamma_n} + \mathcal{R}^{n+1} f. \tag{40}$$

Since $\mathcal{R}^{n+1} f$ is orthogonal to g_{γ_n} thanks to Matching Pursuit properties, an energy conservation relation could be written for each iteration:

$$\|\mathcal{R}^n f\|^2 = |c_{\gamma_n}|^2 + \|\mathcal{R}^{n+1} f\|^2, \tag{41}$$

where $\|g_{\gamma_n}\| = 1$. The eq. (41) means that the norm of the coefficient is strongly related to the energy of the residue. From eq. (19), this norm is thus upper-bounded by an exponential function, which can be written as

$$|c_{\gamma_n}| \leq (1 - \alpha^2 \, \beta^2)^{\frac{n}{2}} \|f\| \, . \tag{42}$$

The upper-bound depends on both the energy of the input function and the construction of the dictionary. More particularly, as described in Sec. III, highly redundant dictionaries induce a very rapid decay of the coefficients norm. Indeed, since the coefficient can obviously not bring more energy than the residual function, the norm of the coefficient is strongly related to the residual energy decay curve.

Figure 9 represents the energy of the Matching Pursuit coefficients for the coding of images with a Gabor dictionary. It clearly shows that the coefficients norm can be upper-bounded by an exponential decay curve. These qualitative result does not depend neither on the signal nor on the dictionary. However, the decay rate directly depends on the redundancy factor given by eq. (33). Moreover, in this case, MP

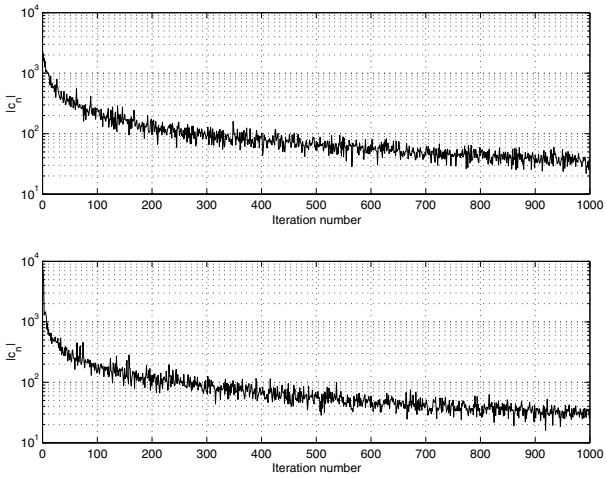

Fig. 9. Coefficient norm versus the iteration number. The MP coefficients using a 2D Gabor dictionary are represented for the *Lena* and *Autumn* images.

search is not optimal in contrast to the experiments presented for the random dictionaries. An exhaustive search would be too demanding in the case of images, for which a sub-optimal method has been provided using Genetic Algorithms [21]. This introduces a sub-optimality factor α (see eq. (21)) smaller than one and influences the decay rate of coefficients.

C. Redundancy-Driven Uniform Coefficients Quantization

The exponential upper-bound on the coefficients is now used to design an efficient quantization scheme. Indeed, there is clearly no need to quantize all coefficients on the same range, since their values decreases exponentially. In other words, the quantization applied to the first coefficients is certainly not efficient on the last ones, as the range of the latters is much smaller. Bits can thus be saved by simply limiting the quantization region between 0 and the exponential decay curve given by the parameters (i.e., β and $\|f\|$). The accuracy of the quantization is also improved, especially for decompositions over highly redundant dictionaries. For the signal to be reconstructed, the decoder needs the parameters of the exponential upper-bound. An additional bit of sign finally suffices to completely characterize the coefficients[5].

The number of coefficients, as well as the number of bits per coefficient have now to be optimized in the context of exponentially upper-bounded quantization range. The distortion at decoding can be bounded by the sum of the quantization error and the approximation error due to the limit on the number of decoded iterations. In the case where atom indexes are correctly received, using eq. (39) the total

[5]The sign bits can be reported on the rotation index in the case of structured dictionary built on anti-symmetric functions.

distortion can be written as :

$$\begin{aligned} D &\leq \sum_{n=0}^{N-1} |\xi_n|^2 + \|\mathcal{R}^N f\|^2 \\ &\leq \sum_{n=0}^{N-1} |\xi_n|^2 + (1 - \alpha^2 \beta^2)^N \|f\|^2 , \end{aligned} \tag{43}$$

where the energy of the residue at iteration N is bounded thanks to eq. (19). Assume now that the distribution of the coefficients norm is uniform between 0 and the exponential upper-bound given by eq. (42). This distribution depends on the Matching Pursuit search algorithm. Keep in mind though that the assumption of a uniform distribution is valid only as a first approximation.

For complexity reasons, and under the previous assumptions, the coefficient c_j is uniformly quantized within the exponentially decaying quantization range $I_j = \nu^j \|f\|$, where $\nu = (1 - \alpha^2 \beta^2)^{\frac{1}{2}}$. Let n_j be the number of quantization steps within I_j for the quantization of the j^{th} coefficient. In the case of fine uniform quantization, the distortion can therefore be written as :

$$D_Q = \sum_{n=0}^{N-1} |\xi_n|^2 = \sum_{j=0}^{N-1} \frac{\nu^{2j} \|f\|^2}{12 \, n_j^2} . \tag{44}$$

The optimal quantization now minimizes the distortion for a bit rate R, or equivalently, minimizes the bit rate for a given distortion. In other words, we have to find the optimal parameters $n_j \geq 1$ and N that minimizes the distortion for a given rate. The Lagrangian multiplier method [22–26] is well suited for this kinds of constrained optimization problems. It defines a cost function $\mathcal{L}(\lambda)$ as the sum of the objective distortion function and the constraint on the rate, weighted by the Lagrangian multiplier λ. In our case the cost function may be written as :

$$\begin{aligned} \mathcal{L}(\lambda) &= D + \lambda \, R \\ &= \sum_{j=0}^{N-1} \frac{\nu^{2j} \|f\|^2}{12 \, n_j^2} + \nu^{2 \, N} \|f\|^2 + \lambda \left(\sum_{j=0}^{N-1} \log_2 (n_j) + N \, B \right) , \end{aligned} \tag{45}$$

where B represents the number of bits needed to code the atom index. The index size is assumed to be constant through the iterations, even if it is possible to perform an entropy coding of the index parameters [14]. The Lagrangian formulation allows to solve the hard constrained problem of finding the optimal set of n_j and N by converting it to a set of unconstrained problems, parametrized by λ.

The optimal quantization is obtained by differentiating $\mathcal{L}(\lambda)$ with respect to both n_j and N. First, solving

$$\frac{\partial \mathcal{L}(\lambda)}{\partial n_j} = 0, \forall j \tag{46}$$

for n_j positive and finite, the optimal quantization is given by

$$n_j = \sqrt{\frac{\|f\|^2 \, \nu^{2j} \, \log 2}{6 \, \lambda}} . \tag{47}$$

The solution of eq. (47) is indeed a minimum of the Lagangian since the second derivative is positive at this point, regardless of the value of j. Hence, the optimal quantization imposes an exponential law to the number of quantization levels :

$$\frac{n_{j+1}}{n_j} = \frac{\nu^{j+1}}{\nu^j} = (1 - \alpha^2\ \beta^2)^{\frac{1}{2}}\ . \tag{48}$$

Interestingly, this bit distribution leads to an equivalent participation of each iteration to the total distortion. Indeed, the distortion per coefficient is equal to :

$$\frac{\lambda}{2\ log(2)},\ \forall n_j > 1\ , \tag{49}$$

independently of the iteration. Notice however that the Lagrangian formulation provides only an approximation to the optimal quantization. Indeed, in the practical case, n_j takes only integer powers of two values.

The appropriate number of MP iterations has now to be found to minimize the reconstruction error for a given rate. The rate can indeed be limited by transmitting only part of the stream. Moreover, it can be seen from eq. (45) that indexes can be transmitted alone, even if no coefficient is coded ($n_j \leq 1$). Depending on the redundancy of the dictionary, indexes often carry more information than coefficients. One can therefore imagine a scheme where the coefficients are simply interpolated from the exponential decay curve, especially for high order iterations (i.e., small coefficients). The optimal number of iterations N is thus given by minimizing the Lagrangian cost function of eq. (45) where the n_j have been replaced by their optimal values from eq. (47) :

$$\mathcal{L}(\lambda) = \sum_{j=0}^{N-1} \mathcal{L}_j(\lambda) + \nu^{2\ N} \|f\|^2 \tag{50}$$

with

$$\mathcal{L}_j(\lambda) = \begin{cases} \frac{\nu^{2j}\ \|f\|^2}{12\ n_j^2} + \lambda\ (\log_2 n_j + B) & \text{if } n_j \geq 1 \\ \frac{\nu^{2j}\ \|f\|^2}{12} + \lambda\ B & \text{otherwise.} \end{cases} \tag{51}$$

The Lagrangian is defined as a piecewise function since indexes can be transmitted without coefficients, as stated before. The break-point occurs at $j = N_B = \lfloor \frac{-\log(n_0)}{\log(\nu)} \rfloor$ (i.e., $n_j = 1$) where $n_0 = \sqrt{\frac{\|f\|^2\ log(2)}{6\ \lambda}}$ from Eq (47). Notice that N_B is positive only if $\lambda \leq \frac{\|f\|^2\ log(2)}{6}$. Otherwise, the weight on the rate in the Lagrangian cost function becomes much more important than the distortion, and the best scheme would be not to transmit any coefficient. The optimal number of iterations, N_{opt}, is therefore obtained by the zeros of the first derivative on both parts of the Lagrangian cost function. Thus,

$$N_{opt} = \begin{cases} N_1 & \text{if } B > -\frac{23\ \log(\nu)+1}{2\ \log(2)} \\ N_2 & \text{otherwise.} \end{cases} \tag{52}$$

where :

$$N_1 = -\frac{1 + 2\,B\,\log(2) - \log(\nu) + \log\left(\frac{\|f\|^2\,\log(2)}{6\,\lambda}\right)}{2\,\log(\nu)} - \frac{W_{-1}\left(3\,2^{3-2\,B}\,\nu\,\log(\nu)\,e^{-1}\right)}{2\,\log(\nu)}, \tag{53}$$

$$N_2 = \frac{\log\left(-\frac{\lambda\,B}{2\|f\|^2\,\log(\nu)\,\left(1+\frac{1}{12\,\nu^2-12}\right)}\right)}{2\,\log(\nu)} \tag{54}$$

are the optimum values of N on both sides of the piecewise function (i.e., for respectively $N \leq N_B$ and $N > N_B$). The choice between N_1 and N_2 depends only on the relation between B and ν. Indeed, the atom index size and the decay rate of the coefficient norm influences the decision of coding an atom without coefficients. In eq. (53), $W_{-1}(x)$ represents the second branch of the Lambert W function [27]. Finally, these solutions are indeed a minimum of the Lagrangian cost function, since the second derivative is positive at $N = N_{opt}$. Details of the computation are given in Appendix B.

Now that the unconstrained problem of eq. (45) has been solved optimally for an arbitrary λ, the next step is to find the optimal λ_{opt} that guarantees a bit budget R_{budget}. The solutions of the optimization problem forms a convex hull of the achievable rate-distortion curve, and the bit budget constraint imposes a λ_{opt} that represents the slope of the rate-distortion function at $R = R_{budget}$. Several bisection methods have been proposed to solve this second problem [23,28,29]. Since in our case we can express the bit rate in terms of the Lagrangian multiplier λ, the constraint on the bit budget directly imposes an approximated value for λ_{opt} from :

$$R_{budget} = \begin{cases} \sum_{j=0}^{N_1-1} \log_2(n_j) + N_1\,B & \text{if } B > -\frac{23\,\log(\nu)+1}{2\,\log(2)} \\ \sum_{j=0}^{N_B-1} \log_2(n_j) + N_2\,B & \text{otherwise.} \end{cases} \tag{55}$$

where N_1, N_2 and n_j are functions of λ from the above equations. The approximation is due to the constraint of integer number of bits for each coefficient. The values of λ_{opt} can finally easily be computed through numerical methods.

Fig. 10 shows the rate-distortion characteristic for a decomposition of a random signal of 10 samples over a random dictionary of 50 vectors. The redundancy-driven quantization scheme is compared to different uniform quantization schemes. The latter quantizes the coefficients over a range defined by the energy of the signal. The proposed exponentially upper-bounded quantization clearly outperforms the uniform quantization, since it adapts to the rate and to the range of coefficients to provide the best approximation for the available rate. Uniform quantization schemes provide good results for low rate, since the first coefficients are finely quantized. For high rates, however, the coefficients become too small compared to the quantization steps. The error therefore increases, since the quantized coefficient is often larger than the true value computed by the MP.

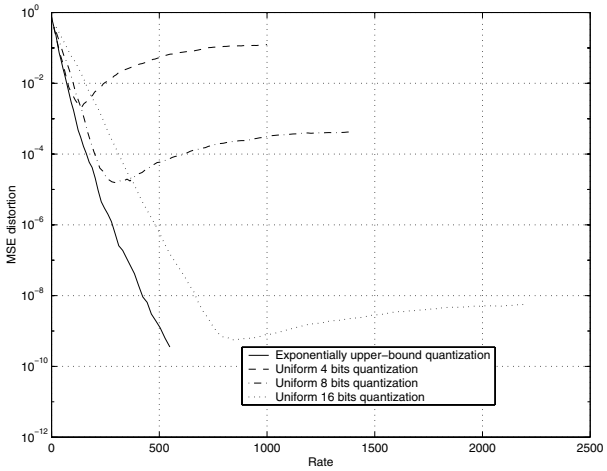

Fig. 10. Average R-D curve for the decomposition of ten random signals of length 10 over a random dictionary of 50 vectors. Comparison of exponential quantization and uniform quantization as a function of the total rate.

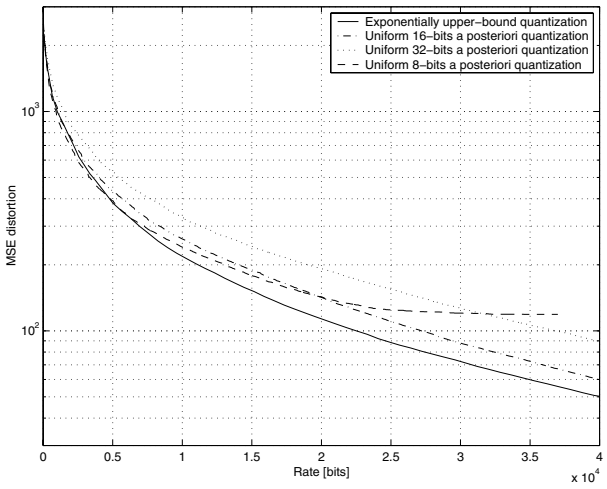

Fig. 11. Rate-distortion curves for the new quantization scheme and uniform a posteriori quantization of *Lena* MP stream. Distortion is represented as a function of the total rate.

Fig. 11 shows the behavior of the redundancy-driven quantization scheme in the practical case of image coding. The *Lena* image has been decomposed onto a structured Gabor dictionary. The coefficients have then been quantized with respectively our scheme and common uniform quantization schemes. The distortion is smaller for the proposed quantization scheme. However, for very low rates, uniform quantization can be slightly better due to mismatch between the coefficients decay rate and the values of the first coefficients (see Fig. 9). The mismatch is due to the sub-optimal search within the MP. The reconstructed image are represented in Fig. 12 to 14 for similar rates. The redundancy-driven quantization provides better reconstructed images compared to *a posteriori* and even *a priori* uniform quantization. The PSNR quality is slightly better, and the visual quality is also better since more components are used for the reconstruction, for the same transmission rate. Notice that the rates given in the previous figures

Fig. 12. Reconstruction of the *Lena* image with Gabor atoms with exponentially upper-bounded quantization. PSNR = 26.7 dB, Rate = 16.4 kbit (without entropy coding).

Fig. 13. Reconstruction of the *Lena* image with Gabor atoms with a posteriori coefficients uniform 16-bits quantization. PSNR = 25.8 dB, Rate = 16.4 kbit (without entropy coding).

Fig. 14. Reconstruction of the *Lena* image with Gabor atoms with a priori coefficients uniform 16-bits quantization. PSNR = 25.7 dB, Rate = 16.4 kbit (without entropy coding).

represents only qualitative values, and can not be used to compare Matching Pursuit with other coding schemes. The efficiency of MP is described for example in [30, 31]. Indeed, in our application, the dictionary is not optimized. Bits can be saved on rotations, positions and scales for example without decreasing the convergence (as shown in the next Section). The bit rate can thus be reduced by 25-30 percent, before additional entropy coding.

Finally, the exponentially upper-bounded uniform quantization is equivalent to the division of the coefficients by an exponential quantization table factor and multiplied by a quantizer scale factor which is given by the bit budget. The optimal number of iterations depends only on the bit budget and the design of the input dictionary. Practically speaking, the only parameters to transmit for the decoder to perform the inverse quantization are ν, $\|f\|$ and n_0, or equivalently the quantizer scale factor.

The redundancy-driven quantization becomes particularly interesting for highly redundant dictionary, where the values of the coefficients decrease very rapidly. In this case, the distribution of bits among coefficients is particularly efficient compared to uniform quantization. Similarly, for small index size the proposed scheme is also more efficient since additional elements can be sent at low price.

Finally, the redundancy-driven quantization could be improved by a periodic refreshing of the exponential parameters, or even by describing a piecewise upper-bound to fit data similar to the previous Gabor decomposition. Moreover, a limit to the maximum number of bits per coefficients can be set to ensure that the quantization of the first coefficient is not too fine.

VI. Quantization of Structured Atoms Indexes

A. A Posteriori Index Quantization

In the context of asymmetric applications, index quantization of a structured dictionary may also limit the rate to transmit the MP stream. Recall however that the *a posteriori* index quantization is not

equivalent to a Matching Pursuit encoding with a sub-dictionary. The coder works in this case with a large dictionary, thus ensuring optimal convergence of the MP. The quality of the reconstruction is then sacrified to respect some rate constraints on the transmission of the MP stream. Hence, the convergence of MP is preserved [32], but obviously the reconstruction suffers from an inexact representation, since the decoded atoms do not necessarily correspond to the atoms used within the encoding process.

The Matching Pursuit properties can help to design an appropriate index quantization scheme. First, successive atom indexes are not related to each other. Even worse, due to the Matching Pursuit encoding, successive atoms are close to orthogonal. However, it has been shown heuristically that atoms with large scale are generally encoded first, since they often capture more energy than small atoms. Second, one can take advantage of the fact that the energy captured by atoms decreases with the iteration number. The quantization of the atom indexes should therefore be finer for the first terms. Under the assumption that each Matching Pursuit iteration should participate equivalently to the reconstruction error, the error on the atom is allowed to increase exponentially with the iteration number. Indeed the error on the atom is multiplied by the respective coefficient as shown in eq. (39), while the coefficient norm decreases exponentially. The evolution of the atom selection error should therefore satisfy:

$$\frac{\|E_{n+1}\|}{\|E_n\|} = \frac{|c_{\gamma_n}|}{|c_{\gamma_{n+1}}|} = (1 - \alpha^2\ \beta^2)^{\frac{1}{2}}, \tag{56}$$

in the worst case where the Matching Pursuit coefficient norm decay follows exactly the exponential upper-bound.

The effects of atom selection error on the reconstruction are first investigated in this section, that is the error arising when the reconstruction formula (8) is biased :

$$\tilde{f} = \sum_{n=0}^{+\infty} \langle g_{\gamma_n} | \mathcal{R}^n f \rangle\, g_{\tilde{\gamma}_n}\ . \tag{57}$$

The decoder receives in this case an approximation $\tilde{\gamma}$ of the indexes chosen by the encoder. The atom selection error is shown to be bounded in the general case. This error is then analyzed in detail for structured dictionaries, and more particularly for Gabor atoms. Finally, the performance of a simple scalar quantization of Gabor atom indexes is studied in terms of rate-distortion characteristics.

B. Bounds on the Atom Selection Error

In this section, the atom selection error is shown to be bounded in the case of structured dictionaries. Let Δ_E be the error between the original signal and its reconstructed version with possible error on indexes:

$$\Delta_E = \|f - \tilde{f_E}\|^2\ . \tag{58}$$

The reconstruction error Δ_E can be bounded. The following result shows that the reconstructed signal is still in $L^2(\mathbb{R})$ provided there exists a uniform bound on each error[6].

[6]This assumption is perfectly realistic in the settings of quantization or channel losses.

Theorem 2: Let $f \in L^2(\mathbb{R})$ and let $\mathcal{D}$ be a complete dictionary of atoms. Let $\{g_{\gamma_n}\}$ be the set of atoms selected by a Matching Pursuit of f and $\{g_{\tilde{\gamma}_n}\}$ be the associated biased set of atoms. The reconstruction error Δ_E is bounded if there exists a constant C such that

$$\|g_{\tilde{\gamma}_n} - g_{\gamma_n}\|^2 < C\,, \quad \text{uniformly in } n. \tag{59}$$

Proof : From eq. (58), the reconstruction error Δ_E can be written as :

$$\Delta_E \leq \sum_{n=0}^{+\infty} |c_{\gamma_n}|^2 \|E_n\|^2\,. \tag{60}$$

Defining $S_n = |c_{\gamma_n}|^2$, eq. (60) becomes

$$\Delta_E \leq \sum_{n=0}^{+\infty} S_n \|E_n\|\,. \tag{61}$$

Now we immediately have $S \in \ell^1$ because

$$\sum_{n=0}^{+\infty} S_n = \|f\|^2\,, \tag{62}$$

and, by Köthe duality [33], the sum on the right-hand side of equation (61) converges if and only if $G \in \ell^\infty$, that is if and only if there exists $C > 0$ such that $\|E_n\| < C$, uniformly in n. Finally, the total error is thus bounded by :

$$\Delta_E \leq C\,\|f\|^2 \sup_n \|E_n\|^2\,. \tag{63}$$

■

Let us now focus on the last term in this equation. Suppose our dictionary was built using the technique described in Section II-B. By definition of E_n, we then have

$$E_n = \mathcal{U}(\gamma_n)g - \mathcal{U}(\tilde{\gamma}_n)g\,, \tag{64}$$

which can be computed as :

$$\begin{aligned} \|E_n\|^2 &= \langle \mathcal{U}(\gamma_n)g - \mathcal{U}(\tilde{\gamma}_n)g | \mathcal{U}(\gamma_n)g - \mathcal{U}(\tilde{\gamma}_n)g\rangle & (65)\\ &= 2\big(1 - \Re\{\langle \mathcal{U}(\gamma_n)g | \mathcal{U}(\tilde{\gamma}_n)g\rangle\}\big) & (66)\\ &= 2\big(1 - \Re\{\mathcal{K}_g[\gamma_n;\tilde{\gamma}_n]\}\big)\,, & (67)\end{aligned}$$

where $\mathcal{K}_g$ stands for the reproducing kernel associated to the dictionary and $\Re\{x\}$ represents the real part of x. Using the unitarity of the representation $\mathcal{U}$, we can also write

$$\|E_n\|^2 = 2\Big(1 - \Re\{\mathcal{K}_g[\tilde{\gamma_n};\gamma_n]\}\Big)\,. \tag{68}$$

eq. (68) shows that the kernel $\mathcal{K}$ endowes the elements of the dictionary $\mathcal{D}$ with a notion of distance. Indeed, when $\tilde{\gamma_n} = \gamma_n$, the error vanishes, $\Delta = 0$. In the next section, the atom selection error is computed in the practical case of Gabor atoms.

C. Gabor Atom Selection Error

Following eq. (68), we now give explicit estimates of the error on the atoms in the settings of a Gabor dictionary [34]:

$$h(\vec{x}) = \frac{1}{2\ \pi\ \sigma^2} \cos\left(\vec{\omega_0} \cdot \vec{x}\right)\ e^{-\frac{\|\vec{x}\|^2}{4\ \sigma^2}}\ , \tag{69}$$

where $\vec{\omega_0}$, generally equal to $(\omega_0, 0)$, with ω_0 even, is the oscillation parameter and σ is a scale parameter. The MP encoder transmits in this case translation, scaling and rotation parameters applied to the elementary Gabor function. The atom selection error corresponds therefore to the error arising from decoding the atoms with distorted translation, scaling and rotation indexes.

The formulation of eq. (69) corresponds to the real part of the complex function:

$$\psi(\vec{x}) = \frac{1}{4\ \pi\ \sigma^2}\ e^{i\ \left(\vec{\omega_0}\cdot\vec{x}\right)}\ e^{-\frac{\|\vec{x}\|^2}{4\ \sigma^2}}\ . \tag{70}$$

Making use of this complex form of the Gabor atom we can write :

$$g(\vec{x}) = \frac{\psi(\vec{x}) + \overline{\psi(\vec{x})}}{\|h(\vec{x})\|} \tag{71}$$

The atom selection error is given by re-writing the inner product :

$$\begin{aligned} \langle g_\gamma\ ,\ g_{\tilde{\gamma}}\rangle &= \frac{\langle \psi_\gamma\ ,\ \psi_{\tilde{\gamma}}\rangle + \langle\overline{\psi_\gamma}\ ,\ \psi_{\tilde{\gamma}}\rangle + \langle \psi_\gamma\ ,\ \overline{\psi_{\tilde{\gamma}}}\rangle + \overline{\langle \psi_\gamma\ ,\ \psi_{\tilde{\gamma}}\rangle}}{\|h(\vec{x})\|^2} \\ &= \frac{2\ \left(\Re\langle \psi_\gamma\ ,\ \psi_{\tilde{\gamma}}\rangle + \Re\langle \psi_\gamma\ ,\ \overline{\psi_{\tilde{\gamma}}}\rangle\right)}{\|h(\vec{x})\|^2}, \end{aligned} \tag{72}$$

since $\|h(\vec{x})\| = \|\tilde{h}(\vec{x})\|$ thanks to the properties of the representation $\mathcal{U}$. Moreover, the norm of h depends only on the oscillation and scale parameters (ω_0, σ) and is given by :

$$\|h(\vec{x})\|^2 = \frac{1\ +\ e^{-2\ \sigma^2\ \omega_0^2}}{4\ \pi\ \sigma^2}. \tag{73}$$

Assume now that a, θ, b_1 and b_2 respectively represents the scale, rotation and positions parts of the error on the index $\tilde{\gamma}$. The index γ is given in a 2-dimensional Gabor dictionary by the quadruplet $(\vec{p}\ ,\ s\ ,\ \varphi)$ which respectively correspond to the position, scale and rotation parameters of the basic Gabor function. Thanks to the group properties of the representation $\mathcal{U}$, the index error is therefore written as [35]:

$$\begin{aligned} (\vec{b}\ ,\ a\ ,\ \theta) &= \tilde{\gamma}^{-1} \circ \gamma \\ &= \left(\tilde{s}^{-1} r_{-\tilde{\varphi}}(\vec{p} - \vec{\tilde{p}})\ ,\ \frac{s}{\tilde{s}}\ ,\ \varphi - \tilde{\varphi}\right)\ . \end{aligned} \tag{74}$$

The inner product between two complex Gabor atoms given by the parameters σ and ω_0 can then be expressed as :

$$\langle \psi_\gamma\ ,\ \psi_{\tilde{\gamma}}\rangle = \frac{a}{4\ \pi\ \sigma^2\alpha^2} e^{-\sigma^2\Gamma^2} e^{-\frac{b_1^2+b_2^2}{4\sigma^2\alpha^2}} e^{-i\frac{\beta_1 b_1 + \beta_2 b_2}{\alpha^2}}\ , \tag{75}$$

where

- $\alpha = \sqrt{a^2+1}$
- $\Gamma^2 = 2\omega_0^2 - \frac{\beta_1^2+\beta_2^2}{a^2+1}$
- $(\kappa_0\ ,\ \eta_0) = (\omega_0\ \cos\theta\ ,\ \omega_0\ \sin\theta)$
- $\beta_1 = a\kappa_0 + \omega_0, \beta_2 = a\eta_0.$

Using the properties of complex numbers, the inner product can be rewritten as :

$$\langle g_\gamma\ ,\ g_{\tilde{\gamma}}\rangle = \frac{a\ e^{-\frac{b_1^2+b_2^2}{4\sigma^2\alpha^2}}}{2\pi\ \|h(\vec{x})\|^2\ \sigma^2\alpha^2}\ \Big[e^{-\sigma^2\Gamma^2}\cos\frac{\beta_1 b_1+\beta_2 b_2}{\alpha^2} + e^{-\sigma^2\hat{\Gamma}^2}\cos\frac{\hat{\beta}_1 b_1+\hat{\beta}_2 b_2}{\alpha^2}\Big]\ , \tag{76}$$

where

- $(\hat{\kappa}_0\ ,\ \hat{\eta}_0) = (\omega_0\ \cos(\theta+\pi)\ ,\ \omega_0\ \sin(\theta+\pi))$
- $\hat{\Gamma}^2 = 2\omega_0^2 - \frac{\hat{\beta}_1^2+\hat{\beta}_2^2}{a^2+1}$
- $\hat{\beta}_1 = a\hat{\kappa}_0 + \omega_0, \hat{\beta}_2 = a\hat{\eta}_0.$

Finally, from eq. (68), the atom selection error is expressed as

$$\|E_n\|^2 = 2\ (1\ -\ \langle g_\gamma\ ,\ g_{\tilde{\gamma}}\rangle)\ . \tag{77}$$

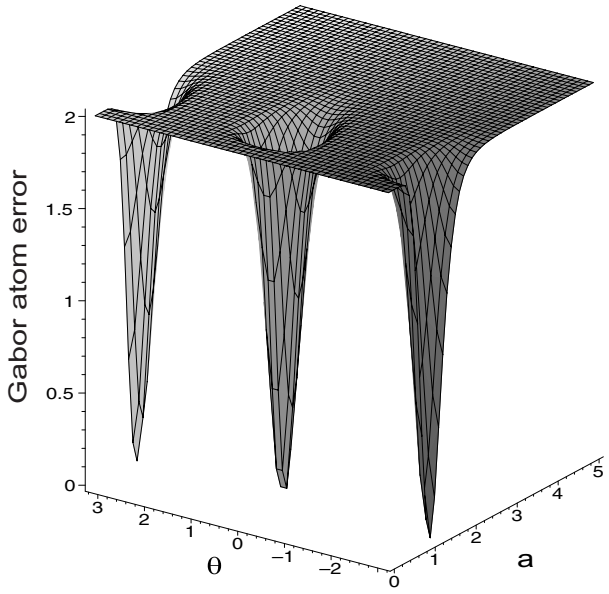

Fig. 15. Gabor atom selection error as a function of a and θ, respectively the scale and rotation part of the error on indexes.

The error on Gabor atoms is therefore clearly bounded since its value can not be larger than 4. Indeed, since atoms have a unit norm by definition, the error arising from a completely wrong atom selection can

Fig. 16. Reconstruction of the *Lena* image with Gabor atoms after 1000 MP iterations (PSNR = 31.17 dB).

Fig. 17. Reconstruction of the *Lena* image with Gabor atoms, where all positions indexes have been quantized (PSNR = 22.35 dB). The number of positions indexes has been divided by two compared to the encoding dictionary.

Fig. 18. Reconstruction of the *Lena* image with Gabor atoms, where all rotation indexes have been quantized (PSNR = 27.72 dB). The number of rotation indexes has been divided by two compared to the encoding dictionary.

not grow indefinitely. The error forms moreover a convex function along the axis representing the error on the scale parameter. The minimum of the error is reached when $a = 1$ (i.e., when the error on the scale is null), and then increases when a goes both towards zero or infinity. Due to Gabor functions properties, the error is clearly periodic with the error onto the rotation parameter. Finally, it can be seen also from eq. (74) that, for a given reconstruction error, the allowable error on position increases with the scale. This can be be explained by the decrease of the relative importance of the position of large scale atoms. The evolution of the reconstruction error as a function of the error on the position index depends on both on the scale and the rotation indexes (see eq. (74)). It is therefore not straightforward to handle it, since it does form neither a convex nor a periodic function. Finally, Gabor atoms are particularly sensitive to position error due to the oscillations within the elementary function.

Fig. 15 shows the variation of the atom selection error as a function of the scale and rotation components of the error on the Gabor atom indexes. As expected from the last section, the reconstruction error is clearly bounded. It can be seen that the reconstruction error is a convex function of a, the error on the scale. Moreover, due to the oscillations in the Gabor atoms, oscillations are also present along the axis of θ, the error on the rotation index. As expected, the reconstruction error is minimal when the error on the rotation parameter correspond to multiples of π, due to the symmetry in Gabor atoms.

D. Scalar Quantization of Structured Atom Indexes

To avoid a complex decoding we restrain the atom quantization to scalar quantization of structured atom indexes. Atom index quantization is a typical candidate to apply Vector Quantization [36], where bits can be saved on the dictionary size while guaranteeing a uniform distribution of the atoms in the

sub-dictionary. Vector Quantization is therefore expected to generate less degradation than scalar quantization of atom indexes for the same rate reduction. However, the quantization and inverse quantization are more complex. Indeed, to avoid the transmission of the complete dictionaries, the decoder should be able to build uniformly distributed dictionaries to perform the inverse quantization.

In the case of structured or semi-structured dictionaries (see Sec. II-B and II-C) the parameters of the structured dictionary can be *a posteriori* quantized to cope with the decoding complexity problems. In other words, the atom index i is quantized as

$$i_q = i_0 + k \, \frac{i_N - i_0}{N} = i_0 + k \, \delta_i \, , \tag{78}$$

where i_0 and i_N are the bounds on the index value, N is the number of quantization steps and k is the transmitted bin ($k \in [0, N]$). A scalar quantization is performed separately on the position, scale and rotation parameters.

Figures 17 and 18 show the effect of *a posteriori* quantization of respectively the position and rotation indexes in a Gabor dictionary. Even though the relative error is not exactly the same in both cases, it is clear that the atom selection error is much more sensitive to position indexes, as explained in the previous section. Indexes quantization can generate annoying distortion effects if not handled carefully, as seen in the same figures. It may provide flexibility in the design of asymmetric applications, but obviously degrades the picture quality compared to a MP with the quantized sub-dictionary.

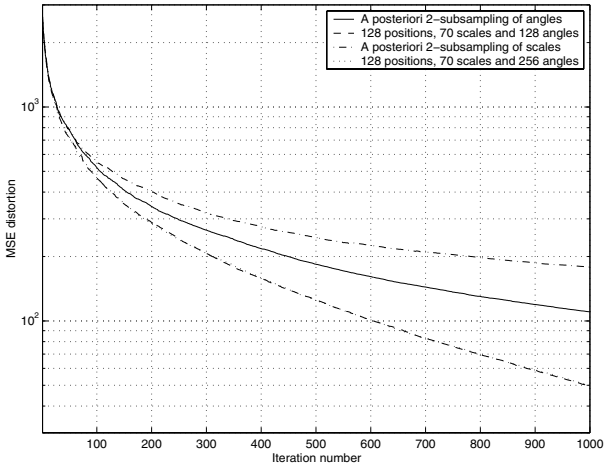

Fig. 19. Evolution of the reconstruction error of *Lena* for Gabor sub-dictionaries and index *a posteriori* quantization. Curves taken by pairs have the same dictionary size.

Fig. 19 shows the reconstruction error as a function of the iteration number of Matching Pursuit, for *a posteriori* quantization of rotation and scale indexes, and for an encoding over the quantized sub-dictionaries. It clearly shows that *a posteriori* quantization, for the same dictionary size, causes more distortion than a MP encoding on the quantized dictionary. For the first iterations, the two schemes are very similar, since generally first atoms are very large and therefore less sensitive to errors. For the higher iterations however, the reconstruction error clearly converges more slowly in the case of *a*

posteriori quantization.

Clearly there is a trade-off between convergence of the MP encoding (i.e., the size or redundancy of the dictionary) and the coding rate of atom indexes. We have seen that the Gabor atoms are really sensitive to index *a posteriori* quantization. However, this sensitivity is driven by the error on the indexes. For the same rate reduction, this error decreases when the size of the dictionary increases. Indeed, saving one bit in a dictionary of R rotation indexes generates more distortion than in a double-sized dictionary, for example.

To illustrate the trade-off between the size of the initial dictionary and *a posteriori* index quantization, Figures 20 and 21 represents the reconstruction distortion in the case of the *Lena* 128×128 image with 1000 Gabor atoms as a function of the coding size of the atom index. It is clear that, for the same index size, the index quantization is worse than an encoding on a smaller dictionary. In this case the approximation converges slower than for large dictionaries, but there is no error propagation, contrarily to the case of the *a posteriori* quantization. For very large dictionaries, the quantization error is small for fine quantization (i.e. gain of one or two bits per index). However, it can be seen from the Figure 20 that the convergence of the reconstruction error with the dictionary size saturates relatively early, for relatively small dictionaries. Hence encoding dictionary has clearly to satisfy a trade-off between its size to limit the transmission rate, and its properties of convergence. In sight of the previous results, only very fine scalar index quantization may be allowed, under penalty of very large reconstruction distortion.

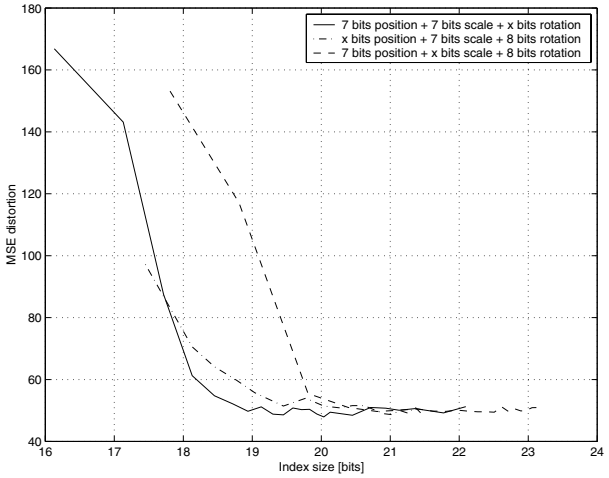

Fig. 20. Rate-distortion representation between size of the Gabor dictionary index and MSE distortion after 1000 iterations on the *Lena* 128×128 image.

Finally, the results presented here above are only valid for Gabor atoms, which are particularly sensitive to position error due to oscillations. A similar development can however be done for any structured dictionary as defined in Sec. II-B. Anisotropic dictionaries, for example, are expected to be less sensitive to index quantization due to their inherent properties.

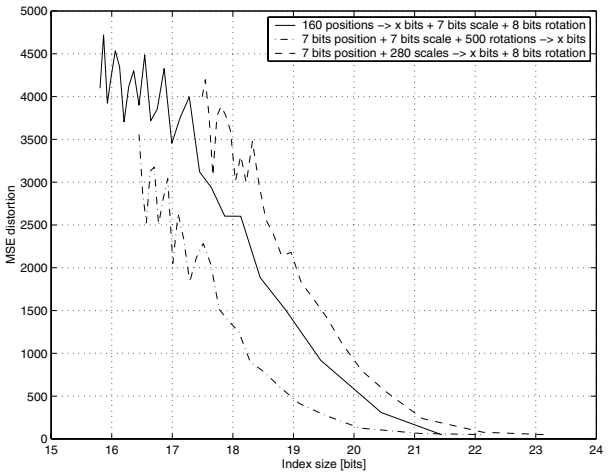

Fig. 21. Rate-distortion representation between size of the Gabor index after quantization and MSE distortion after 1000 iterations on the *Lena* 128×128 image.

E. Redundancy-Driven Gabor Indexes Quantization

The main design features of the scalar index quantization have been drawn in the previous paragraphs. However, the detailed scheme is closely related to the chosen dictionary. The problem becomes now to allocate bits between the different indexes (e.g., positions, scales, angle) so that the error on the atoms given by eq. (56) is respected. In other words, the number of quantization steps of each index has to be properly determined, while respecting the complexity constraints at decoder. Finally, the optimal bit distribution, in a rate-distortion sense, between coefficients and the different atom indexes is obtained by a Lagrangian formulation similar to eq. (45). In particular, the upper-bound on the reconstruction error given by eq. (39) should be minimized for a given encoding rate. One has to minimize the following general Lagrangian formulation :

$$\begin{aligned}\mathcal{L}(\lambda) &= D + \lambda\, R \\ &= \sum_{j=0}^{N-1} \nu^{2j}\, \|f\|^2 \left(\frac{1}{12\, n_j^2} + \|E_j(p_j, s_j, r_j)\|^2 \right) + \nu^{2\,N} \|f\|^2 \\ &+ \lambda \sum_{j=0}^{N-1} \left(\log_2(n_j) + \log_2(p_j) + \log_2(s_j) + \log_2(r_j) \right) ,\end{aligned} \tag{79}$$

where n_j, p_j, s_j and r_j represents the number of quantization steps, at iteration j, for respectively the coefficient and the position, scale and rotation indexes. The solution of the Lagrangian formulation is obtained by differentiating eq. (79) with respect to n_j, p_j, s_j, r_j and N, as presented in the previous section.

Let us now focus on the scalar quantization of Gabor atom indexes. For simplicity reasons, let assume moreover that no error is allowed on the position indexes (i.e., $b_1 = b_2 = 0$ in eq. 77). The reconstruction error is therefore convex around $a = 0$ and $\theta = 0$. Moreover, let us assume a logarithmic distribution

of scale indexes, since accuracy of scale indexes should be finer for small atoms. The scale index s_n is therefore given by :

$$s_n = s_0 \left(\frac{S}{s_0}\right)^{\frac{n}{K}} = s_0 \, \delta_s^{\frac{n}{K}} , \tag{80}$$

where s_0 is the smallest scale, S is the largest scale (generally the size of the image) and N is the number of allowed scales within these two bounds, or equivalently the number of quantization steps. The maximal quantization error onto the scale index is therefore written as :

$$2^{-\frac{S}{2\,K}} \leq a \leq 2^{\frac{S}{2\,K}} , \tag{81}$$

when scales are given by powers of two. The rotation parameter is uniformly quantized between 0 and $2\,\pi$:

$$\varphi_m = m \, \frac{2\,\pi}{L} = m \, \delta_\varphi \, . \tag{82}$$

The redundancy-driven quantization problem is now equivalent to find the optimal step size for each indexes (i.e., δ_s, and δ_φ), so that the reconstruction error is minimized for a given rate. The solution of this optimization problem is again given by a Lagrangian formulation of the form :

$$\begin{aligned} \mathcal{L}(\lambda) &= D + \lambda \, R \\ &= \sum_{j=0}^{N-1} \nu^{2j} \, \|f\|^2 \, \left(\frac{1}{12 \, n_j^2} + \|E_j(2^{\frac{S}{2\,K_j}}, \frac{\pi}{L_j})\|^2 \right) \\ &+ \nu^{2\,N} \|f\|^2 + \lambda \, \sum_{j=0}^{N-1} \left(\log_2 \left(n_j\right) + \log_2 \left(K_j\right) + \log_2 \left(L_j\right) \right) \, , \end{aligned} \tag{83}$$

where the atom selection error $\|E_n\|$ is given by eq. (77). The evolution of the coefficient step-size given by n_j has already been computed in the previous section. The optimal index quantization is given by the minimum of the Lagrangian, obtained by differentiating with respect to K_j and L_j. It can be shown that the minimum of the Lagrangian formulation is given for both K_j and L_j as large as possible, or a quantization error as small as possible, independently of the iteration. Unless the constraint on the rate is very high (i.e., for large λ values), the index quantization has to be really fine and even avoided. When the constraint on the rate is very high, or the coefficient energy is very small, the iterations have to be skipped to reach to optimal distortion under a given rate constraint. The optimal scheme is therefore directly given by the exponentially upper-bounded coefficient quantization presented in the Sec. V, where the number of iterations to be transmitted has been computed.

The interesting behavior of the Lagrangian is due to the very rapid increase of the reconstruction error in comparison to the decrease of coding rate while the quantization step size increases. The atom selection error is moreover bounded, so that coarse quantization do not increase the distortion anymore when the upper-bound has been reached. Similar results were presented in the previous section, where

Gabor atoms where shown to be very sensitive to index quantization. The quantization of Gabor atoms must therefore be avoided, and the encoding dictionary should be carefully chosen, even dynamically, so that coding rate is kept minimal.

Furthermore, a careful position index quantization can however be worthy if quantization is based on the value of the scale index. Indeed, for large scale indexes, the position can be coarsely quantized without introducing a large reconstruction error (see eq. (76)), and thus saving bits onto both position indexes.

The optimal index quantization scheme, in R-D sense, is certainly a Vector Quantization satisfying eq.(56). Another less complex, but suboptimal, quantization scheme could however be proposed, taking part of MP properties. Indeed, as generally larger scale atoms are encoded first, one can think about a hierarchical encoding scheme, where the dictionary is updated periodically. The size of the dictionary is thus kept small for the first iterations, and then grows with the iteration number. Bits can thus be saved to code atom indexes, since the resolution of the dictionary evolves during the encoding. The scheme provides an efficient way to reduce the coding rate, while ensuring a good convergence of the MP and avoiding *a posteriori* index quantization.

Finally, the adaptivity to the rate constraints imposed by the decoder has more to be seen in terms of number of transmitted iterations, besides coefficients quantization. Indeed, as we have seen, the scalar index quantization must be avoided, since it generally induces too important degradations. However, the encoding dictionary has to be carefully designed to avoid the need of index quantization. Note that the results presented in the last section are valid only for index scalar quantization, and more particularly for Gabor atoms. The conclusions however are expected to be the same for other dictionaries, since quantization errors behaves similarly.

VII. Conclusions

A new approach of Matching Pursuit coding has been proposed in this paper. It describes an *a posteriori* quantization of the MP stream to meet the constraints of asymmetric applications targeting an heterogeneous set of decoders. The encoder takes advantage of the intrinsic properties of the MP stream, and more particularly of the exponential decay of the coefficients. The exponential decay has been shown to depend only on the dictionary structure. Depending on the parameters of this exponential upper-bound, a redundancy-driven coefficient quantization scheme has been proposed. This scheme clearly outperforms uniform quantization schemes since it adapts to the coefficient decay to provide the best possible approximation with the lowest coding rate.

Similarly to the coefficient quantization, an simple index quantization scheme was proposed to reduce the size of atom index. However, for sake of simplicity, the quantization was assumed to be uniform and independent for each atom indexes. The Gabor atoms have been shown to be particularly sensitive to index quantization, and do not allow for a simple index quantization scheme. However, a index quantization scheme based on Vector Quantization, and thus ensuring a uniform error distribution, is expected to give much better results than scalar index quantization. The decoding complexity becomes in this case too important for the stream to be decoded on any device in the context of heterogeneous

receivers. Similarly, a non uniform scalar quantization, taking part of perceptual characteristics for example, would surely improve the signal reconstruction by distributing bits where they are most needed.

Finally, entropy coding was not considered in this paper. Entropy coding can be performed on top of the quantization to reduce coding rate [14]. One can also think about an entropy-constrained quantization, where the entropy is considered in the design of the quantization scheme. The design of optimal dictionaries was also not considered in this paper, although its importance has been emphasized.

APPENDIX

I. GROUPS AND GROUP REPRESENTATIONS

In this appendix we collect some definitions and properties of groups and group representations. We also give explicit examples of group constructions that are used in the paper. Interested readers can find further details in [37].

Definition 2: Let G be a group and $\mathcal{H}$ be a Hilbert space. A unitary representation $\mathcal{U}$ of G in $\mathcal{H}$ is a homorphism between G and the set of unitary operators on $\mathcal{H}$ with composition as the group law :

$$\begin{aligned} \mathcal{U} : G &\mapsto U(\mathcal{H}) \\ \gamma &\mapsto \mathcal{U}(\gamma) \end{aligned}$$

The following properties, straightfully derived from this definition, are valid for any γ, γ_1, $\gamma_2 \in G$ and used throughout the paper :

$$\begin{aligned} \mathcal{U}(\gamma_1 \circ \gamma_2) &= \mathcal{U}(\gamma_1)\mathcal{U}(\gamma_2) \\ \mathcal{U}(\gamma^{-1}) &= \mathcal{U}^{-1}(\gamma) \\ \mathcal{U}(e) &= \mathbb{I} \end{aligned}$$

where $\circ$ denotes the group law, e is the neutral in G and $\mathbb{I}$ is the identity operator. An example used in the paper is the *similitude group of the plane* :

$$\mathrm{SIM}(2) = \mathbb{R}^2 \rtimes \left(\mathbb{R}_*^+ \times \mathrm{SO}(2)\right).$$

This group is composed of dilations, rotations (e.g., $\mathrm{SO}(2)$) and 2-D translations. Rotations and dilations commute but have a non-trivial action on translations. Indeed, writing a generic element $\gamma = (\vec{b}, a, \theta) \in \mathrm{SIM}(2)$, the group law is given by :

$$\left(\vec{b}, a, \theta\right) \circ \left(\vec{b'}, a', \theta'\right) = \left(\vec{b} + a r_\theta \, \vec{b'}, a \cdot a', \theta + \theta'\right), \tag{84}$$

where r_θ is a rotation matrix. Accordingly, the neutral and inverse elements read :

$$e = \left(\vec{0}, 1, 0\right), \quad \left(\vec{b}, a, \theta\right)^{-1} = \left(-a^{-1} r_{-\theta} \, \vec{b}, a^{-1}, -\theta\right).$$

Eq. (84) shows that $\mathbb{R}^2$ is endowed with a structure of G-space which enables us to compute an action σ of $\mathrm{SIM}(2)$ on the plane :

$$\begin{aligned} \sigma[\gamma] \; &: \; \mathbb{R}^2 \to \mathbb{R}^2 \\ \sigma\left[\left(\vec{b}, a, \theta\right)\right]\vec{x} &= a \cdot r_\theta \vec{x} + \vec{b}. \end{aligned} \tag{85}$$

Finally a unitary representation of SIM(2) in the Hilbert space $L^2\left(\mathbb{R}^2\right)$ is given by :

$$\mathcal{U}(\gamma) f\left(\vec{x}\right) = J\left(\sigma[\gamma]\right)^{-1/2} f\left(\sigma\left[\gamma^{-1}\right]\vec{x}\right),$$

where $J(\sigma)$ is the Jacobian of the transformation (85).

II. Optimal Number of Iterations

In this appendix the details of the computation of the optimal number of MP iterations N_{opt} are presented in the case of a redundancy-driven quantization of the coefficients.

Recall that the Lagrangian cost function of eq. (50) is piecewise defined, where the break-point corresponds to the first non-transmitted coefficient. The minimum are now computed for the right and left sides of the piecewise Lagrangian cost function. The solutions are shown to depend on the atom index size and the convergence rate, respectively the parameters B and ν.

The first derivative of the right part of the Lagrangian (i.e., for $N > N_B$) can be written as :

$$\frac{\partial \mathcal{L}(\lambda)}{\partial N} = \lambda\, B + 2\, \|f\|^2 \, \left(1 + \frac{1}{12\, \nu^2 - 12}\right)\, \nu^{2\, N}\, \log \nu\, . \tag{86}$$

The zero of the eq. (86) is given by :

$$N_2 = \frac{\log\left(-\frac{\lambda\, B}{2\|f\|^2\, \log(\nu)\, \left(1+\frac{1}{12\, \nu^2-12}\right)}\right)}{2\, \log(\nu)}\, . \tag{87}$$

However, this equation is valid only for $N > N_B$, by definition. The condition for N_2 to be larger than the value of N_B is given by :

$$\frac{\log\left(-\frac{\lambda\, B}{2\|f\|^2\, \log(\nu)\, \left(1+\frac{1}{12\, \nu^2-12}\right)}\right)}{2\, \log(\nu)} > \frac{-\log\left(\sqrt{\frac{\|f\|^2\, \log(2)}{6\, \lambda}}\right)}{\log(\nu)}\, , \tag{88}$$

or, for $\nu \in]0,1[$ and B positive,

$$B < -\frac{12\, \log(\nu) + \frac{\log(\nu)}{\nu^2-1}}{\log(2)}\, . \tag{89}$$

The second derivative of the right part of the Lagrangian cost function (i.e., for $N > N_B$) is always positive. Indeed,

$$\frac{\partial^2 \mathcal{L}(\lambda)}{\partial N^2} = 4\, \|f\|^2 \, \left(1 + \frac{1}{12\, \nu^2 - 12}\right)\, \nu^{2\, N}\, \log^2 \nu\, \geq 0\, , \tag{90}$$

for $\nu \in]0,1[$. Hence N_2 is a minimum of the Lagrangian cost function provided that the constraint of eq. (89) is satisfied. The constraints indicates whether it is worth transmitting the indexes without coefficients and depends only on the size of indexes B and the decay rate of the energy of the coefficients ν.

The first derivative of the left side of the Lagrangian cost function (i.e., for $N < N_B$) can be written as :

$$\begin{aligned}\frac{\partial \mathcal{L}(\lambda)}{\partial N} &= \frac{\lambda \, \log(\nu)}{\log(2)} \, N + 2 \, \nu^{2 \, N} \, \log(\nu) \, \|f\|^2 \\ &+ \frac{\lambda}{\log(2)} \left(\frac{1}{2} + \log \sqrt{\frac{\|f\|^2 \, \log(2)}{6 \, \lambda}} - \frac{\log(\nu)}{2} + B \, \log(2) \right) .\end{aligned} \tag{91}$$

For the sake of clarity, let rewrite eq. (91) as :

$$\frac{\partial \mathcal{L}(\lambda)}{\partial N} = a + b \, x + d \, e^x , \tag{92}$$

with

- $a = \frac{\lambda}{2 \, \log(2)} + \lambda \, B - \frac{\lambda \, \log(\nu)}{2 \, \log(2)} + \frac{\lambda}{2} \, \log\left(\frac{\|f\|^2 \, \log(2)}{6 \, \lambda} \right)$,
- $b = \frac{\lambda}{2 \, \log(2)}$,
- $d = 2 \, \|f\|^2 \log(\nu)$,
- $x = 2 \, N \, \log(\nu)$.

The eq. (92) presents possibly several zeros given by :

$$x = - \frac{a + b \, W\left(\frac{d}{b} \, e^{-\frac{a}{b}} \right)}{b} , \tag{93}$$

where $W(z)$ is the Lambert W function [27], which can present two real branches $W_0(z)$ and $W_{-1}(z)$ for $z \in [-e^{-1}, 0]$. First it can be shown that the argument of W is indeed in the region where two branches are defined. It can be shown that, for $B > 1$ [7], the argument of W is negative and larger than $-e^{-1}$. Indeed,

$$-e^{-1} \leq \left(3 \, 2^{3-2 \, B} \, \nu \, \log(\nu) \, e^{-1} \right) < 0 , \tag{94}$$

for $\nu \in]0, 1[$. This ensures the existence of two solutions for N_{opt} and hence that the Lagrangian cost function presents two optimum which are studied now. From the definition of the Lambert W function, $0 \geq W_0(z) \geq -1$ and $W_{-1}(z) < -1$. This means that $\overset{0}{N}_1 > N_{11}$ since $\log(\nu) < 0$, where $N_{10} = \frac{x_0}{2 \log(\nu)}$ and $N_{11} = \frac{x_{-1}}{2 \log(\nu)}$ are the solutions of eq. (93) for respectively the first and second branches of the Lambert W function.

The second derivative of the left part of the Lagrangian cost function is given by :

$$\frac{\partial^2 L}{\partial N^2} = 2 \, b \, \log(\nu) + 2 \, d \, \log(\nu) e^{2 \, \log(\nu) \, N} , \tag{95}$$

where $b \geq 0$ and $d < 0$ for $\nu \in]0, 1[$ and $\lambda \geq 0$. The second derivative is therefore positive when

$$2 \, d \, \log(\nu) e^{2 \, \log(\nu) \, N} > -2 \, b \, \log(\nu) , \tag{96}$$

[7] The atom length is larger than one bit in our application since the number of atoms is clearly larger than two to form over-complete dictionaries

or equivalently

$$e^{2\ \log(\nu)\ N} > -\frac{b}{d}\,. \tag{97}$$

From eq. (93),

$$e^{2\ \log(\nu)\ N_1} = e^{-\frac{a}{b}}\ e^{-W\left(\frac{d}{b}\ e^{-\frac{a}{b}}\right)}\,. \tag{98}$$

Since, by definition, $e^{-W(z)} = \frac{W(z)}{z}$ for $z \neq 0$, we have :

$$e^{2\ \log(\nu)\ N_1} = \frac{b}{d}\,W\left(\frac{d}{b}\,e^{-\frac{a}{b}}\right)\,. \tag{99}$$

Finally, from eq. (97),

$$e^{2\ \log(\nu)\ N_{10}} = \frac{b}{d}\,W_0\left(\frac{d}{b}\,e^{-\frac{a}{b}}\right) < -\frac{b}{d}\ , \tag{100}$$

and

$$e^{2\ \log(\nu)\ N_{11}} = \frac{b}{d}\,W_{-1}\left(\frac{d}{b}\,e^{-\frac{a}{b}}\right) > -\frac{b}{d}\ , \tag{101}$$

since by definition $0 \geq W_0(z) \geq -1$ and $W_{-1}(z) < -1$. The second derivative is therefore positive for $N = N_{11}$ and negative for $N = N_{10}$. Hence the Lagrangian cost function presents a minimum at $N = N_{11}$ and a maximum at $N = N_{10}$. Since $N_{11} < N_{10}$, the Lagrangian cost function is first a decreasing function of N, for N small. Moreover, the maximum of the left part of the function is larger than N_B, since

$$-W_0\left(\frac{d}{b}\,e^{-\frac{a}{b}}\right) < 1 + 2\ B\ \log(2) - \log(\nu)\,, \tag{102}$$

for $\nu \in\ ]0,1[$ and B positive.

Finally, it remains to show that the optimal solution for the left part of the Lagrangian function is positive and smaller than N_B. The minimum N_{11} is first positive if the following condition is respected :

$$\lambda < \frac{4\ \|f\|^2\ \log(2)\ \log(\nu)}{W_{-1}\left(3\ 2^{3-2\ B}\ \nu\ \log(\nu)\ e^{-1}\right)}\,. \tag{103}$$

This means that, in the worst case for the Lambert function,

$$\lambda < 4\ \|f\|^2\ \log(2)\ \log(\frac{1}{\nu})\,. \tag{104}$$

This imposes a limit on the weight of the rate in the Lagrangian cost function. Second, N_{11} is smaller than N_B if :

$$-W_{-1}\left(3\ 2^{3-2\ B}\ \nu\ \log(\nu)\ e^{-1}\right) > 1 - \log(\nu) + 2\ \log(2)\ B\,. \tag{105}$$

Since $W(z) = a$ is equivalent to $a\ e^a = z$, by definition, the previous condition is equivalent to :

$$B > -\frac{23\ \log(\nu) + 1}{2\ \log(2)}\ . \tag{106}$$

It has to be noted that both conditions from Eqs (89) and (106) are mutually exclusive. If one constraint is respected, the other cannot be satisfied. Moreover, it can be shown that one of the constraint is always satisfied. This ensures the existence of one and only one minimum to the Lagrangian cost function. Indeed, the product of both conditions is always positive.

$$(-\frac{23\ \log(\nu) + 1}{2\ \log(2)} - B)\ (-\frac{12\ \log(\nu) + \frac{\log(\nu)}{\nu^2-1}}{\log(2)} - B) \geq 0\ . \tag{107}$$

The previous relation is always positive since both terms have the same sign. Indeed, if

$$B + \frac{12\ \log(\nu) + \frac{\log(\nu)}{\nu^2-1}}{\log(2)} \leq 0\ , \tag{108}$$

then

$$B + \frac{23\ \log(\nu)}{2\ \log(2)} \leq -\frac{\log(\nu)}{2\ \log(2)} - \frac{2\ \log(\nu)}{2\ \log(2)\ (\nu^2 - 1)}\ . \tag{109}$$

Furthermore, if eq. (108) is respected, then

$$B + \frac{23\ \log(\nu) + 1}{2\ \log(2)} \leq 0\ . \tag{110}$$

From eq. (109), the last relation can be rewritten as :

$$\frac{2\ \log(\nu)}{\nu^2 - 1} \leq 1 - \log(\nu)\ . \tag{111}$$

Setting $\nu = e^{-\frac{t}{2}}$, for $t \in]0, \infty[$, it is equivalent to prove that

$$t \geq (e^{-t} - 1)\ (1 + \frac{t}{2})\ , \tag{112}$$

where both left and right terms are null for $t = 0$. The left terms increases with t with a slope 1. The right term also increases with t, since its first derivative is always positive, but the increasing rate is always smaller or equal to 1. This proves the inequality of eq. (112).

It can be shown in the same way that if one of the factor of eq. (107) is negative, the other one is negative also. Developing this conditions ends up in proving again the relation of eq. (112). This means either the first conditions is satisfied and not the second one, or the reverse. This ensures the existence of one and only one solution to the optimal number of iterations to transmit for a given rate.

REFERENCES

[1] Cvetkovic Z., "Accurate Subband Coding with Low Resolution Quantization," in *Proceedings of the IEEE Data Compression Conference*, 1998, pp. 448–457.

[2] Mallat S.G. and Zhang Z., "Matching Pursuits With Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[3] Goyal V.K., Vetterli M. and Thao N.T., "Quantization of Overcomplete Expansions," in *Proceedings of the IEEE Data Compression Conference*, 1995, pp. 13–22.

[4] Goyal V.K., Vetterli M. and Thao N.T., "Quantized Overcomplete Expansions in R^N: Analysis, Synthesis and Algorithms," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 16–31, January 1998.

[5] Cvetkovic Z., "Source Coding with Quantized Redundant Expansions: Accuracy and Reconstruction," in *Proceedings of the IEEE Data Compression Conference*, 1999, pp. 344–358.

[6] Cvetkovic Z. and Vetterli M., "Overcomplete Expansions and Robustness," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 1996, pp. 325–328.

[7] Mallat S., *A Wavelet Tour of Signal Processing*, Academic Press, 2 edition, 1999.

[8] Davis G., *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, September 1994.

[9] Davis G., Mallat S. and Avellaneda M., "Adaptive Greedy Approximations," *Journal of Constructive Approximations*, vol. 13, pp. 57–98, 1997.

[10] Cvetkovic Z. and Vetterli M., "Tight Weyl-Heisenberg Frames in $l^2(Z)$," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1256–1259, May 1998.

[11] Cvetkovic Z., "Analysis of Errors in Quantization of Weyl-Heisenberg Frame Expansions and Oversampled A/D Conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, vol. 3, pp. 1435–1438.

[12] Goyal V.K. and Vetterli M., "Dependent coding in quantized matching pursuit," in *Proceedings of the SPIE - Visual Communication and Image Processing*, 1997, vol. 3024, pp. 2–12.

[13] Neff R. and Zakhor A., "Adaptive Modulus Quantizer Design for Matching Pursuit Video Coding," in *Proceedings of the IEEE International Conference on Image Processing*, 1999, vol. 2, pp. 81–85.

[14] Gharavi-Aikhansari M., "A model for entropy coding in matching pursuit," in *Proceedings of the IEEE International Conference on Image Processing*, 1998, vol. 1, pp. 778–782.

[15] Ali S.T., Antoine J.-P. and Gazeau J.-P., *Coherent States, Wavelets and Their Generalizations*, Springer-Verlag, 2000.

[16] Antoine J.-P., Vandergheynst P. and Murenzi R., "Two-dimensional directional wavelets in image processing," *International Journal on Imaging Systems and Technology*, vol. 7, pp. 152–165, 1996.

[17] Bernier D. and Taylor K., "Wavelets from square-integrable representations," *SIAM Journal on Mathematical Analysis*, vol. 27, no. 2, pp. 594–608, 1996.

[18] Jones L.K., "On a Conjecture of Huber Concerning the Convergence of Projection Pursuit Regression," *The Annals fo Statistics*, vol. 15, no. 2, pp. 880–882, 1987.

[19] Sloane N.J.A., Hardin R.H. and Smith W.D., "Nice arrangements of points on a sphere in various dimensions," http://www.research.att.com/ njas/packings/.

[20] Sloane N.J.A., "The Sphere Packing Problem," in *Proceedings ot the International Congress on Mathematics*, Berlin, 1998, vol. 3, pp. 387–396, Documenta Mathematika, III.

[21] Figueras R.M., "Image coding with Matching Pursuit," M.S. thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, August 2000.

[22] Everett H., "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research*, vol. 11, pp. 399–417, 1963.

[23] Shoham Y. and Gersho A., "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 9, pp. 1445–1453, September 1988.

[24] Ramchandran K., Ortega A. and Vetterli M., "Bit Allocation for Dependent Quantization with Applications to Multiresolution and MPEG Video Coders," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 533–545, September 1994.

[25] Ortega A. and Ramchandran K., "Rate-Distortion Methods for Image and Video Compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, November 1998.

[26] Schuster G.M. and Katsaggelos A.K., *Rate-distortion based video compression: optimal video frame compression and object boundary encoding*, Kluwer Academics Publishers, 1997.

[27] Corless R.M., Gonnet G.H., Hare D.E.G., Jeffrey D.J. and Knuth D.E., "On the Lambert W Function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[28] Schuster G.M. and Katsaggelos A.K., "An Optimal Quadtree-Based Motion Estimation and Motion-Compensated Interpolation Scheme for Video Compression," *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1505–1523, November 1998.

[29] Ramchandran K. and Vetterli M., "Best Wavelet Packet Bases in a Rate-Distortion Sense," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 160–175, April 1993.

[30] Al-Shaykh O.K., Miloslavsky E., Nomura T., Neff R. and Zakhor A., "Video Compression Using Matching Pursuits," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 123–143, February 1999.

[31] Neff R., Nomura T. and Zakhor A., "Decoder Complexity and Performance Comparison of Matching Pursuit and DCT-Based MPEG-4 Video Codecs," in *Proceedings of the IEEE International Conference on Image Processing*, Chicago, IL, October 1998, vol. 1, pp. 783–787.

[32] Gribonval R., "A counter-example to the convergence of greedy algorithms under very weak conditions," *Journal of Approximation Theory*, March 2000, submitted paper.

[33] Kothe G., *Topological Vector Spaces I*, Springer-Verlag, Berlin, 1969.

[34] Gabor D., "Theory of Communication," *The Journal of the Institution of Electrical Engineers*, vol. 93, pp. 429–457, 1946.

[35] Antoine J.-P., Murenzi R., Vandergheynst P. and Ali S.T., *2-D wavelets and their relatives*, Cambridge University Press, 2000, to appear.

[36] Gray R.M., "Vector Quantization," *IEEE ASSP Magazine*, pp. 4–29, April 1984.

[37] Barut A.O. and Raczka R., *Theory of Group Representations and Applications*, ISBN: 997150216X. World Scientific Pub Co, February 1987.